Using Soccer Goals to Motivate the Poisson Process

Singfat CHU
NUS Business School
National University of Singapore
1 Business Link
SINGAPORE 117592

bizchucl@nus.edu.sg

ABSTRACT

In introducing analytic queuing models, textbooks rarely justify the assumption of the exponential and the Poisson distributions. This paper demonstrates how a real-life phenomenon familiar to many students, namely the occurrences of soccer goals, can drive the ideas. An Excel worksheet is made available for further analysis.

Editor's note: This is a pdf copy of an html document which resides at
http://ite.pubs.informs.org/Vol3No2/Chu/

1 INTRODUCTION

A standard topic in OR and OM courses is the management of queues. Analytics developed for tractable queuing models assume random arrivals according to a Poisson process with constant rate λ. Equivalently, inter-arrival times are said to be independent and to follow an exponential distribution with mean 1 / λ.The appropriateness of the exponential and Poisson distributions, their linkage and their properties which lead to simple analytics, often escape our students as textbooks rarely provide empirical evidence to justify them.

Recently, Grossman (1999) expounded on the utility of process-driven spreadsheet queuing simulation to help students experience queue dynamics. Inter-arrival times and service times (Figure 2, in the paper) are generated according to parameters specified by the user. To appreciate the practicality of the simulation exercise, the user must therefore have a feel for the relevance of the exponential distribution used to generate the inter-arrival times. This simulation approach, coupled with some analytical insights, is undoubtedly fruitful to the pedagogy.

Experience has shown that analytical concepts are best driven by real-life illustrations. An early example of this pedagogical paradigm is Lafleur et al. (1972) who report a student project which modeled the number of individuals encountered during the approximately 30 seconds required to walk through a dormitory entrance as a Poisson distribution. Another example is Schmuland (2001), who uses the Poisson model to explain the phenomena of bursts in shark attacks and the scoring patterns of ice hockey legend Wayne Gretzky.

Based on his observation that the Poisson distribution provides a good fit for goals scored in ice hockey games, Berry (2000) assumes an exponential distribution for the times between goals to estimate the strategic time to "pull the goalie" when a team is down in a game. This problem was recently re-visited by Zaman (2002) from a Markov Chain angle.

The link between the Poisson and exponential distributions can only be illustrated empirically if data on the sequential timing of the random occurrences are available. Call centers, for example, often collect statistics on the number of customer calls received and not necessarily the times of individual calls. The same situation may prevail at toll booths where only the rates of car arrivals may be recorded. Likewise, Larsen and Marx (1981) have an interesting case study on the incidence of war outbreaks during the period 1500-1931. But they also do not provide data on when each war started.

This paper makes available a dataset on the timing of a competitive phenomenon, namely the scoring of soccer goals in a sequence of 232 games. Soccer, with its global following, provides an excellent channel to sell the Poisson and the exponential distributions to our students. The paper will demonstrate their relevance, their inter-relationship, their properties and their implication towards a better understanding of the soccer game. The occurrences of touchdowns in American football or home runs in baseball in sequential games played by a particular team or player can also lend themselves to a similar analysis.

  ISSN: 1532-0545

2 THE WORLD CUP DATASET

The World Cup tournament pitching the best national teams is the ultimate in the soccer world. It is held every four years. South Korea and Japan co-hosted the latest tournament between 31 May and 30 June 2002. That tournament involved 32 countries, which were initially divided into 8 groups of four. In this first stage, each country played against each of its 3 group peers. The top 2 countries within each group then advanced to a knockout second stage. Ultimately, Brazil beat Germany 2-0 in the 64^{th} and final game.

In 1998, the same tournament format was used in France. The host country was ultimately crowned as the champion. However, when Italy and USA respectively hosted the World Cup in 1990 and 1994, only 24 countries were showcased. In these two gatherings, the respective winners, Germany and Brazil, emerged after 52 games.

Data on all the goal occurrences in these four World Cup tournaments are available at Fifa's World Cup website. Previous tournaments are ignored as the website does not provide information on the sequence of games. As such, they would not help in the illustration of the exponential distribution.

