Multiple Vehicle Collisions:
An Application of the Monte Carlo Method

John Daily
Jeremy Daily

28-30 September, 2004

Presented at
Pensylvania State Police
Traffic Crash Reconstruction Seminar

Abstract

A case study of a four vehicle collision is used to illustrate the use of the Monte Carlo simulation method in traffic crash reconstruction. The evidence gathered at the scene along with scene photographs and witness statements must be translated into random variables. These random variables are used to solve for the initial vehicle speeds based on the principles of the conservation of linear momentum and rectilinear motion. A common approach would be to use low and high values to compute a range of initial speeds. This leads to large variations with complex problems and the probability associated with the extreme cases is very small. Also, multiple vehicle collisions require special consideration because the results of the first collision effect the second collision and so forth. As such, the Monte Carlo method is used to propagate the uncertainty of the evidence through to the initial speeds of the involved vehicles. In order to use the Monte Carlo method, a brief overview of descriptive statistics and probability theory is presented. The simulation results are given as estimated probability density functions from which speeds are determined using confidence intervals. The most important aspect of using a simulation technique in traffic crash reconstruction is input variables and simulation results must correlate to the actual physical evidence.

Contents

1 Introduction and Background

1.1 The Crash

The traffic crash for this case study occurred at 1623 hours on a clear, warm spring day. Driver A was driving a 1995 Pontiac Grand Am 2 door and was leaving a technical school to head southbound on a US highway. Driver B was driving northbound on the US highway in a 1993 Dodge Dakota pickup. Driver A pulled into the path of the Dodge. There was a collision between the two vehicles in the northbound left lane. The Pontiac spun away from this collision, rotating about 270 degrees counter-clockwise. Driver A was ejected during this collision and slid westerly across the roadway. The Dodge spun about 180 degrees and came to rest in the southbound lanes.

Before this crash equilibrated, a 1992 lowered Chevy pickup, driven by driver C, collided with the Pontiac in the right rear tire area. This collision caused the Pontiac to rotate about 30 degrees clockwise. This re-directed the path of the Chevy, and it exited the road to the west. Before this crash equilibrated, a 1997 GMC pickup, driven by driver D, entered the crash scene and drove over driver A. This was confirmed by the investigating officer by identifying biological material on the front and under the GMC. Shortly after its collision with driver A, the GMC collided with the front of the Pontiac. This caused the vehicles to leave the road to the west. Driver A was drug onto the west shoulder by the GMC, which deposited his body there. The GMC continued on in an arcing path and came to rest just past the right-of-way fence. The Pontiac came to rest just off the west shoulder area. The final rest positions are shown on the diagram shown in figure 1.1 on the following page.

1.2 The Investigation

There was concern that the initial report listed a speed of 69 mph for the Dodge Dakota driven by driver B. Depending on the time of day, the speed limit through this section of this US highway is either 50 mph (when designated by flashing lights), or 65 mph (normal traffic conditions).

Figure 1.1 This scene diagram was produced from at-scene measurements, post-scene measurements and an aerial photograph. The original drawing was on legal paper so scale was violated to create this illustration. The vehicles are depicted multiple times to illustrate each trajectory.

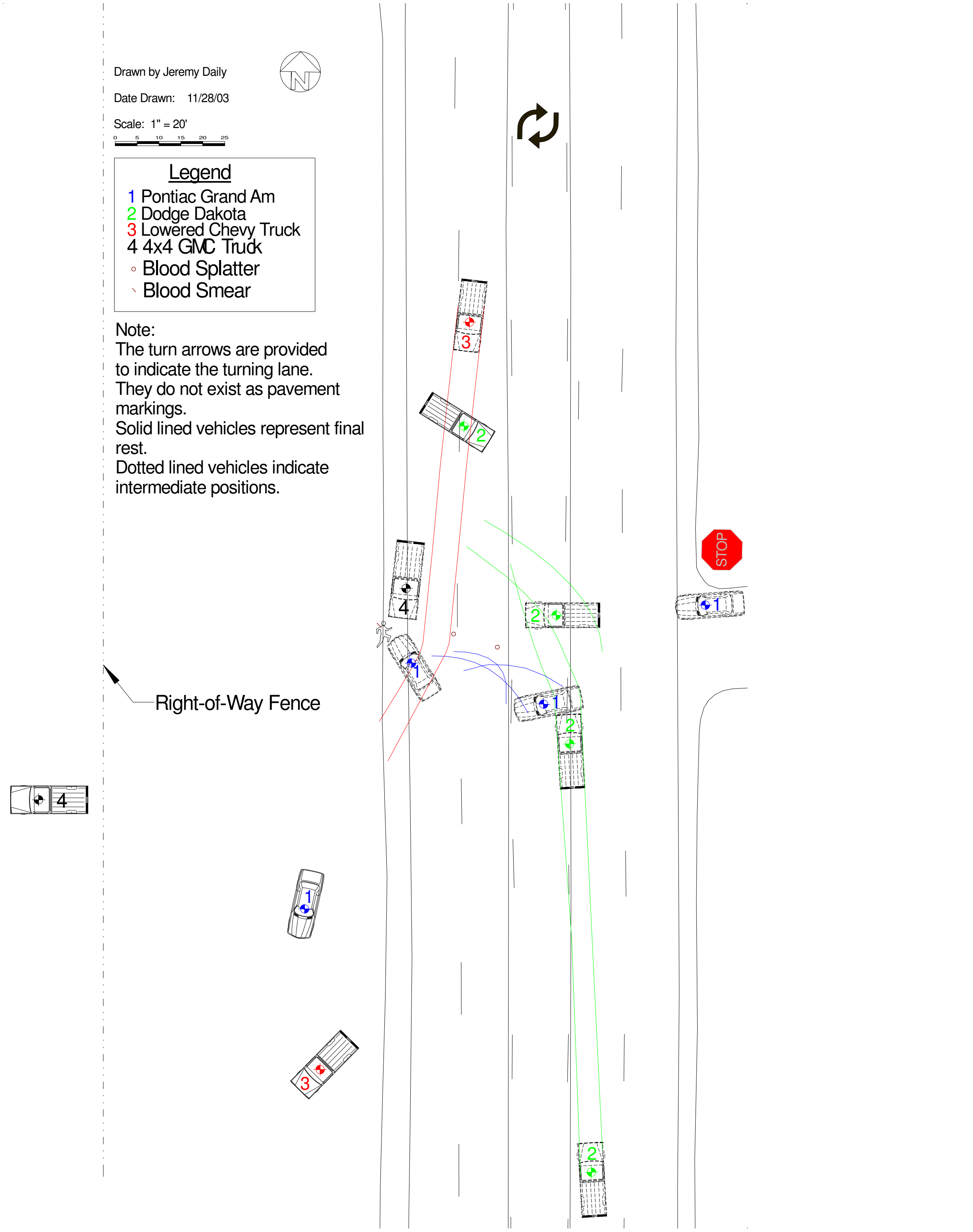


Photo 1 This photo shows the skid marks left by the lowered Chevy truck as it skidded to impact with the Pontiac. Note how the skid marks of the Chevy truck are offset to the right from the collision. This indicates the Pontiac was still in motion when the collision between the two vehicles occurred.

According to the information received from the local Department of Transportation, the flashing light indicating a speed limit of 50 mph would have been flashing. The flashing light would have been on from 1600 hrs to 1645 hrs.

All of the evidence strongly supports the collision sequence outlined in the synopsis and initially reported by the work of the at-scene investigating officer. Photos 1 and 2 on the following page give examples of the scene photos that corroborate the collision sequence.

1.3 Analysis and Reconstruction

The investigation has shown the collision sequence is as reported, and that the speed limit in effect at the time of the crash for traffic on the US Highway was 50 mph. The purpose of the reconstruction will be to examine the speeds of the vehicles both at their respective impacts as well as their speeds coming into impact (start of skids). Because this crash did not reach equilibrium until after the third crash between the GMC and the Pontiac, we must analyze each impact in reverse order. The results of the third impact must be used to analyze the second impact, which must then be used to look at the first impact. Because of the dependence each

Photo 2 Final rest position of the Dodge Dakota. Note the skid marks from the Chevy pickup go underneath the Dodge, yet there was no collision between the Chevy and the Dodge. This is clear evidence the collision between the Dodge and the Pontiac had not come to equilibrium before other vehicles interacted with the first crash vehicles.

impact has on another, each collision will be analyzed stochastically using the Monte Carlo simulation method.

We will begin by making the assumption the friction coefficient μ of the roadway ranges between 0.70 and 0.80 on the road surface and μ is between 0.2 to 0.3 off the road with no braking. We will also assume the collisions are inelastic, and may be solved using conservation of linear momentum techniques. The collision between the Pontiac and the Chevy pickup was essentially a sideswipe, and a momentum solution is the only viable technique to use for this collision. A damage analysis is not appropriate, because a damage analysis requires the centroids of the damage areas to reach a common velocity, which does not happen in a sideswipe. We will consider the impact velocities calculated will form a normal (Gaussian) distribution in order to perform the momentum confidence interval analysis. Finally, the collisions will be modeled with a Monte Carlo simulation to examine the effect each variable has on the impact velocity computation for that specific collision. Once the collision speeds are known, it is a simple matter to calculate the speeds of the various vehicles at the start of their respective skids.

2 Dealing With Uncertainty

This section leaves the case study to introduce or refresh some basic concepts of statistics and probability to the reader. The concepts of descriptive statistics, random variable and probability functions are discussed. Some small examples are provided to illustrate some of the concepts. The discussion provide on percentiles is not necessary for understanding the Monte Carlo method but the information is valuable as a reference.

