# Examples in Probability and Statistics using NFL data

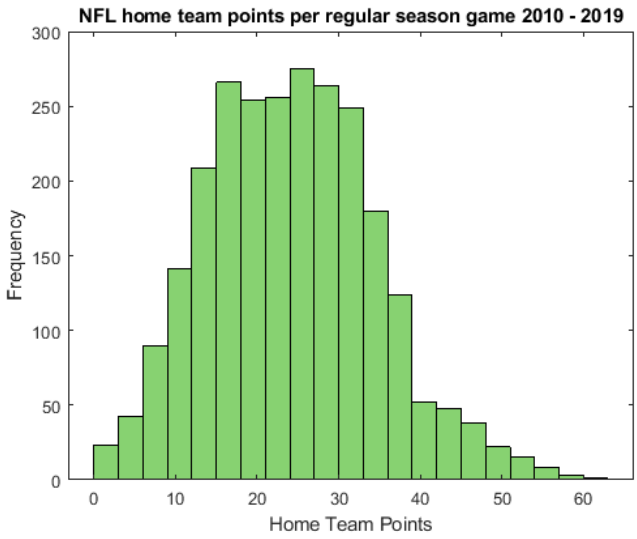
We will compute various statistics using custom Java code and MATLAB, and we will also use MATLAB to plot and visualize data and the results of certain calculations. The data set on which all the calculations are to be perfomed is a collection of data from every regular season NFL game from the ten seasons spanning 2010 - 2019. The total number of games played was 2560. A typical row in the data set looks like:

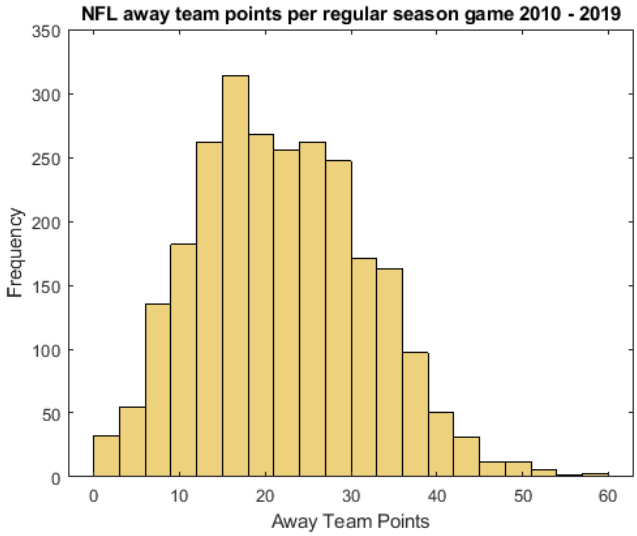
| Week | HomeTeam | AwayTeam | Total | H-RushAtt | H-RushYards | H-PassYards | H-Turnover | H-Score | A-RushAtt | A-RushYards | A-PassYards | A-Turnover | A-Score | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | NOR | MIN | 49.5 | 25 | 79 | 237 | 0 | 14 | 23 | 91 | 171 | 1 | 9 | -1 |