This paper therefore focuses on the 232 games played in the 1990-2002 World Cup tournaments. Only the goals scored in the 90 minutes regulation time are considered. This leaves out goals scored in extra time or in penalty shoot-outs. A regular soccer game consists of two halves scheduled for 45 minutes. However, injury time is often added at the end of each half to compensate for game stoppages arising from player injuries. The extent of injury time in each game is unfortunately not available. For consistency in the analysis, a goal scored in injury time, say at the 92^{nd} minute, is recorded as occurring at the 90^{th} minute. This is because until 1998, goals scored in added times were always recorded at either the 45^{th} or 90^{th} minute. This may have some effect on the fit of the Poisson and exponential distributions to the data.

The first game in Italy90 saw Cameroon scoring a single goal against Argentina at the 67^{th} minute. In the second game, Romania scored 2 goals against the then Soviet Union at the 42^{nd} and 57^{th} minutes. The time to these goals are respectively 65 $(= (90-67)+42)$ and 15 $(= 57-42)$ minutes. Proceeding in the same way, 574 inter-goal times were obtained by the end of the 2002 Final game (game 232). Figure 1 illustrates the computation of the time between goals.

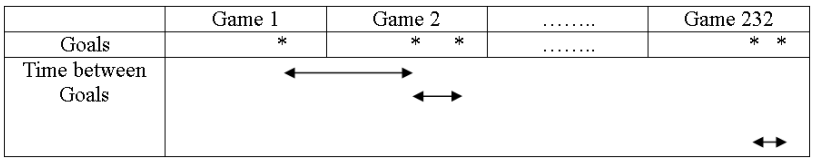

Figure 1: Illustration of time between goals

3 MOTIVATING THE EXPONENTIAL DISTRIBUTION

3.1 Mean and standard deviation coincide

The suitability of an exponential fit to the 574 inter-goal time intervals can be demonstrated in many ways. For example, a special property of the exponential distribution is that its mean equals its standard deviation. These two statistics were found to be respectively 36.25 and 36.68 minutes. These compare favorably against their common theoretical expectation, namely $90/\lambda$ or 36.31 minutes where $\lambda = 575/232$ is the mean number of goals scored per 90-minutes regulation game.

3.2 Chi-square test for the exponential fit

A more thorough validation is to compare the empirical distributions of the times between goals against an exponential distribution with the estimated mean. For example, the theoretical probability of observing an inter-goal time interval between 0 and 10 minutes is 0.2407 $(= 1 - exp(\lambda * 10/90))$. This implies that among 574 inter-goal intervals, we would expect about 138 $(= 574 * .2407)$ to lie between 0 and 10 minutes.

  ISSN: 1532-0545

Inter-goal Duration (minutes)	Actual	Empirical Probability	Theoretical Probability	Expected
0-10	144	0.2504	0.2407	138
10-20	106	0.1843	0.1828	105
20-30	86	0.1496	0.1388	80
30-40	52	0.0904	0.1054	60
40-50	46	0.0800	0.0800	46
50-60	27	0.0470	0.0607	35
60-70	35	0.0626	0.0461	26
70-80	16	0.0278	0.0350	20
80-90	22	0.0383	0.0266	15
90-100	12	0.0209	0.0202	12
100-110	3	0.0052	0.0153	9
110-120	3	0.0052	0.0116	7
120-130	6	0.0104	0.0088	5
130 or more	16	0.0278	0.0279	16
Total	574	1	1	574