When dealing with uncertain phenomenon, it is necessary to describe the uncertainty in some fashion. The study of probability and statistics is used to articulate the amount of uncertainty. In accident reconstruction, uncertainty looms in every calculation performed. There are a number of reasons for this and two obvious reasons stand out:

1. Simplification of the actual physics into manageable mathematical expressions will induce uncertainty.

2. The evidence is never perfect.

Take the second case above and think about having 20 different people measuring the same skid mark. All of the numbers will be different (see reference [1]). What if the mark was measured 30 times? Each measurement is a sample of the all the measurements. The total number of measurements that *could* be made is the population of measurements. Somewhere within the range of measured values presumably lies the actual value. Since measuring something an infinite number of times is infeasible, a sample must be taken. The same holds true for all experimental and statistical studies. The field of statistical inference is devoted to describing the population based on a few samples.

2.1 Descriptive Statistics

Two traffic crash reconstruction classes participated in a field study during which each member was asked to walk 100 feet. The time for each person was recorded from which speed could be calculated. The data are as follows (in mph):

Pedestrian Walking Speeds (mph)									
3.34	3.13	3.06	3.24	2.92	2.87	2.93	3.31	3.14	3.36
3.18	3.01	3.14	3.08	3.67	3.56	3.24	3.21	3.57	3.76
3.63	3.13	3.62	3.28	3.00	3.79	2.51	3.82	3.25	3.12
2.95	3.25	3.41	2.63	3.13	2.97	3.17	2.97	3.17	2.95
3.16	3.26	3.29	3.10	2.56	4.00	3.29			

This section aims to illustrate the tools used to make sense of these data with just a few estimators.

2.1.1 Central Tendency

2.1.1.1 Mean

The arithmetic mean is often called the "average." There are a number of ways to compute an average, but the mean value is defined as:

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \tag{2.1}$$

Here the summation notation is used. The capital Greek letter sigma (Σ) means to add all the numbers starting from x_1 and finishing with x_n. The variable n denotes the total number of values in the set and $\bar{x}$ is the mean of the variable x. For example, let a sorted set of numbers be defined as

$$X = \{1, 2, 3, 4, 4, 5, 6, 10\} \tag{2.2}$$

so $x_1 = 1$, $x_2 = 2$ and so forth. The total number of entries is $n = 8$ so applying equation (2.1) gives

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1+2+3+4+4+5+6+10}{8} = 4.375 \tag{2.3}$$

If an extreme outlying value is in the set of data, then the arithmetic mean will be influenced by that outlier. Any measure of central tendency cannot account for the spread of the data. The advantage of using the mean is it is an unbiased estimator of the true population mean. This means the population mean and the sample mean become the same number as the number of sample are approach the number in the population.

2.1.1.2 Median

The median is the central number. It is used when dealing with percentiles and is a good way of determining central tendency for human performance. The median is not as influenced by outliers when compared to the mean. Finding the median is as simple as finding the middle value in a sorted list. This is also the 50th percentile. From the data set described in equation (2.2) the median value is 4.

2.1.1.3 Mode

The mode is the value that is the most frequent. If a case arises wherein a data set has more than one mode, then the set is said to be multi-modal. The data set described in equation (2.2) has a mode of 4. An example of a multi-modal data set with a mode of 1.5 and a mode of 2.1 is

$$X = \{1.2, 1.5, 1.45, 1.5, 1.5, 1.5, 1.6, 1.8, 2.1, 2.1, 2.1, 2.5\} \tag{2.4}$$

2.1.2 Quantifying Variation

2.1.2.1 Range

The range values are the easiest to understand and have been the hallmark of crash reconstruction for years. All that is needed is the maximum and the minimum of the data set and the range is simply the difference.

2.1.2.2 Sorting Data

Sorting is necessary to determine some statistical qualities. There are two appropriate ways to sort: from maximum to minimum or from minimum to maximum. The term “index” is used to refer to the position of a value when sorted from minimum to maximum and the lowest value is labeled as #1. Use the term “rank” to refer to the position of a value when sorted from maximum to minimum and the highest value is labeled as #1.

2.1.2.3 Quartiles and Percentiles

Percentile measure tells how much of the data is below and how much is above a certain value. To calculate the index of any percentile use the formula:

$$index = \frac{P}{100}(n-1)+1 \tag{2.5}$$

$$rank = n - \frac{P}{100}(n-1) \tag{2.6}$$

where P is the percentage and n is the total count. Equation (2.5) is used if the data are sorted with the lowest value being labeled as #1 and equation (2.6) is used if the highest values is the first data point.

The value of the percentile or the percentile point (PP) is the number corresponding the the index or rank. If the value of *index* or *rank* has a fractional part (i.e. not a whole number) then the percentile point must be interpolated between the two closest points. For example, take the 65th percentile of the dataset in equation (2.2). Since X is sorted ascending, the calculated index is $\frac{65}{100}(8-1)+1 = 5.55$ and the percentile point needs to be between the values at the 5th and 6th position ($x_5 < PP < x_6$). For the example, $4 < PP < 5$.

Here is how to interpolate:

1. Take the calculated quantity *index* and split it at the decimal. Example: $index = 5.55$

2. Call the whole number i and call the fraction or remainder R. Example: $i = 5$ and $R = 0.55$

3. Now x_i is the lower value before the index and x_{i+1} is the higher value after the index. Example: $x_5 = 4$ and $x_6 = 5$

4. The percentile point (PP) follows:

$$PP = x_i + R(x_{i+1} - x_i) \tag{2.7}$$

   So for the example: $PP = x_5 + R(x_6 - x_5) = 4 + 0.55(5-4) = 4.55$

The procedure is slightly different at step 3 if using data sorted with #1 being the maximum.

1. Take the calculated quantity *rank* and split it at the decimal.

2. Call the whole number k and call the fraction or remainder R.

3. Now x_k is the *higher* value before the rank and x_{k+1} is the *lower* value after the rank.

4. The percentile point (PP) follows:

$$PP = x_k + R(x_{k+1} - x_k) \tag{2.8}$$

A quartile is a measure of the quarter points of the spread. The first quartile Q_1 is the 25th percentile so the 25% of the values are lower and 75% of the values are higher. The middle quartile is also the median and the third quartile Q_3 is the 75th percentile so 75% of the values are lower. These quartiles are used to construct the box for a box and whisker plot. The formulas for the index or rank of the quartiles are determined by letting $P = 25$ or $P = 75$ in equations (2.5) or (2.6).

$$index(Q_1) = \frac{n+3}{4} \tag{2.9}$$

$$index(Q_3) = \frac{3n+1}{4} \tag{2.10}$$

or

$$rank(Q_1) = \frac{3n-1}{4} \tag{2.11}$$

$$rank(Q_3) = \frac{n-1}{4} \tag{2.12}$$

A decile is a measure of the data in tenths whose indicies or ranks can be determined with equations (2.5) or (2.6).

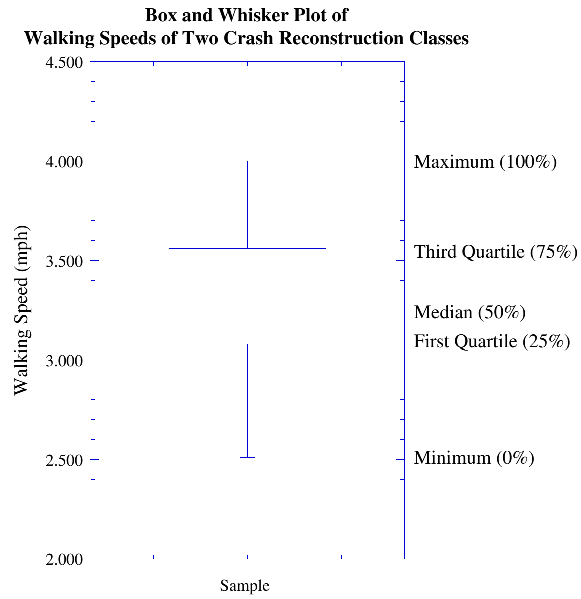
2.1.2.4 Box and Whisker Plot

Figure 2.1 on the next page shows a box and whisker plot that gives a five number summary and visualization of the data. That summary includes: the maximum, the minimum, the median, the first quartile and the third quartile. All values are based on relative standing (Percentile) and give a good picture of the data distribution.

2.1.2.5 Variance

At first glance, a way of quantifying the amount of spread of a set of data would be to take the difference from each data point to the mean then find the mean of those differences. However,

Figure 2.1 A Box and Whisker plot showing the results of the five number summary.



the differences can be both positive and negative so the average would be zero. As such, the variance (s^2) is quantified by the mean value of the *square* of the differences.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{2.13}$$

Notice the denominator of equation (2.13) is $n-1$ and not just n. This is because the variance is calculated based on a sample mean and not the population mean. Think of it this way, if the sample mean $\overline{x}$ is known, then how many more points are required to completely describe the data? The answer is $n-1$ because the last data point can be determined using equation 2.1 on page 6. This is the concept of *degrees of freedom* in statistics. The total number of degrees of freedom from a set of data is n and calculation a moment such as the variance only requires n unknowns. If the mean is already known, the only, $n-1$ more points are required. If the number of samples become very large then the difference between n and $n-1$ is negligible. Since the Monte Carlo method has a large number of samples, it does not make a difference which form of the denominator is used. The variance equation is also an unbiased estimator of the true population because, like the mean, it can be shown that the mathematical expectation of the estimated variance is the same as the population variance.

Another point of interest is the variance formula is often expressed in a different form, namely

$$s^2 = \frac{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}{n-1} \tag{2.14}$$

This give the same result as equation (2.13) but only requires one subtraction instead of n. If you apply a binomial expansion to the numerator of equation (2.13), the result is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n-1} \tag{2.15}$$

By a property of a sum that basically says $C+C+C+C = 4C$ when C is a constant we get

$$s^2 = \frac{\sum_{i=1}^{n}(x_i^2) - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2}{n-1} \tag{2.16}$$

If both sides of equation (2.16) are multiplied by $\frac{n}{n}$ it doesn't change the value but make things simpler

$$s^2 = n\left[\frac{\sum_{i=1}^{n}(x_i^2)}{n(n-1)} - \frac{2\bar{x}\frac{\sum_{i=1}^{n} x_i}{n}}{n-1} + \frac{\bar{x}^2}{n-1}\right] \tag{2.17}$$

Recall the equation for the mean and simplify

$$s^2 = n\left[\frac{\sum_{i=1}^{n}(x_i^2)}{n(n-1)} - \frac{2\bar{x}^2}{n-1} + \frac{\bar{x}^2}{n-1}\right] \tag{2.18}$$

The last two terms combine and the n is multiplied through to get

$$s^2 = \frac{\sum_{i=1}^{n}(x_i^2)}{n-1} - \frac{n\bar{x}^2}{n-1} \tag{2.19}$$

Combine with a common denominator to get equation (2.13). These formulas may seem daunting but a modern spreadsheet, such as Excel, can compute these quickly.