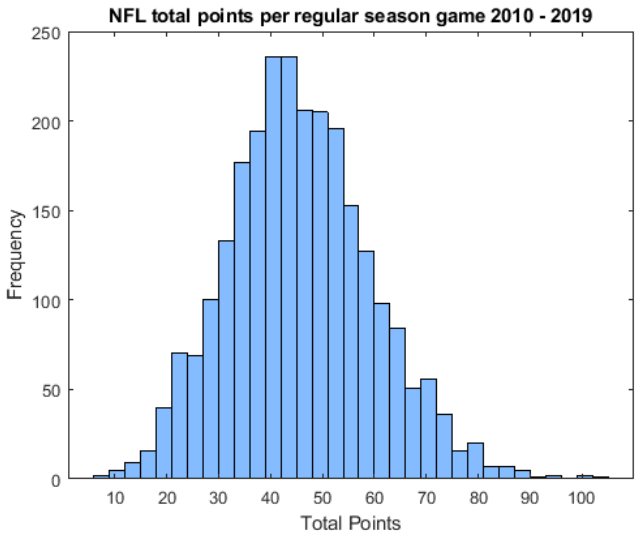
Let us begin by using MATLAB to create a histogram of the total points scored in each game, and also a histogram for the total points for both the home team and the away team. Since the data are stored in a particular manner, we will use Java to manipulate the data into a convenient form in order for us to have MATLAB process the data more easily. Therefore we first use Java to generate a text file that contains the home team points, the away team points, and the total points scored in each game, each in a separate column. We then have MATLAB read the file and store each column as a separate matrix. To produce the histograms, we use the code

```
data = readmatrix('gameTotals.txt');
home = data(:,1);
away = data(:,2);
total = data(:, 3);

figure
histogram(home, 'FaceColor', '#38B513')
title('NFL home team points per regular season game 2010 - 2019')
xlabel('Home Team Points')
ylabel('Frequency')

figure
histogram(away, 'FaceColor', '#E3B327')
title('NFL away team points per regular season game 2010 - 2019')
xlabel('Away Team Points')
ylabel('Frequency')

figure
histogram(total, 'FaceColor', '#3390FF')
title('NFL total points per regular season game 2010 - 2019')
xlabel('Total Points')
ylabel('Frequency')
```

which produces




We can also compute the mean, median, mode, variance and standard deviation using the MATLAB commands

```
mean median mode var std
```

The results are given in the following table.

| | Mean | Median | Mode | Variance | Standard Deviation |
|---|---|---|---|---|---|
| Home | 23.74 | 23 | 24 | 107.05 | 10.35 |
| Away | 21.55 | 21 | 17 | 97.66 | 9.88 |
| Total | 45.29 | 44 | 51 | 193.99 | 19.53 |

Let us now consider some calculations involving elementary probability. Suppose we want to know the number of games in which the home team had at least 100 rushing yards. Since the total number of games is finite, we can use the sample space method and a brute force approach to count the number of such games that meet our criterion. Using the method

```
public double atleast100RushingYards("home")
```

we find that the number of games is 1499, or about 58% of the games. Similarly, the number of games in which both the home team and the away team each had at least 100 rushing yards is 677, or about 26%.

The NFL's 32 teams are divided into two conferences, each conference contains four divisions, and each division contains four teams. Suppose that we are restructuring the NFL and we are given the task of assigning each team to a conference. In how many distinct ways can this be done? Note that we are partitioning a set of 32 teams into 8 groups of 4 teams each. Hence the answer is given by the multinomial coefficient

$$\binom{32}{\underbrace{4\,4\cdots4}_{8\text{ factors}}} = \frac{32!}{4!^8} = 2390461829733887910000000.$$

The same technique can be used to solve the following problem: in a given week, how many possible matchups are there (assuming that no team has a bye)? In this case the 32 teams are being partitioned into 16 groups each of size 2, so there are

$$\binom{32}{\underbrace{2\,2\cdots2}_{16\text{ factors}}} = \frac{32!}{2!^{16}} = 401505793661031387584256000000 0$$

possible matchups.

We consider next conditional probability. Previously we found that the probability of the home team having at least 100 rushing yards was about 58%. However, what is the probability that the home team had at least 100 rushing

yards *given that the home team had at least 20 rushing attempts*? Now the sample space has been reduced and consists solely of games in which the home team had at least 20 rushing attempts, and *of those* we count the number of games in which the home team had at least 100 rushing yards. The method

```
public double homeConditionalRushingYards()
```

calculates this probability to be about 68%.