Table 1: Time between goals

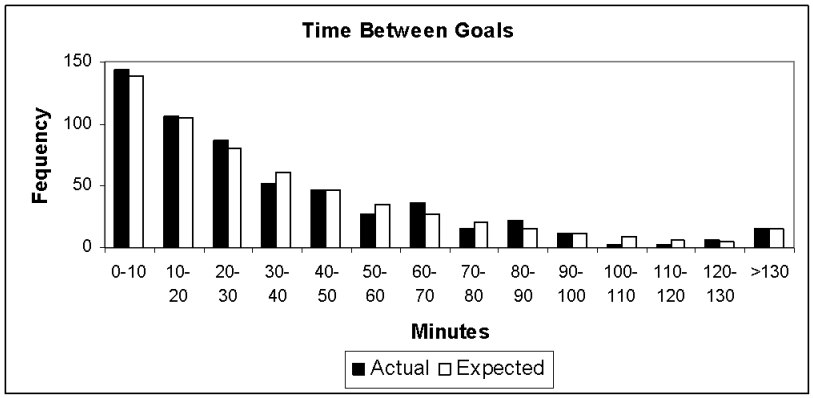

Figure 2: Time between goals

The actual number of intervals was 144. Table 1 reports similar calculations for other intervals while Figure 2 provides a graphical comparison.

The exponential fit to the actual distribution of time between goals can be assessed by a chi-square test. This returns a p-value of 0.2008 thereby suggesting that the suitability of the exponential fit to the data can not be rejected. This interfacing with Statistics is an effective means to illustrate the link between different disciplines to the students.

3.3 Memoryless Property

The dataset also allows for a demonstration that the exponential distribution lacks memory. Mathematically, it is written as $P(T > t + s \text{ — } T > t) = P(T > s)$, where T is the time between consecutive goals. This implies that if the last goal occurred t minutes ago, then the chance of observing the next goal beyond the next s minutes only depends on s and bears no relationship to t. In other words, the relative chance is a function of the length of the time interval s irrespective of the relative location of the origin t.

  ISSN: 1532-0545

Using Table 1 we find, for example, that empirically P(T $>$ 10) = 1-144/574 = 0.7491 while P(T $>$ 20 — T $>$ 10) = (574 - 144 - 106) / (574 - 144) = 0.7534 and likewise, P(T $>$ 30 — T $>$ 20) = 0.7346 etc. The empirical probabilities are again close to their theoretical value of 0.7593, assuming that $\lambda = 575/232$.

Alternatively, a soccer-centric approach to motivate the memoryless property of the exponential distribution is to ask the following, "Given that the last goal was scored 20 minutes ago, what is the chance of observing a goal within the next 10 minutes?" The above calculations indicate that this probability is about 0.25.

More examples can be developed from Table 1 to illustrate that no matter where a new origin is positioned, the exponential distribution regenerates itself as if the origin is at time 0. Intuitively, this implies that the relative variability or the coefficient of variation in both the conditional and unconditional distributions must be the same. The exponential distribution satisfies this requirement since its mean and standard deviation coincide in theory.

3.4 Independence of inter-goal times

Finally, another assumption in queuing theory is that inter-arrival times are independent. To justify this, we show that the autocorrelation function of the times between goals does not depart significantly from 0. This is indeed the case as none of the autocorrelations listed in Table 2 lie outside a band of plus or minus twice the standard error i.e. $2/\sqrt{574}$ or 0.0835.

The first-order or serial correlation can be visualized via a scatter plot of consecutive inter-goal times, as depicted in Figure 3.

Lag	1	2	3	4	5	6	7	8	9
Autocorrelation	-0.0112	0.0199	-0.0564	0.0063	0.0058	0.0065	-0.0301	0.0249	-0.0387