2.1.2.6 Standard Deviation

The units from a variance calculation are squared. So to alleviate this problem, the standard deviation is defined as the square root of the variance and it has the same units as the mean.

$$s = \sqrt{s^2} \tag{2.20}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.21}$$

2.1.2.7 Coefficient of Variation

Since both the standard deviation and the mean have the same units, their ratio can be taken. This value is called the *coefficient of variation* is defined as

$$C = \frac{s}{\bar{x}} \tag{2.22}$$

The coefficient of variation is useful in describing the relative spread of some distribution, regardless of the magnitude of the numbers.

2.1.3 The Difference Between a Population and Sample

A population is the collection of all possible values. The total number can be a discrete amount like number of cards in a deck. However, in most crash reconstruction applications, there are an infinite number of values each variable can be assigned within some range. When dealing with a population, Greek letters are used to describe the population characteristics. The values determined for the mean and standard deviation are considered unbiased because as the number of data increase, they approach the true population parameters.

$$\bar{x} \rightarrow \mu$$
$$s^2 \rightarrow \sigma^2$$

where μ is the mean and σ^2 is the variance.

2.1.4 Skewness

Skewness is a measure of how asymmetric the data may be. Often human performance data is skewed. This paper does not make use of the skewness and a brief overview of the skewness can be found in reference [2] and a more in depth discussion is in [3]. In a nutshell:

- If mean > median, then right skewed (high maximum)
- If mean < median, then left skewed (low maximum)
- If mean = median, then symmetric (balanced)

2.1.5 Histograms

Generating and understanding histograms is paramount to understanding how data is distributed. Histograms provide an excellent visual aid to communicate the results of a study. Figure 2.2 on the next page show a typical histogram. For a definition:

A histogram is a bar graph showing the number of occurrence of all possible events. It is also known as a frequency diagram

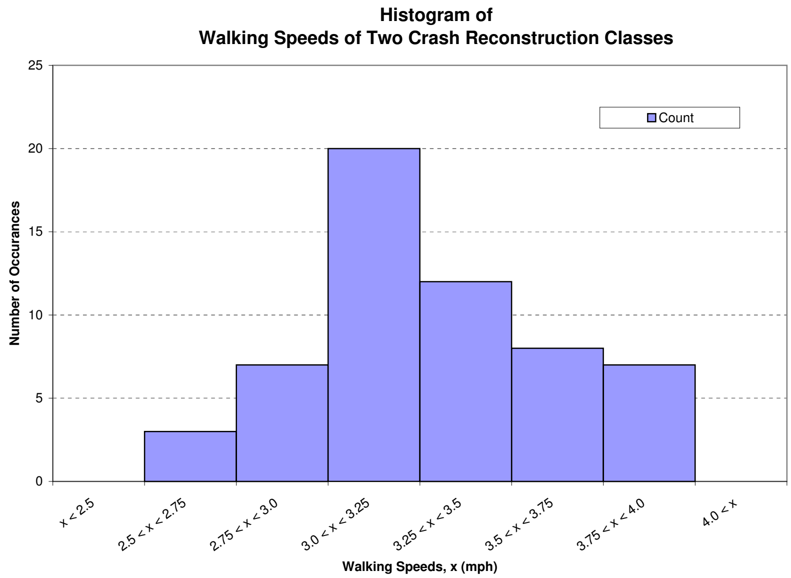
In most cases, the event is a number falling in a certain range. However, an event could be the occurrence of a 6 on a die or some other discrete event. To construct a histogram, follow this procedure:

1. Find the maximum and minimum values of a set. Call them u_{max} and u_{min}. The max can be rounded up and the min can be rounded down if desired.
2. Divide the x-axis into into m bins of a fixed width Δx, so $\Delta x = \frac{u_{max} - u_{min}}{m}$. The number of bins, m is to be determined at your discretion, however a guideline formula for the number of bins is $m = \text{ceil}[1 + 3.3\log_{10}(N)]$ where ceil means to round up to the next integer and N is the number of data entries. [1]
3. Sort the data and place it in the appropriate bin.
4. Count how many points are in each bin to determine the frequency. The ith bin will have c_i points.
5. Make a bar chart of the frequency vs. bin range.

For a good looking histogram, use between 10 and 30 bins. Also, if the number of entries in the data set is small, the histogram may not look very good.

[1] The ceil[argument] operator stands for ceiling and is a common function found in many computer programming languages that rounds the argument to the next integer.

Figure 2.2 A histogram of the walking speed data from this section. The histogram will become more smooth as the number of data points are increased.



2.2 Random Variables

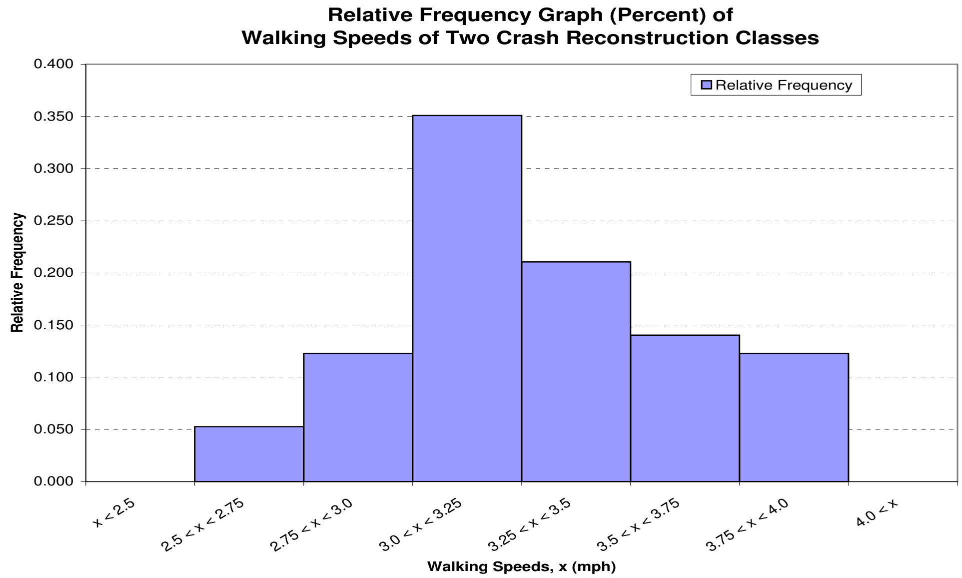
Defining a random variable is the process of assigning a random event a numerical value. For a simple coin toss, the head could be mapped to a 1 and the tail mapped to a 0. This sets up the random variable that is predictive in nature. In other words, we can use the laws of probability to predict the outcome of an event or series of events.

In accident reconstruction, an abstraction is required because the random variable represents a value that is not really random. Consider the drag factor of the vehicle/road system during the crash. That average drag factor has a unique value for that crash. In fact, every vehicle within a crash has its own unique values, we just are never there to measure them. So, in reality, the drag factor is not a random variable. However, we can treat it like a random variable because we lack the ability to identify its actual value. To put it another way, the real value of the drag factor is unknown, but we have a good idea of what it could be.

2.2.1 Histograms to Estimated Probability Density Functions

The classic definition of probability says the likelihood of an event happening is the ratio of the event to the number of tries. This assumes the number of tries goes to infinity. Again, this is the

Figure 2.3 The relative frequency diagram from the histogram of figure 2.2.



concept of a population.

$$P(event) = \frac{events}{N} \tag{2.23}$$

For a continuous random variable the total number of events (N) approaches infinity and events are written as an inequalities. In other words, an event could be picking a value greater than or less than some variable, e.g. $P(value < x) = \lim_{N\to\infty}(value < x)/N$. Describing the event as a random variable falling in a given range is even more useful $P(x_1 < value < x_2)$.

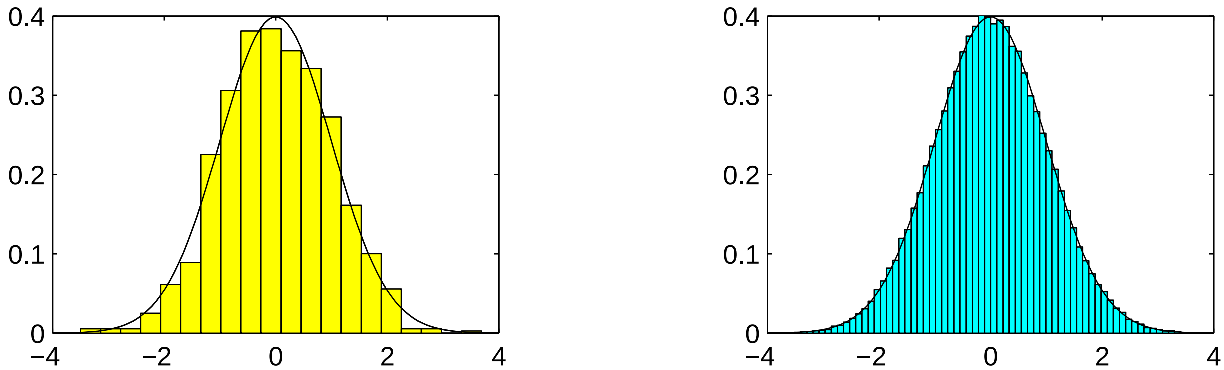
Considering the histogram in figure 2.2, we may convert it into a relative frequency diagram by dividing the y-axis by the total number of samples. The result of this operation is shown in figure 2.3. Now look at the definition of probability from equation (2.23) and recognize that the relative frequency diagram is now able to give us probabilities of the intervals.