Now let $A$ be the event that the home team scored at least 10 points and let $B$ be the event that the away team scored at least 10 points. Using the methods

```
public double team10Points("home")
public double team10Points("away")
```

we find that $P(A) \approx 92.7\%$ and that $P(B) \approx 89.7\%$. Thus $P(A) + P(B) > 1$. What are the minimum and maximum possible values for $P(A \cap B)$? In other words, what is the minimum probability that both teams scored at least 10 points, and what is the maximum probability that both teams scored at least 10 points? Since

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 182.4 - P(A \cap B) \leq 1,$$

we see that $P(A \cap B) \geq 82.4\%$, which establishes the lower bound. The maximum in general of $P(A \cap B) \leq \min\{P(A), P(B)\}$ and equality holds when one of $A \subset B$ or $B \subset A$ is true. In our case, we obtain $P(A \cap B) \leq 89.7\%$. We can verify our conclusion via explicit computation using the method

```
public double each10Points()
```

which gives the answer of $\approx 82.9\%$, consistent with our analysis.

Let us recalculate one of the probabilities just computed via the general rule which states that for any event $A$, we have $P(A) = 1 - P(A^c)$. If the probability is 92.7% that the home team scored at least 10 points, then we expect that the probability that the home team scored fewer than 10 points should be 1 - 92.7% = 7.3%. We compute this explicitly using the method

```
public double teamUnder10Points("home")
```

which outputs .073, as expected.

Now define the random variable $X$ as follows:

$$X = \begin{cases} 1, & \text{if } 150 \leq \text{away team's passing yards} \leq 200 \\ 0, & \text{otherwise.} \end{cases}$$

Then the *probability mass function*, PMF, of $X$, denoted by $p_X(x)$, gives the number of occurrences of each value in the range. In our case, $p_X(x)$ tells us the number of games in which the away team's passing yards was in the interval [150, 200] and how many were outside this interval. The method

```
public int passingYardsInterval(150, 200, "away")
```

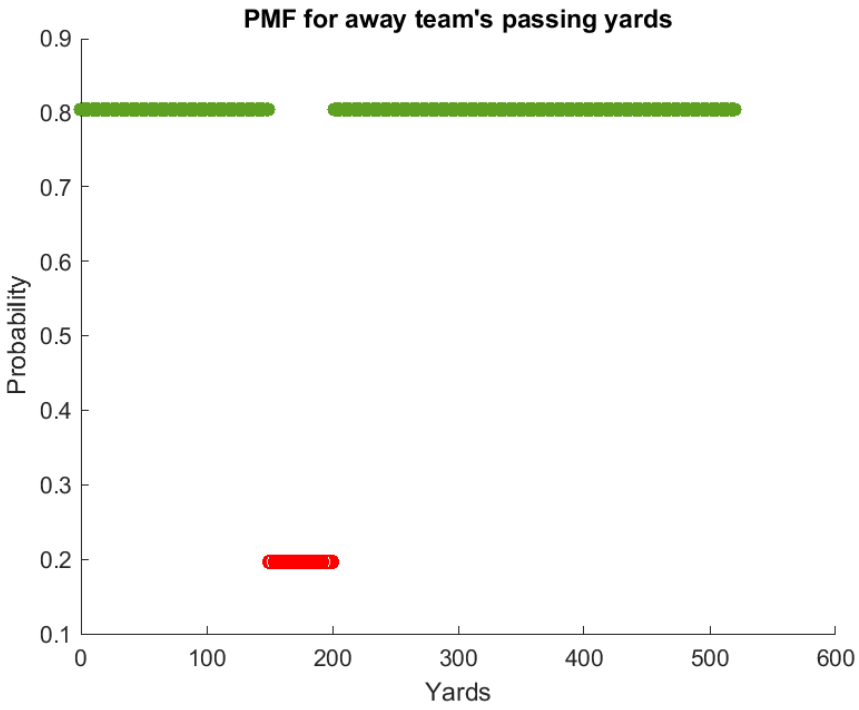
tells us that 505/2560 games, or $\approx 19.7\%$, meet this criterion. Hence

$$p_X(x) = \begin{cases} .197, & 150 \leq x \leq 200 \\ .803, & x \in [0, 150) \cup (200, \infty). \end{cases}$$