Table 2: Autocorrelations for time between goals

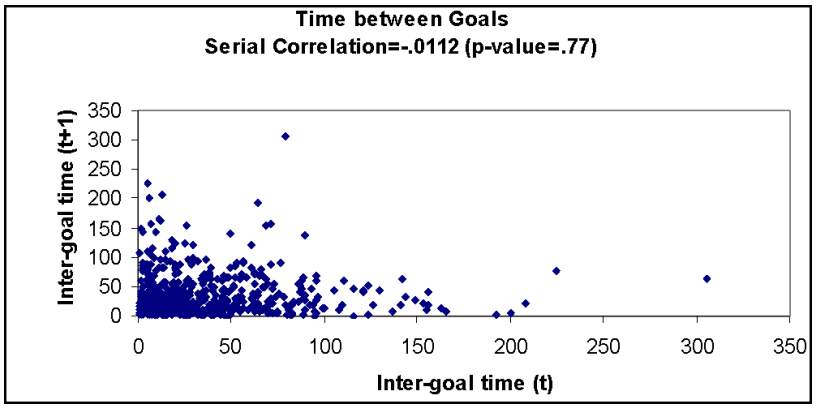

Figure 3: Serial correlation of time between goals

4 ILLUSTRATING THE POISSON DISTRIBUTION

4.1 Mean and Variance coincide

If the times between goals are exponentially distributed, then the number of goals within a fixed period should follow a Poisson distribution. The estimated rate λ is 575/232 or about 2.4784 goals per 90 minutes regulation game. The variance of the number of goals per game is found to be 2.4584. This is in close conformance to the theoretical result that the mean and the variance of a Poisson distribution coincide. The

  ISSN: 1532-0545

fit of the Poisson distribution can be further assessed by benchmarking a theoretical frequency distribution of goals per game against the actual frequencies.

4.2 Chi-square test for Poisson Fit

The expected number of games with x goals is obtained by multiplying the Poisson probability of x goals by 232, the total number of matches. The results are displayed in Table 3. A chi-square test fails to reject the validity of the Poisson fit (p-value=0.9745).

A visual contrast of the actual distribution of goals against the Poisson fit is provided in Figure 4.

#Goals	Actual # Games	Empirical Probability	Theoretical Probability	Expected # Games
0	19	0.0819	0.0839	19
1	49	0.2112	0.2079	48
2	60	0.2586	0.2576	60
3	47	0.2026	0.2128	49
4	32	0.1379	0.1319	31
5	18	0.0776	0.0654	15
6 or more	7	0.0302	0.0406	9
Total	232	1	1	232

Table 3: Actual and theoretical Distribution of goals over 90 minute intervals

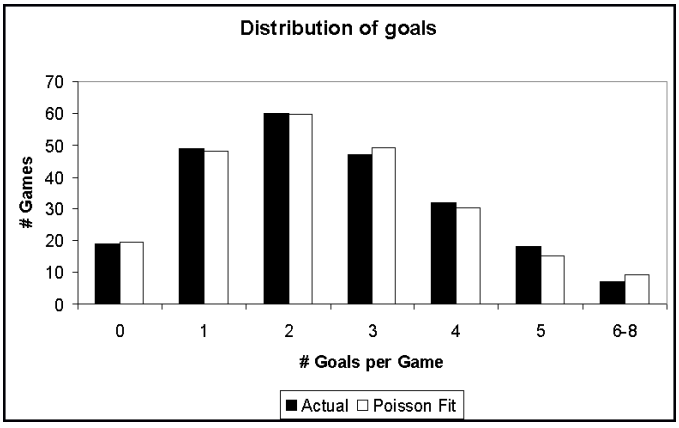

Figure 4: Actual and expected distribution of goals

  ISSN: 1532-0545

4.3 Is the Poisson rate constant?

One property of the Poisson process is that the number of events within alternative time intervals is also Poisson but with a proportionally adjusted mean. The validity of this property can be ascertained in Table 4 which contrasts the actual and expected distributions of goals in different time intervals fixed at 45, 15, 10, 5 and 1 minutes. Generally, the discrepancies between the actual and expected frequencies are statistically insignificant. The Poisson property of coincident mean and variance is also evident for the different time intervals.

	Time Intervals (mins)									
	45		15		10		5		1	
#Goals	Actual	Expected	Actual	Expected	Actual	Expected	Actual	Expected	Actual	Expected
0	141	134	910	921	1576	1585	3629	3639	20305	20313
1	157	167	397	380	453	437	520	501	575	567
2	100	103	77	79	55	60	26	34		
3	48	43	8	11	4	6	1	2		
4	16	13								
5	2	3								
Total	464		1392		2088		4176		20880	
Mean	1.2392	1.2392	0.4131	0.4131	0.2754	0.2754	0.1377	0.1377	0.0275	0.0275
Variance	1.2580	1.2392	0.3878	0.4131	0.2639	0.2754	0.1327	0.1377	0.0267	0.0275
p-value for chi-square test of fit	0.88		0.66		0.60		0.22		0.74	

Table 4: Actual and theoretical Distribution of goals over alternative time intervals. p-values obtained after consolidating classes such that their expected count is 5 or more.