The *probability density* is given as the probability of some interval divided by the width of the interval. The formula below gives the estimated probability density given a histogram of a sample .

$$f_x(x_i) = \frac{P(x_i < x \leq x_i + \Delta x)}{\Delta x} \tag{2.24}$$

where $f_x(x_i)$ is the probability density of the variable x.
Δx is the width of the bins of the histogram.
$P(x_i < x \leq x_i + \Delta x)$ is the probability of finding a value (x) in the ith bin of the histogram.

Figure 2.4 This figure shows the transition of an estimated probability density graph to a continuous probability density function (PDF).



(a) Low Resolution

(b) High Resolution

Equation (2.24) is a great definition but not very usable. To create an estimated probability density function, use the definition of probability density and the procedure in creating a histogram (section 2.1.5 on page 13). The probability of an item falling in the ith interval of a histogram is the value of the histogram (c_i) divided by the total number. In light of equation (2.23), the *event* is x falling in the ith interval that is Δx wide and has occurred c_i times. The value for Δx is simply $(u_{max} - u_{min})/m$. Now equation (2.24) can be written in a useful form to estimate a probability density function (PDF).

$$f_x(x_i) = \frac{c_i/N}{(u_{max} - u_{min})/m} \tag{2.25}$$

Now to make the last transition to a continuous probability density function, pretend both the number of bins and the number of values increases significantly. In the limit, the bin width will be come zero and the number of bins will be infinite. Also the bar chart of the estimated PDF will become a continuous smooth function. This concept is illustrated with figure 2.4. Now we can write a function $f_x(x)$ to represent the probability density.

2.2.2 Obtaining Cumulative Density Functions.

The reason we want to obtain a cumulative density function (CDF) is it enables us to get actual probability values from the function. This plays an important role in sampling the random variables in the Monte Carlo simulation.

To understand what a cumulative density function is, start with some properties of a probability density function (PDF):

1. The total area under the PDF curve is equal to 1. Using the concept of integration yields

$$\int_{-\infty}^{+\infty} f_x(x)\,dx = 1 \tag{2.26}$$

2. The probability density can never be negative.

If we have a probability density function ($f_x(x)$), then we can get the probability of x being in some interval by finding the area under the PDF curve and between the interval bounds. The cumulative density function (CDF) is the area under the PDF curve starting at its lowest point. Mathematically the expression uses an integral function and is expressed with a capital letter:[2]

$$F_x(x) = \int_{-\infty}^{x} f_x(x)\,dx \tag{2.27}$$

This also implies that the PDF is the derivative of the CDF.

$$f_x(x) = \frac{dF_x(x)}{dx} \tag{2.28}$$

The cumulative density function gives the actual probability of a value being less than x

$$F_x(x) = P(value \leq x) \tag{2.29}$$

Some properties of a CDF include:

1. The values of the CDF must begin at 0 on the left side and end at 1 on the right.

2. The CDF can never decrease in value as x increases.

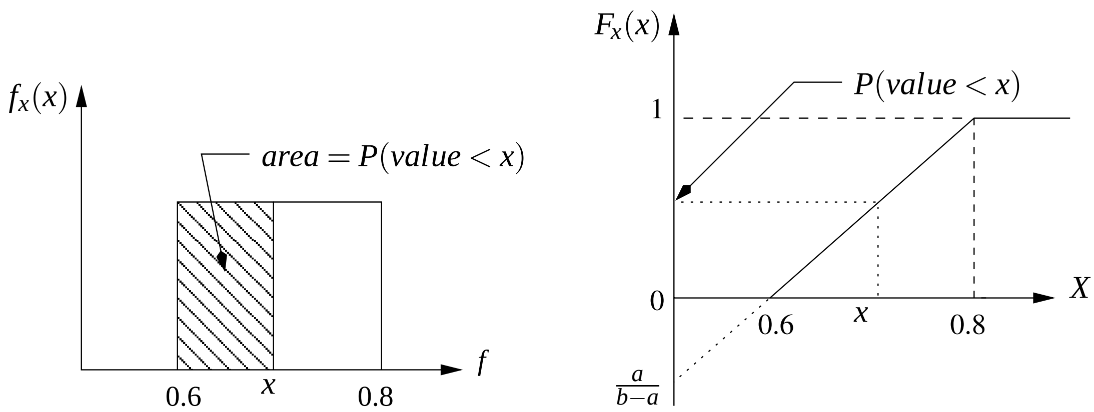
2.2.3 Common Distributions

2.2.3.1 Uniform Distribution

The uniform distribution is the easiest distribution to use and also is the most "conservative" because it gives an equal probability to all numbers within the distribution range. The formula

[2] It is common notation for the probability density function to be represented with a lower case f subscripted with the random variable x and the cumulative density function is denoted as a subscripted uppercase F_x

Figure 2.5 An example of the uniform distribution representing a range of drag factors where $a = 0.6$ and $b = 0.8$.



(a) Probability Density Function

(b) Cumulative Density Function

for the PDF of a uniform distribution from a to b.

$$f_X(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \le x \le b \\ 0 & x > b \end{cases} \tag{2.30}$$

Since the uniform distribution is so common, it is often written in shorthand as $U(a,b)$ as shown in equation (2.30). The cumulative density function for the uniform distribution is

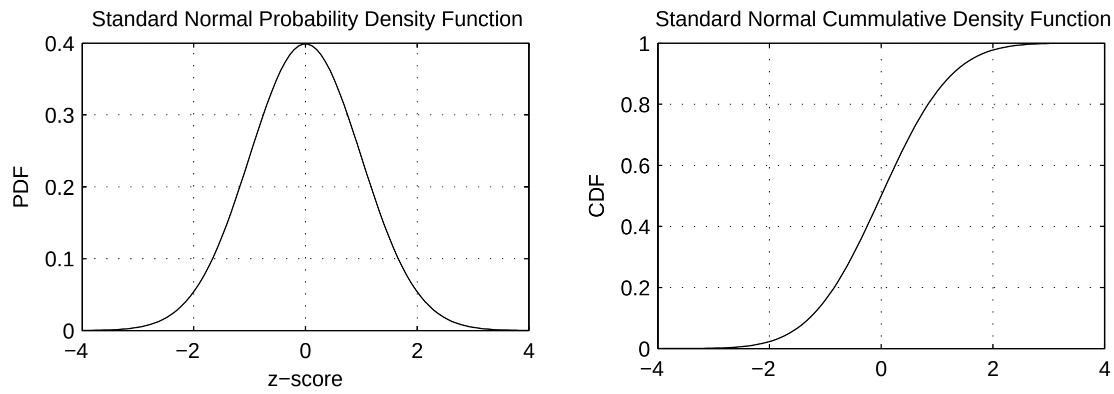
$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a}(x-a) & a \le x \le b \\ 1 & x > b \end{cases} \tag{2.31}$$

Both the PDF and the CDF are shown in figure 2.5. All of the properties are satisfied. Further analysis shows the derivative of the CDF is the same as the PDF. For the uniform case, the derivative of a line is a constant.

$$\frac{dF_X(x)}{dx} = \frac{1}{b-a} \tag{2.32}$$

when x is between a and b.

Figure 2.6 Standard normal probability density and cumulative density functions. The CDF was created using the `NORMSINV` command in a spreadsheet.



(a) (b)

2.2.3.2 Normal Distribution

Figure 2.6 shows the common "bell" curve known as the normal or Gaussian distribution. The function plotted in figure 2.6a is the probability density formula shown below

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.33}$$

This is written shorthand as

$$N(\mu, \sigma) \tag{2.34}$$

The normal distribution is symmetric and only needs the mean and the standard deviation to define it. Since it is symmetric, the mean, median and mode are all the same value.

The CDF of the normal distribution (figure 2.6b) is not known in a closed form so it has to be solved using numerical methods. As a result, the values of the CDF were solved once for a *standard normal* whose mean was zero and standard deviation is 1. Since the standard normal is known, it is useful to translate any normally distributed variable into the standard normal. To do this, use the z- score of a number as defined by

$$z = \frac{x-\mu}{\sigma} \tag{2.35}$$

Once a z-score is known, the value of the CDF can be looked up in a table, solved using a computer, or taken off a graph. Also, there are many text books with z tables in the appendix [4].

2.2.3.3 Other Distributions

There are many more distributions, both analytical and experimental that all satisfy the criteria for a PDF or a CDF. They can be broken down into two broad categories with a few examples of each.

1. Discrete: Uniform, Binomial, Poisson's, and Multinomial

2. Continuous: Uniform, Triangular, Normal, Log-Normal, Exponential, Rayleigh, and Weibull

Please see a basic text on probability for examples of these distributions [3, 4].

2.2.4 Combining PDFs

Characterizing a single random variable is necessary, but is not useful for solving problems. To make use of the random variables, the technique of analytically combining PDFs involves convolution integrals and function inverses. The procedure is very long and complicated for any real problem so instead, alternative methods, like Monte Carlo, are used. This is why we'll use the Monte Carlo method to solve the momentum problems required for this reconstruction.