The suboptimal MATLAB code for the PMF $p_X(x)$ is

```
x1 = [0:1:149];
x2 = [150:1:200];
x3 = [201:1:520];

y1 = zeros(150) + .803;
y2 = zeros(51) + .197;
y3 = zeros(320) + .803;

figure
hold on
first = scatter(x1, y1,'filled','MarkerFaceColor', '#5EA022');
second = scatter(x2,y2, 'r');
third = scatter(x3,y3, 'filled', 'MarkerFaceColor', '#5EA022');
title('PMF for away team''s passing yards')
xlabel('Yards');
ylabel('Probability');
```

and produces the following figure:




♠ ♠ ♠ ♠ ♠

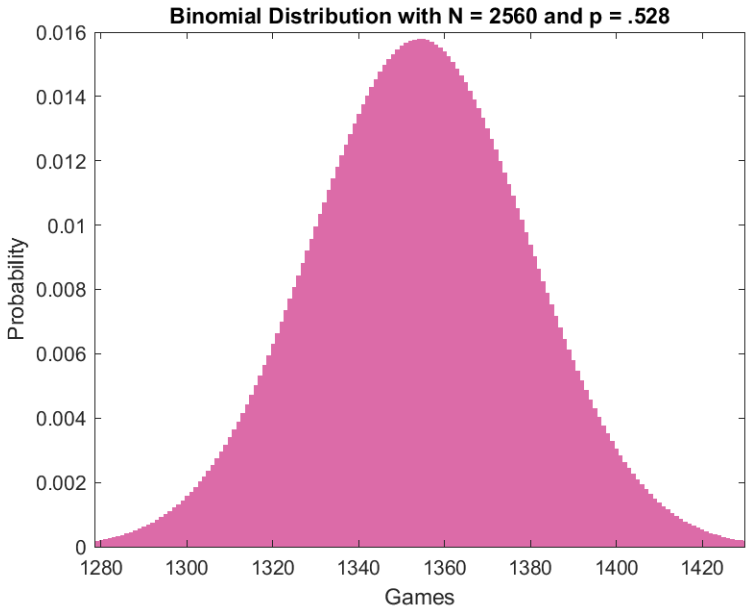
We consider next distributions of discrete random variables. We first count the number of games in which the total points scored was odd, and the number of games in which the total points scored was even using the method

```
public int[] oddEven()
```

and find that 1353 (52.9%) games had odd totals and 1207 (47.1%) had even totals. Since every game's total points is either even or odd, and assuming that the results of our data set are representative, we can model this using a binomial distribution with $N = 2560$, $p = .528$, and $\sigma = \sqrt{Np(1-p)} = 25.26$. The MATLAB code

```
x = linspace(0,2560,2561);
y = binopdf(x,2560,.528);
mean = 2560*.528;
stdev = sqrt(2560*.528*.471);

bar(x,y,1, 'FaceColor', '#DC6BA8')
xlim([mean - 3*stdev, mean + 3*stdev]);
title('Binomial Distribution with N = 2560 and p = .528');
xlabel('Games');
ylabel('Probability');
```

produces the following figure, which includes points $\pm 3\sigma$ about $\mu = 1354.24$:




To find the probability that exactly 1000 games chosen at random all had odd totals, we compute