Tables 3 and 4 are useful for contrasting empirical and theoretical probabilities pertaining to the number of goals. For instance, one can read off the empirical probability of 2 goals within 10 minutes as 55/2088 or 0.0263. The probability under an ideal Poisson process would have been 0.0288.

Soccer fans may question the constancy of the rate of goals production as follows: If a goal has just been scored, does this alter the rate of goals scored in the remainder of the game? The empirical evidence points interestingly to NO. In other words, the rate of goal production stays constant. Out of the 232 games, there were 213 with at least one goal scored. The mean time to the first goal was 33.71 minutes. This translates to a mean production of 2.7 goals over the 90 minutes duration of a game. Following the first goal, a total of 362 goals were scored i.e. at a mean production of 2.7173 goals over 90 minutes. The mean rates of goal production before and after the first goal do not appear to be different. In layman term, the result suggests that after a goal is scored, changes in strategy by the teams are effectively neutralized. As a result, the total rate of goal scoring stays constant. Students could be asked to investigate whether this result still holds after say a 1-1 or a 2-0 score.

4.4 Axioms of the Poisson Process

Table 4 also lends itself to the illustration of the axioms of the Poisson process, namely

1. $P\{1 \text{ goal in interval } (t, t+\Delta t)\} = \lambda * \Delta t + o(\Delta t)$
2. $P\{0 \text{ goal in interval } (t, t+\Delta t)\} = 1 - \lambda * \Delta t + o(\Delta t)$

  ISSN: 1532-0545

3. The numbers of goals in non-overlapping time periods are independent.

As the goal occurrences are recorded in minutes, the minimal time period that can be studied is 1 minute. From Table 4, the empirical probability of a goal in a 1 minute time interval is 575/22880, which is the product of λ =575/232 and Δt=1/90. This compares favorably with theoretical probability of 0.02716. As there was no 1-minute interval with 2 or more goals scored, both the first and second axioms are verified. The autocorrelation function in Table 5 supports the third axiom of independence in the number of goals in these 22880 non-overlapping 1-minute time intervals.

Lag	1	2	3	4	5	6	7	8	9
Autocorrelation	-0.0122	-0.0086	0.0021	-0.0033	0.0128	-0.0051	0.0003	0.0092	0.0033

Table 5: Autocorrelation for number of goals in 1-minute time intervals (margin of error=0.0138)

5 CONCLUSION

This contribution demonstrates that the Poisson and exponential distributions can be convincingly illustrated using an application of interest to many soccer fans worldwide. Students have an easier time grasping the concepts because they can see how they arise in a domain for which they may have a deep passion. As a bonus, the findings provide them with some possibly non-intuitive insights into the game of soccer. It is a win-win for both educators and students. They are invited to probe the dataset to investigate further problems of interest to them.

ACKNOWLEDGEMENT

I thank the two referees whose helpful feedback considerably improved the paper.

REFERENCES

Berry S. (2000), "A Statistician Reads the Sports Pages: My Triple Crown," *Chance*, Vol. 13, No. 3, pp. 56-61.

Grossman T.A. (1999), "Spreadsheet Modeling and Simulation Improves Understanding of Queues," *Interfaces*, Vol. 29, No. 3, pp. 88-103.

Lafleur M.S., P.F. Henrichsen, P.C. Landry, and R.B. Moore (1972), "The Poisson Distribution: An Experimental Approach to Teaching Statistics," *Physics Teacher*, Vol. 10, pp. 314-321.

Larsen R.J., and M.L. Marx (1981), *An Introduction to Mathematical Statistics and its Applications*, Prentice Hall, p. 148.(Englewood Cliffs)

Schmuland B. (2001), "Shark Attacks and the Poisson Approximation," (http://www.pims.math.ca/pi/issue4/page12-14.pdf)

Zaman Z. (2002), "Coach Markov Pulls Goalie Poisson," *Chance*, Vol. 14, No. 2, pp. 31-35.

  ISSN: 1532-0545