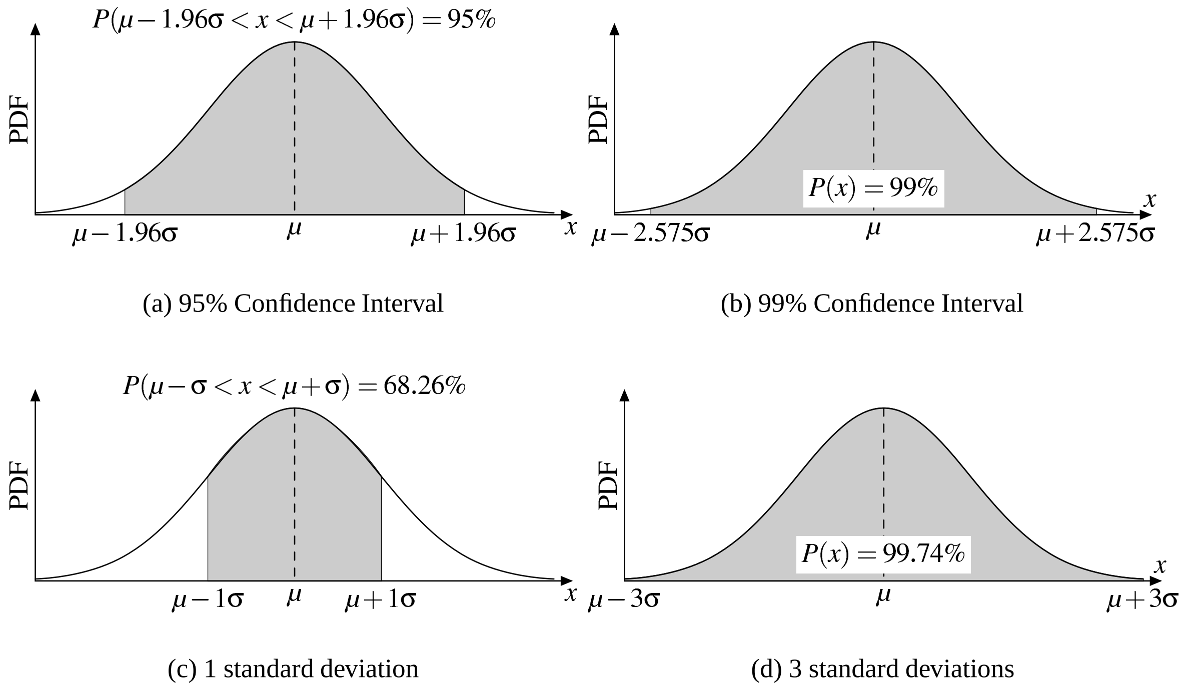
2.2.5 Central Limit Theorem

One of the reasons the normal distribution is so popular is the phenomenon called the central limit theorem.

The Central Limit Theorem leads to the conclusion that large combinations of independent random variables lead to a normal distribution.

This fact is a very nice conclusion and allows us to treat many random variables as normal. This is also why the normal distribution shows up in so many different fields. There are, of course, rigorous mathematical proofs of the central limit theorem, but knowing the proofs exist is all that is required to use it.

Figure 2.7 Common confidence intervals for a normal distribution. The coefficients of the σ terms are the *z*-score corresponding to the area under the curve.



(a) 95% Confidence Interval

(b) 99% Confidence Interval

(c) 1 standard deviation

(d) 3 standard deviations

2.3 Relating Error Estimates to Probability

2.3.1 Confidence Analysis

A confidence interval is defined by the interval that contains a certain probability of occurrence. The best way to show this is through an example. If a 95% confidence interval is desired, then 95% of the area under the PDF curve is contained in an interval $(x_1 \leq x \leq x_2)$. The unknowns are x_1 (the lower limit) and x_2 (the upper limit). Figure 2.7a shows how the 95% confidence interval is made from the bounds $x_1 = \mu - 1.96\sigma$ and $x_2 = \mu + 1.96\sigma$ where μ is the mean and σ is the standard deviation.

A confidence interval can be one sided or two sided. For the remainder of this paper, only the two sided confidence interval is considered because the values can go either way. This means the amount of area not in the confidence interval is divided equally among the tails. The tails are the unshaded regions under the curves in figure 2.7. We use confidence intervals to both set up the random variables as input to a simulation as well as to interpret the output of the simulation.

2.3.2 Picking a Distribution

The evidence from a traffic crash does not usually give any indication which distribution to use for a particular variable. There are instances, such as perception reaction times, where a special distribution is appropriate, but it must be justifiable based on experimental data and previous research. Hopefully, the researchers in the human performance field will be able to define an appropriate probability function to use in a Monte Carlo simulation.

As a rule of thumb, use a uniform distribution for drag and acceleration factors and normal distribution for everything else. When using a normal distribution, the 99.74% confidence interval ($\pm 3\sigma$) is sufficient for a definite range. If there is desire for some leeway on the input ranges then a lower confidence interval is acceptable. The details of creating a random sample from a range is in section 2.4.

2.4 Simulation Techniques

2.4.1 Random Number Generators

The simplest random number generators are dice or coins. Poke your head into a casino and random number and event generators abound. These methods of generating random numbers are valid, but limited by feasibility issues.

Computer generated random numbers require a seed number and an algorithm to generate the numbers. Since the list will be the same if the seed is the same, the numbers are really not truly random. Instead, they are called pseudo-random numbers. A typical seed value is the time/date stored on the computer and a basic algorithm will multiply two 8 bit numbers together to get a 16 bit number. The 8 most significant bits are the answer to the multiplication problem and the 8 least significant bits are “random”. See reference [5] for details on random number generation in a spreadsheet.

Pseudo random numbers are just fine for Monte Carlo simulations and nearly every computer (and some calculators) is capable of generating a uniform distribution from 0 to 1, $U(0,1)$. Many are also capable of generating random numbers from a standard normal distribution $N(0,1)$ as well.

2.4.2 Random Distribution Sampling

To convert from a unit uniform distribution on the computer to any arbitrary distribution of a random variable, we must make use of the fact that a cumulative density function (CDF) will never decrease. This means there is a unique mapping from the y-axis to the x-axis. In other words, if a random number is generated, say 0.57, then we can let that number be the probability of some value less than x.

$$U(0,1) \rightarrow 0.57 = P(value \leq x) = F_x(x) \tag{2.36}$$

Now just solve for x in equation (2.31).

Here is an example using the uniform distribution in figure 2.5 on page 18. Again, assume our random number generator spits out a value of 0.57. Use the equation for the CDF of a uniformly distributed variable.

$$\begin{aligned} F_x(x) = 0.57 &= \frac{1}{0.8-0.6}(x-0.6) \\ 0.57(0.2) &= x-0.6 \\ x &= 0.714 \end{aligned}$$

Performing the above calculation over and over with a randomly generated $F_x(x)$ will transform a uniform distribution from 0 to 1 to a uniform distribution from 0.6 to 0.8. This procedure can be used for any distribution, but solving the CDF for x may require a numerical method.

3 The Monte Carlo Solution Procedure

3.1 Construct the Governing Equations

This multiple vehicle collision is solved using the principal of Conservation of Linear Momentum (COLM). The derivation of the COLM equations can be found in [6, 7] and are presented below:

$$v_2 = \frac{w_1 v_3 \sin\theta}{w_2 \sin\psi} + \frac{v_4 \sin\phi}{\sin\psi} \tag{3.1}$$

$$v_1 = v_3 \cos\theta + \frac{w_2}{w_1} v_4 \cos\phi - \frac{w_2}{w_1} v_2 \cos\psi \tag{3.2}$$

The inputs into equations (3.1) and (3.2) are the magnitudes of the post-impact velocities (v_3, v_4), the weights of the vehicles (w_1, w_2) and the approach and departure angles (θ, ϕ, ψ). The best solution is to weigh the actual vehicles if they are available. If the vehicles are not available, the weights can be looked up but are still assumed unknown where the lower limit is the curb weight and the upper limit is an additional 400 lbs for one passenger and 600 lbs for two passengers. This includes additional baggage and fuel also. Consider all extra weight such as a truck camper shell or a heavy tool box when determining the range of probable weights. The angles are determined from the scale diagram and assigned ranges accordingly. In some cases, the evidence may be very clear for the angles, other times there is a wide possible range.

The exit velocities are obtained using the Work-Energy theorem which reduces to

$$S = \sqrt{30 d f + S_o^2} \tag{3.3}$$

where d is the distance traveled, f is the average drag factor over the distance, and S_o is the speed at the end of the skid. The speed at the end of the skid will be solved by the momentum equations for all but the last collision where the vehicles come to rest. This equation will have to be solved 6 times.

The drag factor f is estimated by experience or determined from a spin analysis. The spin analysis considers the amount of rotation, the distance traveled and the number of wheels locked. The objective of the spin analysis is to find a reduction factor to apply to the coefficient of friction. The reduction factor is presented as a distribution which makes it conducive to Monte Carlo techniques. For more information on the spin analysis see the work presented at Special Problems 2003 by Dr. Gordon Bigg and John Daily [8]. In summary,

$$f = \eta\mu \pm m \tag{3.4}$$

where η is the reduction factor from the spin analysis, μ is the coefficient of friction and m is the slope. This was used when measurable vehicle spin was present. The spin factor is simply 1 when there is no spin.

3.2 Generate the Random Variables

Generating random variable requires an art as well as a science. The random variable is usually articulated in English first and may go something like this:

> "The distance d is most likely 90 feet but may be anywhere from as low as 81 as high as 99."

This kind of understanding of the random variable d would lead to a normal distribution with a mean of 90 and $6\sigma = range = 18$ or the standard deviation would be 3. The random variable can be written in shorthand as

$$d \sim N(90, 3)$$

If a variable is articulated with the phrase "The drag factor f can be anywhere between 0.6 and 0.8" then the best distribution to use is a uniform distribution because it makes each value in the range equally likely. Written in shorthand as

$$f \sim U(0.6, 0.8)$$

One important concept is that of rejecting non-physical values. The normal distribution has the ability to give a number below some threshold which is impossible for weights, drag factors, and reaction time. As such, it may be necessary to validate the random number and the result before using it. See references [9, 10] for more details on rejecting runs. Careful selection of the random variable distributions should mitigate the validation. An example for this case is the exit angle of the Pontiac on the first collision cannot be less than the exit angle of the Dodge. Some unscrupulous people may try to use the Monte Carlo method in ways to better their views by ignoring this concept of rejecting non-physical values.

3.2.1 Third Collision

The third collision occurred when the the 4x4 pickup hit the Pontiac. The post impact trajectories occurred off road with no braking. Neither vehicle spun significantly, as such the coefficient of friction was assigned a uniform distribution from 0.2 to 0.3.