$$\binom{2560}{1000}.528^{1000}.472^{1560}$$

which in MATLAB is

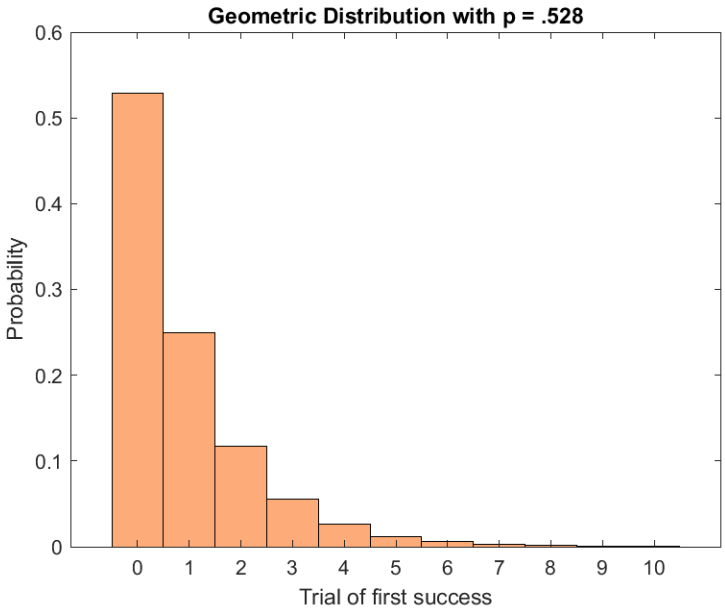
```
binopdf(1000, 2560, .528)
```

which is approximately $1.02 \times 10^{-44}$.

If we choose games at random from the data set, we can ask for the probability that first game with an odd total shows up on the $n^{\text{th}}$ choice. This is modeled by a geometric distribution with probability of success $p$ = .528, which can be requested from MATLAB with

```
x = 0:10;
z = geopdf(x,.528);
bar(x,z,1, 'FaceColor', '#FDAC79')
title('Geometric Distribution with p = .528')
xlabel('Trial of first success');
ylabel('Probability');
```

which produces




♦ ♦ ♦ ♦ ♦

We turn now to the hypergeometric distribution. Suppose we choose ten games at random from the data set, and we seek the probability that exactly five of them had totals that exceeded the projected total. We first count the number of games $r$ that meet this criterion using the method

```
public int gamesOver()
```

which produces $r$ = 1261.

The desired probability can be computed using the formula

$$\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}},$$

where $N$ is the number of games (2560), $n$ is the number of games drawn at random (10), $r$ is the number of games that exceeded the projected total (1261), and $y$ is the number of successes (5). In our case we have

$$\frac{\binom{1261}{5}\binom{1299}{5}}{\binom{2560}{10}},$$

which in MATLAB is

```
nchoosek(1261,5)*nchoosek(1299,5)/nchoosek(2560,10)
```

and is approximately 24.6%.

We can check our answer with MATLAB using

```
hygepdf(5, 2560, 1261, 10)
```

and indeed the output is identical.

♥ ♥ ♥ ♥ ♥

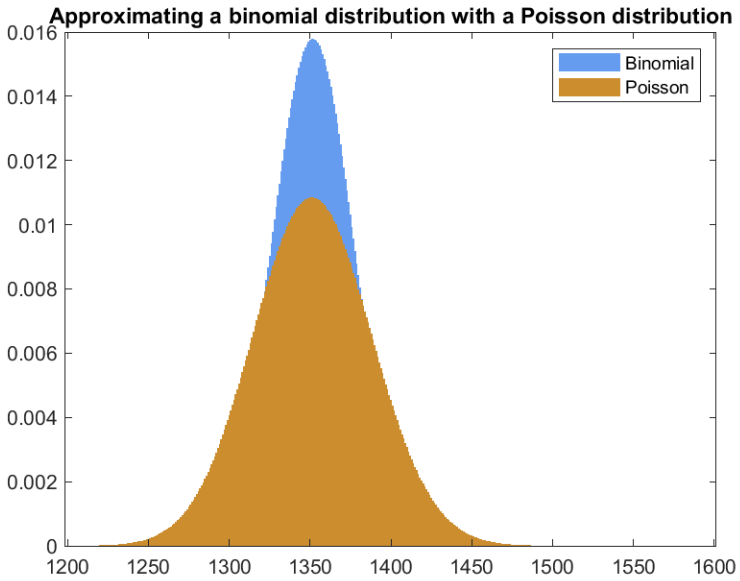
Let us now see how a Poisson distribution can closely approximate a binomial distribution under certain conditions. Recall that if a set of values has a Poisson distribution with parameter $\lambda$, then its PMF is given by