The following table gives the random variables and their distributions for use in analyzing the third collision. The low and high values were computed using the 99.74% (3σ) confidence interval. All distances are in feet, weights are in pounds and angles are in degrees. The first number in the parentheses after Normal is the mean value and the second is the standard deviation according to the shorthand notation for a normal distribution.

Quantity	Variable	Distribution	Low Value	High Value
Weight of 4x4 GMC Truck	w_1	Normal(4100,10)	4070	4130
Weight of Pontiac	w_2	Normal(3024,10)	3084	2994
Coefficient of Friction	μ	Uniform(0.2,0.3)	0.2	0.3
Spin Factor for the 4x4 GMC	η_1	N/A	1	1
Spin Factor for the Pontiac	η_2	N/A	1	1
Distance of the 4x4 GMC	d_1	Normal(90,3)	99	91
Distance of the Pontiac	d_2	Normal(60,2)	54	66
Exit Angle of the 4x4 GMC	θ	Normal(15,5)	0	30
Exit Angle of the Pontiac	ϕ	Normal(10,1.5)	13	7
Entrance angle of the Pontiac	ψ	Normal(80,5)	95	65
Entrance angle of the 4x4 GMC	0	N/A	0	0

3.2.2 Second Collision

The second collision between the lowered Chevy and the Pontiac changed the skid direction of the Chevy until it came to a stop off the road. Because the skids transitioned from the pavement to the shoulder, a combined speed formula is required with two distances and two drag factors. Again, little spin was involved in this crash so $\eta = 1$.

The following table gives the random variables and their distributions for use in analyzing the second collision when the lowered Chevy hit the Pontiac. The low and high values were computed using the 99.74% (3σ) confidence interval. All distances are in feet, weights are in pounds and angles are in degrees. The first number in the parentheses after Normal is the mean value and the second is the standard deviation according to the shorthand notation for a normal distribution.

Quantity	Variable	Distribution	Low Value	High Value
Weight of the lowered Chevy	w_1	Normal(3680,10)	3710	3650
Weight of Pontiac	w_2	Normal(3024,10)	3084	2994
Coefficient of Friction (Off Road)	μ_1	Uniform(0.2,0.3)	0.2	0.3
Coefficient of Friction (On Road)	μ_2	Uniform(0.7,0.8)	0.7	0.8
Spin Factor for the Chevy	η_1	N/A	1	1
Spin Factor for the Pontiac	η_2	N/A	1	1
Dist. of the Chevy (Off Road)	d_1	Normal(60,1)	57	63
Dist. of the Chevy (On Road)	d_2	Normal(30,2)	24	36
Distance of the Pontiac	d_3	Normal(5,1)	2	8
Exit Angle of the lowered Chevy	θ	Normal(25,0.5)	23.5	26.5
Exit Angle of the Pontiac	ϕ	Normal(80,5)	65	95
Entrance angle of the Pontiac	ψ	Normal(100,5)	85	115
Entrance angle of the Chevy	0	N/A	0	0

3.2.3 First Collision

The following table gives the random variables and their distributions for use in analyzing the first collision where the Dodge hit the Pontiac. The low and high values were computed using the 99.74% (3σ) confidence interval.[1] All distances are in feet, weights are in pounds and angles are in degrees. The first number in the parentheses after Normal is the mean value and the second

[1] The results for the spin factor were presented with a 95% confidence interval, not a 99.74%.

is the standard deviation according to the shorthand notation for a normal distribution.

Quantity	Variable	Distribution	Low Value	High Value
Weight of the Dodge Dakota	w_1	Normal(3910,10)	3880	3940
Weight of Pontiac	w_2	Normal(3024,10)	3084	2994
Coefficient of Friction	μ	Uniform(0.7,0.8)	0.7	0.8
Spin Factor for the Dodge	η_1	Normal(0.5,0.05)	0.402	0.598
Spin Factor for the Pontiac	η_2	Normal(0.4,0.05)	0.302	0.498
Distance of the Dodge Dakota	d_1	Normal(65,2)	59	71
Distance of the Pontiac	d_2	Normal(25,1)	22	28
Exit Angle of the Dodge Dakota	θ	Normal(10,3)	1	19
Exit Angle of the Pontiac	ϕ	Normal(80,5)	65	95
Entrance angle of the Pontiac	ψ	Normal(97,6)	79	115
Entrance angle of the Dodge	0	N/A	0	0

3.3 Iterate the Governing Equations

Begin with the concept of a sampling the random variable. That means to pick a single value at random from the distributions. The chances are higher that a number near the mean will result from a normal distribution and any number within the range could result from the uniform distribution. Then solve the three momentum equations (3.1-3.2) and the initial speed equations (3.3). These results should be gathered as a list and stored for analysis and plotting. The procedure is outlined below:

1. Generate a random instance of each variable.

2. Solve the third collision.

   a) Compute the exit speeds, equation (3.3).

   b) Compute the impact speeds, equations (3.1-3.2).

   c) Store the impact speed of the 4x4 truck in a list so we can plot the histogram and reduce the data later.

3. Solve the second collision.

   a) Compute the exit speeds using the impact speeds from step (c) above.

b) Compute the impact speeds, equations (3.1-3.2).

c) Store the impact speed of the lowered truck in a list.

d) Compute the initial speed of the lowered Chevy because of the pre-impact skids. Store this value in a list.

4. Solve the first collision.

   a) Compute the exit speeds using the results from step (d) above.

   b) Compute the impact speeds, equations (3.1-3.2).

   c) Store the pre-skid speed of the Dodge in a list.

   d) Store the impact speed of the Pontiac in a list.

5. Repeat steps 1-4 until the desired accuracy is obtained.

6. Use descriptive statistics and confidence intervals to interpret the results of the list stored in the above steps.

A general rule of thumb is the number of iterations required is expressed as

$$N = \frac{10}{1 - C.I.}$$

where $C.I.$ is the greatest desired confidence interval. So if 99.99% is the greatest confidence interval then the Monte Carlo must have at least $10/(1-0.9999) = 100,000$ iterations. There are more rigorous methods for guaranteeing convergence, but this works quite well.

4 Results and Discussion

The final results are presented as an estimated PDF with a 95% confidence interval in this section. They were generated using equation (2.25). These graphs provide an excellent way to present the results. Vector diagrams of the linear momentum solution are very useful to ensure the impulse vectors are equal and opposite while giving a good representation of the physics involved.

4.1 First Collision

The Dodge and the Pontiac collided first.

1. The impact speed of the Dodge was a nominal 38 mph.

2. The impact speed of the Pontiac was a nominal 36 mph.

3. The speed of the Dodge at the start of skid was 54 to 63 mph.

Figure 4.1 The results of the linear momentum vector diagram based on the means give $\Delta v_1 = 13.22$ mph and $\Delta v_2 = 17.10$ mph. Vehicle #1 is the Dodge Dakota and vehicle #2 is the Pontiac.

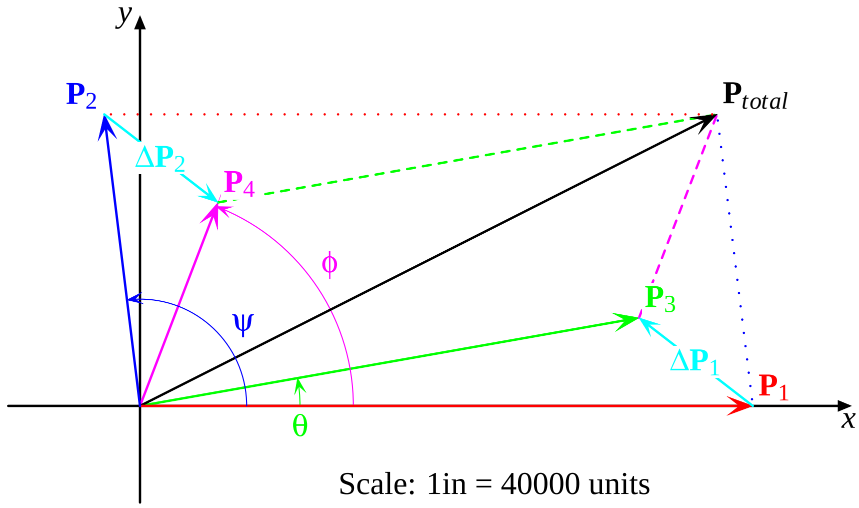


Figure 4.2 95% confidence interval: 29-43 mph.

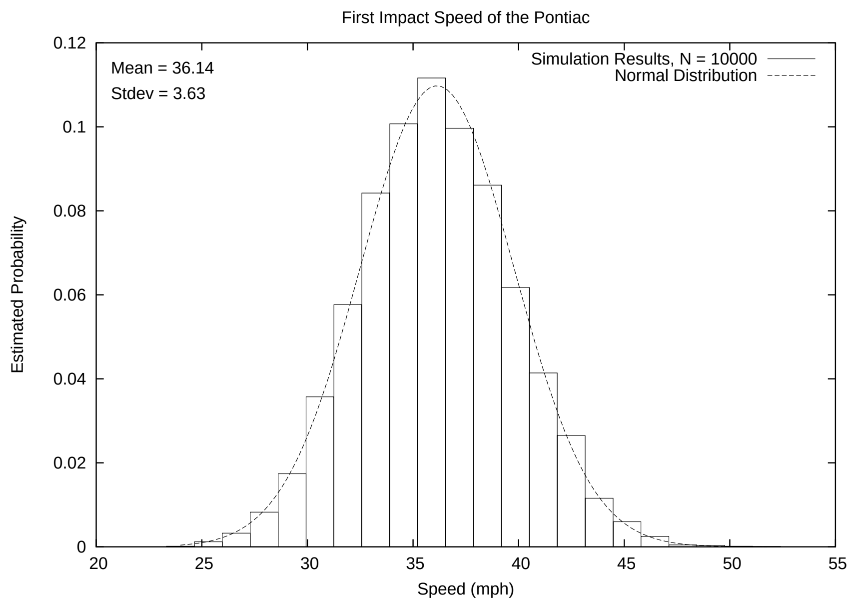


Figure 4.3 95% confidence interval: 31-45 mph.

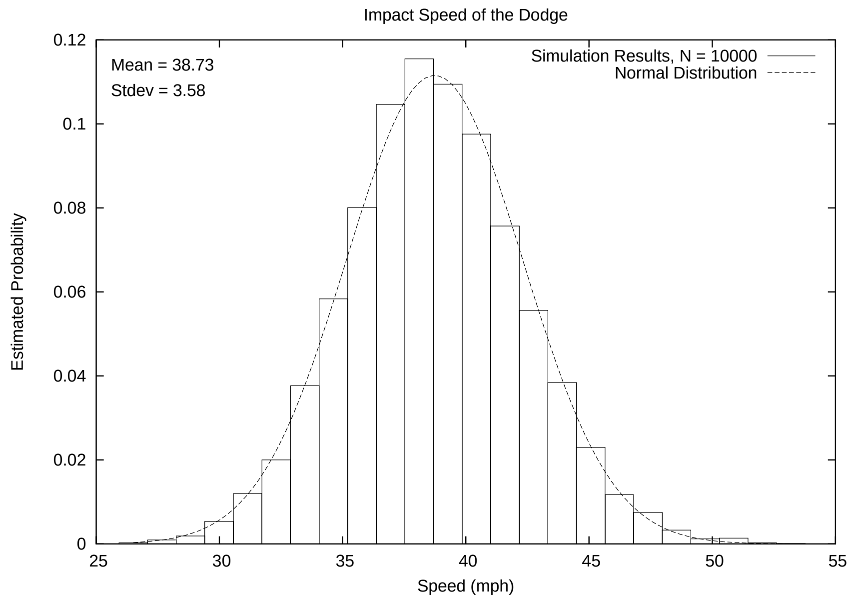
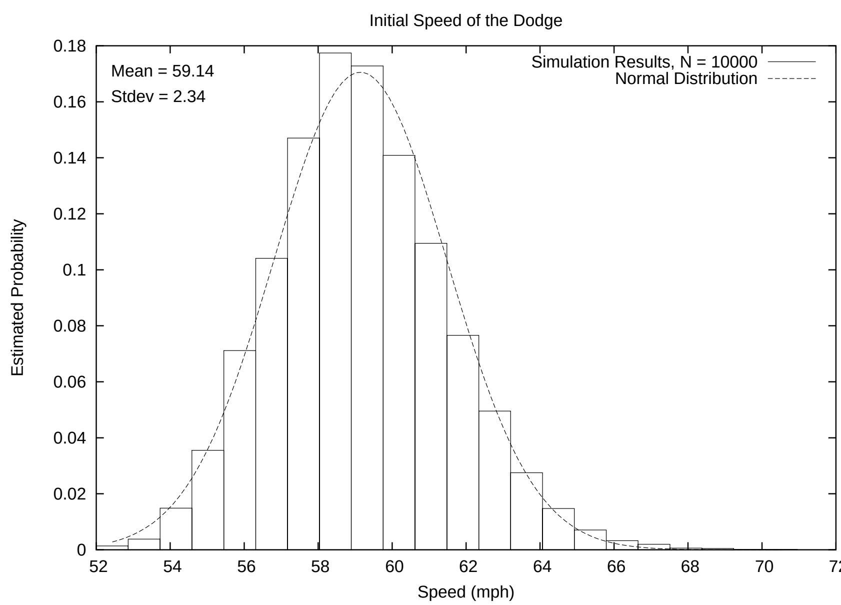


Figure 4.4 95% confidence interval: 54-63 mph.



4.2 Second Collision

The lowered Chevy pickup and the Pontiac collided second.

1. The impact speed of the Chevy was a nominal 29 mph.
2. The impact speed of the Pontiac was a nominal 28 mph.
3. The speed of the Chevy at the start of the skid was 52 to 54 mph.

Figure 4.5 The results of the linear momentum vector diagram based on the means give $\Delta v_1 = 12.69$ mph and $\Delta v_2 = 15.45$ mph. Vehicle #1 is the lowered Chevy truck and vehicle #2 is the Pontiac.

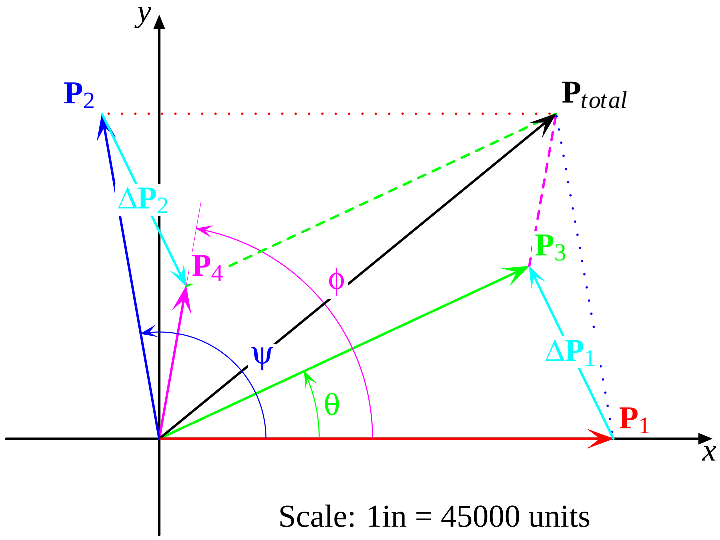


Figure 4.6 95% confidence interval: 22-34 mph.

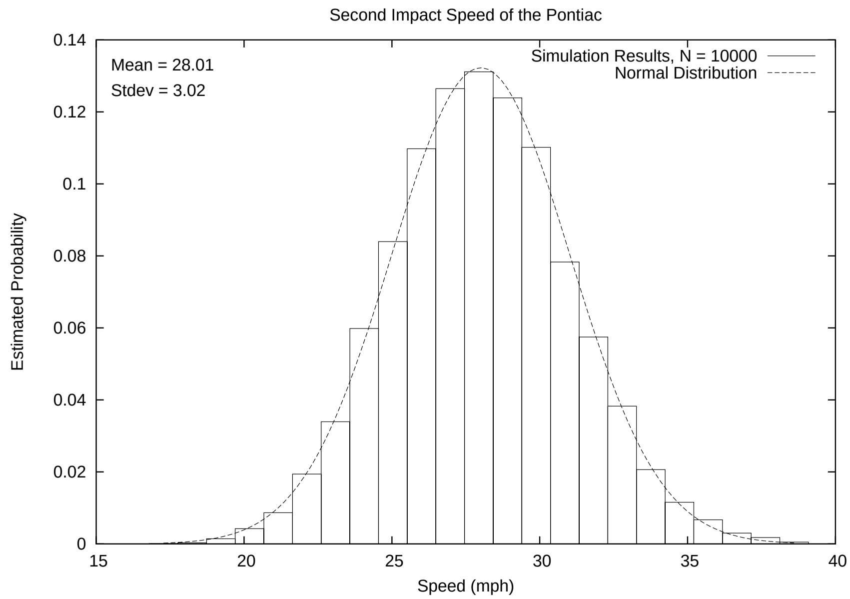
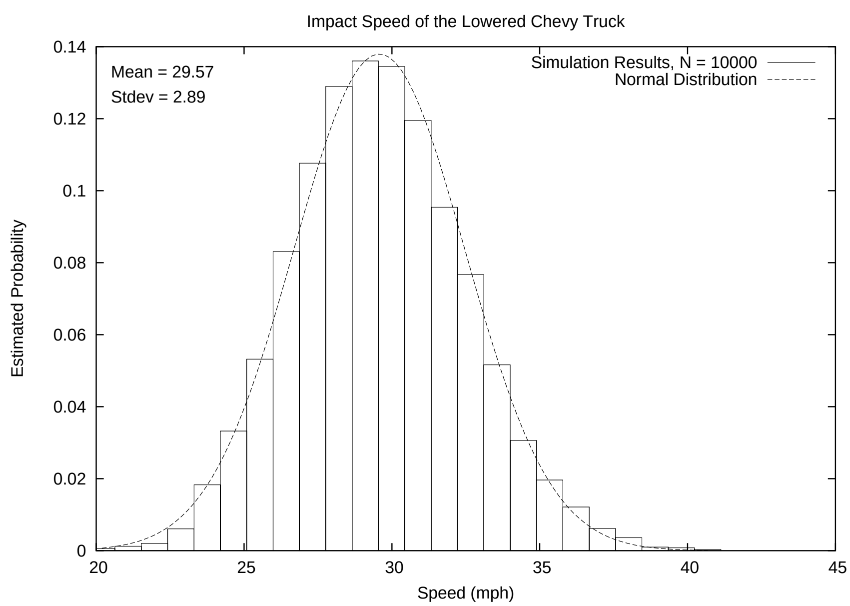


Figure 4.7 95% confidence interval: 23-35 mph. The speed of the lowered Chevy truck was around 52 mph when it began to skid.



4.3 Third Collision

The GMC pickup and the Pontiac collided third.

1. The front of the GMC hit driver A while he was on the road.
2. The impact speed of the GMC with the Pontiac was a nominal 38 mph.
3. The impact speed of the Pontiac was a nominal 12 mph.
4. There were no pre-impact skids from the GMC pickup.

Figure 4.8 The results of the linear momentum vector diagram based on the means give $\Delta v_1 = 15.41$ mph and $\Delta v_2 = 20.90$ mph. Vehicle #1 is the 4x4 truck and vehicle #2 is the Pontiac.