$$\frac{\lambda^k}{k!}e^{-\lambda},$$

for $\lambda > 0$ and $k$ a nonnegative integer. When $N$ is large and $p$ is small, the Poisson distribution with $\lambda = Np$ closely approximates a binomial distribution with $N$ trials and probability of success $p$ on each trial. We can compare the binomial distribution we encountered above, which had $N = 2560$ and $p = .528$, with a Poisson distribution that has parameter $\lambda = Np = 2560(.528)$. The following MATLAB code

```
x = 1200:1600;
y = binopdf(x, 2560, .528);
z = poisspdf(x, 2560*.528);

bar(x,y, 1,'FaceColor', '#669CEF');
hold on;
bar(x,z, 1, 'FaceColor', '#CB8D2E');
title('Approximating a binomial distribution with a Poisson distribution');
legend('Binomial', 'Poisson');
```

produces a visual rendering of the approximation:



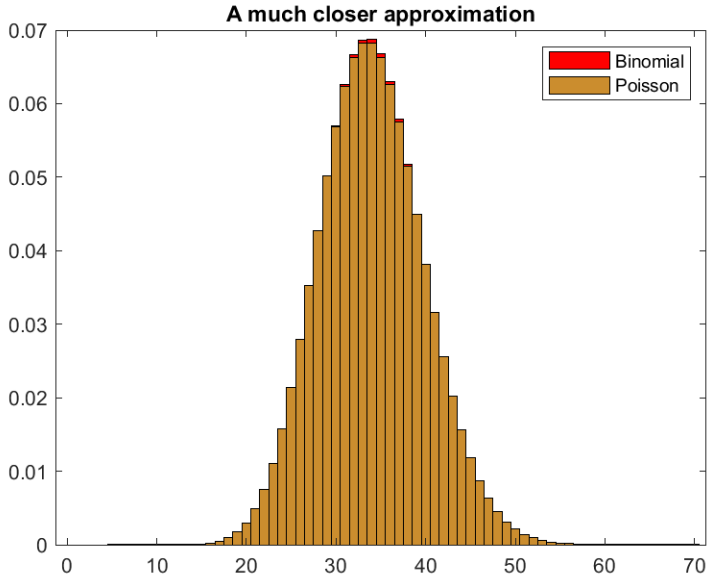

While this is a fairly close approximation, the Poisson distrubution does a better job of approximating a binomial distribution as $N$ gets larger and $p$ gets smaller. To demonstrate this, we might ask the following question: how many games in the data set had total scores that were exactly equal to the projected total? Such games are indicated by a 0 in the last column of any row in the data set. We first count the number of such occurrences using the method

```
exactlyProjectedTotal()
```

which gives a total of 34 games out of the 2560, or $\approx .013$. Thus we can approximate a binomial distribution with $N = 2560$ and $p = .013$ with a Poisson distribution that has $\lambda = Np = 34$. Using the MATLAB code

```
x = 0:70;
y = binopdf(x, 2560, 34/2560);
z = poisspdf(x, 34);

bar(x,y,1,'FaceColor', 'r');
hold on;
bar(x,z,1, 'FaceColor', '#CB8D2E');
title('A much closer approximation')
legend('Binomial', 'Poisson');
```

we obtain the Poisson distribution superimposed on the binomial distribution, and it overlaps almost perfectly:

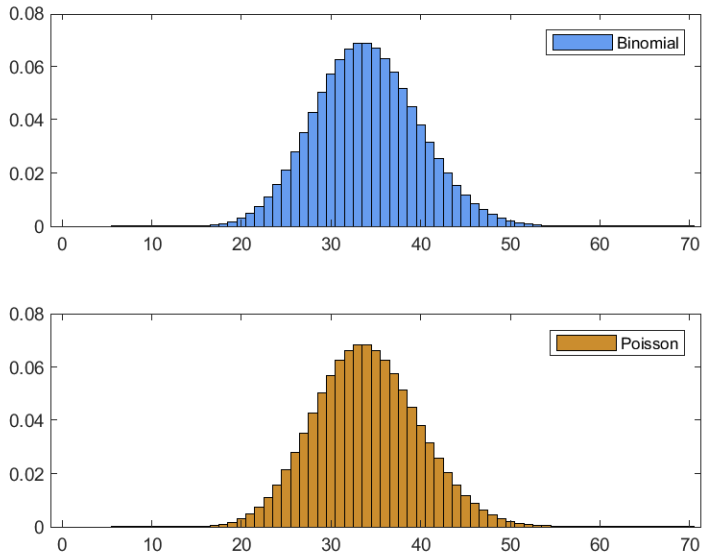



For an alternate visualization, we use the MATLAB code

```
x = 0:70;
y = binopdf(x, 2560, 34/2560);
z = poisspdf(x, 34);

tiledlayout(2,1);

nexttile;
bar(x,y,1,'FaceColor', '#669CEF');
```

```
legend('Binomial');

nexttile;
bar(x,z,1, 'FaceColor', '#CB8D2E');
legend('Poisson');
```

which produces the figures




which look almost identical.

Suppose now that we are interested in the probability of the away team having fewer than $n$ turnovers in a game, where $n$ is a nonnegative integer. We first count the number of games $k_n$ that have $n$ away team turnovers, for each $n = 0, 1, \ldots, n_{\max}$, where $n_{\max}$ is the maximum number of away team turnovers in any game in the data set. We write the data to a file using the method

```
public void turnoverDistributionFunction()
```

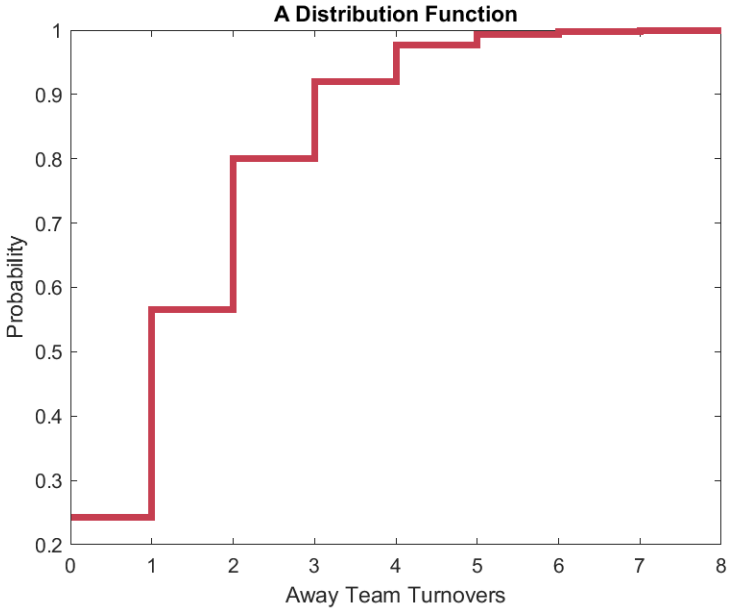
for processing by MATLAB. The distribution function is

$$F(x) = \begin{cases} 0, & x < 0 \\ .2434, & 0 \le x < 1 \\ .5652, & 1 \le x < 2 \\ .8004, & 2 \le x < 3 \\ .9207, & 3 \le x < 4 \\ .9770, & 4 \le x < 5 \\ .9930, & 5 \le x < 6 \\ .9977, & 6 \le x < 7 \\ .9988, & 7 \le x < 8 \\ .1, & 8 \le x < \infty \end{cases}$$

Because this is slightly difficult to visualize, we use the MATLAB code

```
mapValues = readmatrix('turnoverDistribution.txt');
cumulative = [];
x = mapValues(:,1);
sum = 0;

for i = 1:size(mapValues(:,2))
    cumulative = [cumulative, mapValues(i,2) + sum];
    sum = cumulative(i);
end

cumulative = cumulative/2560;
stairs(x, cumulative, 'LineWidth',3, 'Color','#C63E50')
title('A Distribution Function')
xlabel('Away Team Turnovers')
ylabel('Probability')
```

which produces


A Distribution Function
Probability
Away Team Turnovers
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
1
2
3
4
5
6
7
8

---