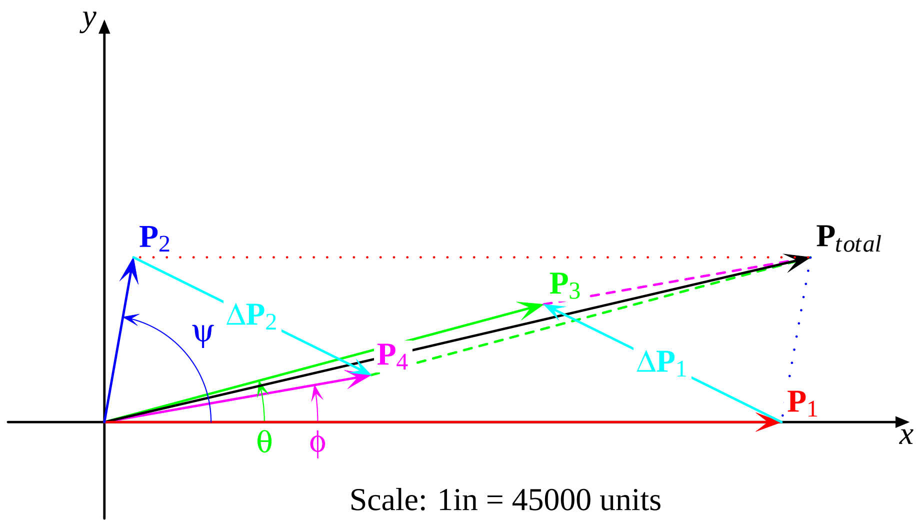


Figure 4.9 95% confidence interval: 6-19 mph.

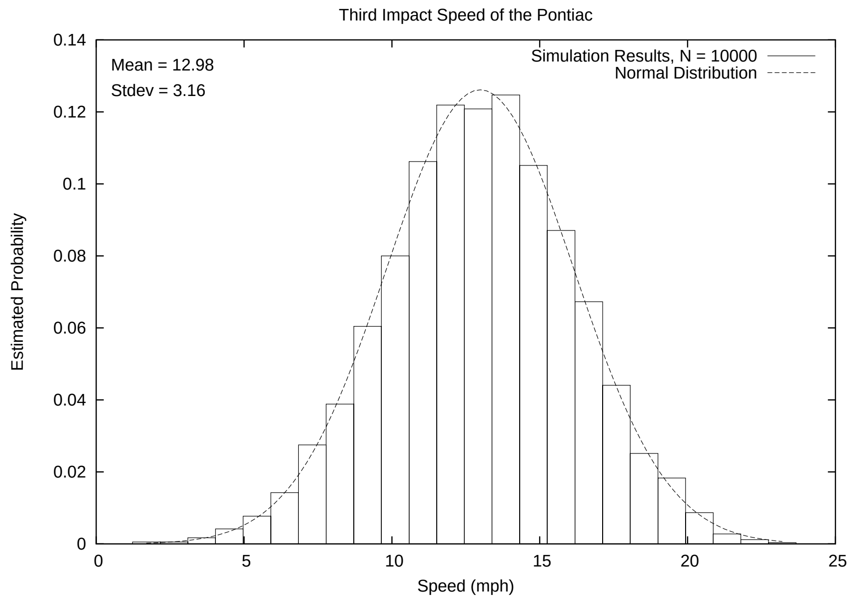
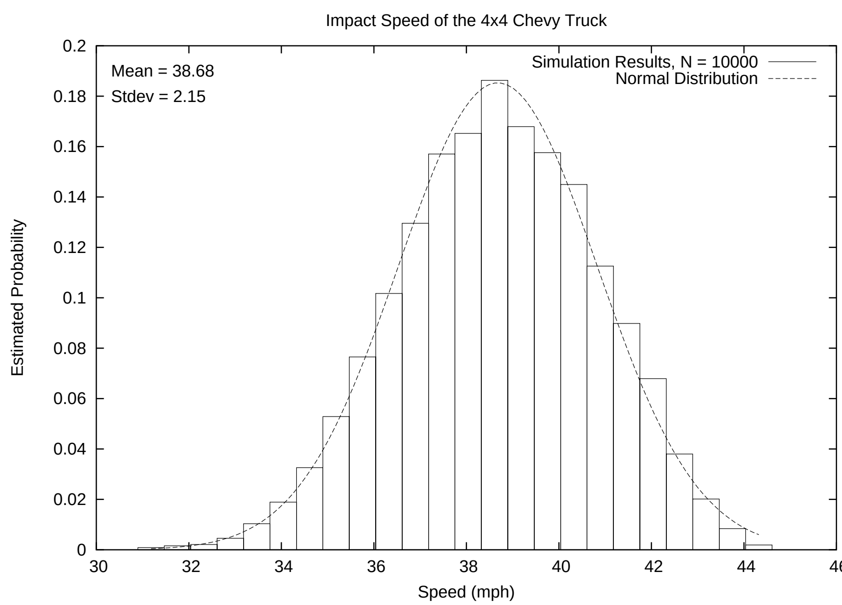


Figure 4.10 95% confidence interval: 34-42 mph.



4.4 Avoidance Analysis

The analysis of the impact speeds of the Dodge Dakota and the Pontiac Grand Am raise two questions:

1. If the Dodge was going 50 mph at the beginning of the driver's perception-response time, would the collision have still occurred?

2. Did the Pontiac stop for its stop sign?

We will address the second question first. The Pontiac had to move about 50 feet from the stop sign and attain a speed of 35 mph in that distance if it had stopped. This would require an acceleration factor of 0.81g. The published value for the acceleration capability of the Pontiac is 0.52g from 0-30 mph at sea level. Since this collision occurred at 7000 ft, the acceleration performance would be reduced. Thus, it was not possible for the Pontiac to have stopped for its stop sign.

Using the Monte Carlo method, a "stopping" criteria can be set up where whenever an instance of the calculated acceleration value is greater than the allowable acceleration then a flag is raised. If the computer keeps track of all the raised flags and divides this number by the total number of runs, then the probability of failure to stop is determined. In this case, every iteration produced a flag so the probability was 100% that the Pontiac did not stop. This technique can be useful in civil trials when a percentage of blame is to be determined. This concept is very similar to a probability of failure in a reliability problem.

The first question may be answered by looking at the time it took the Dodge to skid to its impact with the Pontiac in the actual crash situation and then adding to this time the perception-reaction time of 1.6 seconds. We may also use the skid information from the Dodge along with its beginning speed to compute its distance from impact when the driver's perception-reaction time started. We will use a perception-reaction time of 1.6 seconds for this time at constant velocity. We will then place the Dodge at the beginning of its perception-response distance from impact at a speed of 50 mph, and see if it will reach the impact area. If it does, then we will look at how much time this will take overall, and look at the time difference to see if the Pontiac would have moved out of the way.

- Distance the Dodge will move in 1.6 seconds at 58.65 mph: 137.56 feet

- Distance the Dodge will move in 1.6 seconds at 61.08 mph: 143.26 feet

Add to these distances the 97 foot skid:

- Total distance of the Dodge: 234.56 feet to 240.26 feet

The total stopping distance (including P/R time) for the Dodge at 50 mph:

- 236.32 feet if $\mu = 0.70$
- 221.44 feet if $\mu = 0.80$

Thus, even without doing a time analysis, it is clear a collision would have been avoided had the Dodge been going the speed limit. Just as clearly, had the Dodge been going straight down the road at even its actual speed, it would have skidded straight, rather than across the lanes at an angle, and there would have been no collision. It was also the duty of driver A (Pontiac) to obey the stop sign and to yield the right of way to any oncoming traffic. Had he done this, there would have been no collision.

Clearly, this wreck has several contributing factors. Had driver A obeyed the stop sign and waited for oncoming traffic, there would have been no initial traffic crash. Had driver B (Dodge) been driving the speed limit and if he had not turned to the left before braking, there would have been no traffic crash. The subsequent crashes would not have happened if the first crash had not occurred.

5 Conclusion

The Monte Carlo method is the de facto standard for dealing with complex problems containing uncertainty. We have taken a complex problem in the form of a multiple vehicle collision and addressed the issues of uncertainty using stochastic methods, namely the Monte Carlo method. The basic premise of Monte Carlo is to solve the governing equations (conservation of linear momentum for this case) many times using random inputs. The critical link in performing the analysis is making sure the random variables going into the Monte Carlo simulation are justifiable with physical evidence or previous experience. The results of each run within a simulation must be physically able to happen or else that run must be discarded. This paper also presents some basic concepts of descriptive statistics and probability that are necessary to use and understand the Monte Carlo method.

Contact Information

John Daily	Jeremy Daily
P.O. Box 2206	7485 Timber Hill Dr.
Jackson, WY 83001	Huber Heights, OH 45424
john@jhscientific.com	jeremy@jhscientific.com

Bibliography

[1] W. Bartlett *et al.*, “Evaluating the uncertianty in various measurement tasks common to accident reconstruction,” *SAE Technical Paper Series*, no. 2002-01-0546, 2002.

[2] L. Gonick and W. Smith, *The Cartoon Guide to Statistics*. New York: HarperCollins, 1993.

[3] K. S. Shanmugan and A. M. Breipohl, *Random Signals: Detection Estimation and Data Analysis*. New York: John Wiley & Sons, 1988.

[4] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimentors: An Introduction to Design, Data Analysis and Model Building*. New York: John Wiley & Sons, 1978.

[5] W. Bartlett, “Conducting monte Carlo analysis with spreadsheet programs,” *SAE Technical Paper Series*, no. 2003-01-0487, 2003.

[6] J. G. Daily, *Fundamentals of Traffic Accident Reconstruction*. Jacksonville: Institute of Police Technology and Management, 1988.

[7] J. Daily and N. Shigemura, *Fundamentals of Applied Physics for Traffic Accident Investigators*. Jacksonville: Institute of Police Technology and Management, 1997.

[8] G. Bigg and J. Daily, “Spinning trajectory drag factor validation tests,” in *Special Problems*, (Jacksonville, FL), IPTM, UNF, May 2003.

[9] D. P. Wood and S. O’Riordain, “Monte carlo simulation methods applied to accident reconstruction and avoidance analysis,” *SAE Technical Paper Series*, no. 940720, 1994.

[10] G. Kost and S. M. Werner, “Use of Monte Carlo simulation techniques in accident reconstruction,” *SAE Technical Paper Series*, no. 940719, 1994.