Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

IN SEARCH OF THE MEMORYLESS PROPERTY

Timothy S. Vaughan

Department of Management and Marketing
University of Wisconsin – Eau Claire
Eau Claire, WI 54702-4004, USA

ABSTRACT

This paper describes a distribution fitting exercise that has been used in an undergraduate introductory simulation class. The intent is for students to collect data describing a customer arrival process, with the goal of determining whether the exponential distribution is a good fit. The paper briefly reviews the memoryless property, and presents some teaching tips directed toward students' understanding of the concept. The data collection exercise is then presented, with attention to common pitfalls students have encountered in the past. A presentation of the variety of distributions that emerge as "best fitting", when samples are actually drawn from an exponential distribution, serves as a warning against over-reliance on goodness of fit measures.

1 INTRODUCTION

Analysis of input distributions is an important component of the introductory simulation course. The availability of graphical distribution fitting software enables students to quickly test the fit of a number of alternative distributions, using a visual comparison of the fitted distribution against a histogram of the data, supplemented by examination of alternative goodness of fit measures.

When available, a prior understanding of which distribution is "logically correct" is immensely valuable. For example, theory tells us the exponential distribution is logically consistent for time between arrivals in certain customer arrival processes. This paper presents a data collection and distribution fitting exercise that provides students an opportunity to confirm this theory. The exercise also provides an excellent caution regarding over-reliance on goodness of fit measures as the sole consideration in distribution fitting.

2 THE MEMORYLESS PROPERTY

It is well known that the exponential distribution is the only continuous distribution that has the "memoryless property", and that in many cases customer inter-arrival times should display this property. While the memory-less property is easy to define, it is also somewhat counter-intuitive at first glance. As a result, students have trouble coming to an understanding of when the memory-less property (and hence the exponential distribution) should and should not be expected.

Within a customer arrival scenario, the memoryless property can be stated as, "at any point in time, the time until the next customer arrival does not depend on how much time has passed since the last customer arrival". Assessment has shown that students will describe this as, "successive inter-arrival times are mutually independent". This statement sounds similar, but describes a weaker property that could be realized under any inter-arrival time distribution.

2.1 Reinforcing the memoryless property

I have found that working through a progression of the geometric distribution approaching the shape of the exponential, (and corresponding progression of binomial distribution to the poisson) within the context of inter-arrival times provides compelling logic Suppose, for example, that a one-hour period is divided into 60 one-minute intervals, where each one-minute interval will receive zero or one arrivals. Assume further that each one-minute interval represents an independent trial. Using a probability of 1/6 that an interval sees one arrival, and 5/6 that the interval sees no arrivals, provides the analogy of each minute being an independent "roll of the die" as to whether the interval sees an arrival or not. This works well because most students are comfortable with the concept of independent rolls of a die reflecting a memory-less property, e.g. the number of times it takes to roll a specific value does not depend on the number of trials since the last time that value was rolled. Under this model the expected number of arrivals per hour would be $np = 60(1/6) = 10$ arrivals per hour, and the average time between arrivals would be on average 6 one-minute intervals, or 6 minutes. The time until the next arrival would be a geometrically distributed number of one-minute intervals, with probability mass as depicted in Figure 1. Recall that under the geometric distribution,

978-1-4244-2708-6/08/$25.00

each successive probability is a constant fraction (5/6 in the case of Figure 1) of the previous value. This is an important observation that will be useful later when interpreting the shape of the exponential distribution.

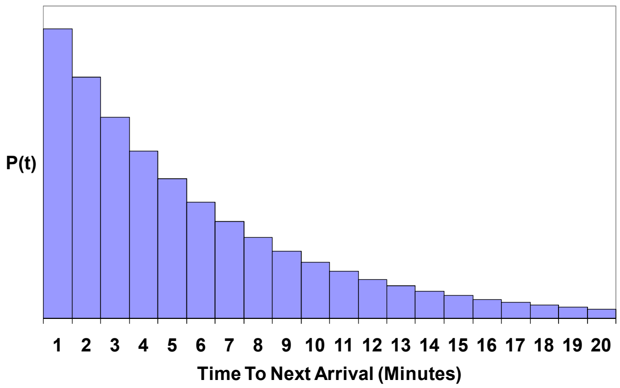

Figure 1: Time until next arrival, modeled as a geometrically distributed number of one-minute intervals with p = 1/6.

A similar model is produced by splitting the one-hour period into 120 30-second intervals, where each 30-second interval has a 1/12 probability of seeing one arrival, and 11/12 probability of seeing no arrivals. Again, we have on average $np = 120(1/12) = 10$ arrivals per hour, and average time between arrivals is 12 30-second intervals, or 6 minutes. If each 30-second interval represents an independent trial, then again the memory-less property applies to time between arrivals under this model. The probability mass for the number of 30-second intervals is depicted in Figure 2.

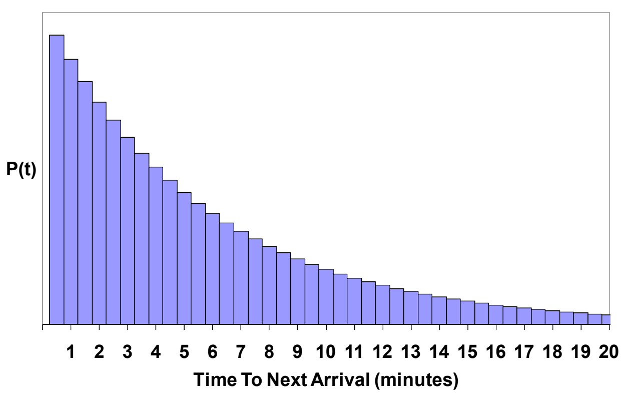

Figure 2: Time until next arrival, modeled as a geometrically distributed number of 30-second intervals with p = 1/12

Similar logic applies to breaking the hour into 360 independent 10-second intervals, where each has probability 1/36 of seeing one arrival and 35/36 of seeing no arrivals, as shown in Figure 3.

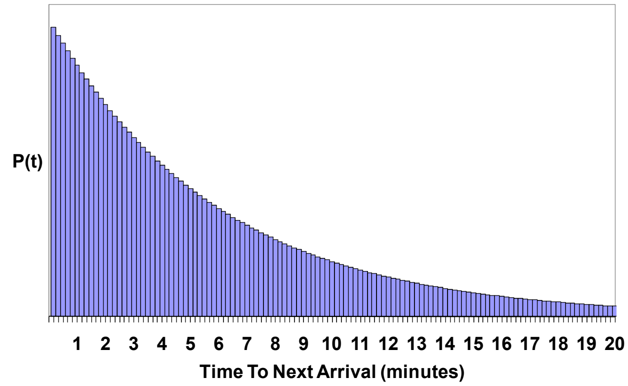

Figure 3: Time until next arrival, modeled as a geometrically distributed number of 10-second intervals with p = 1/36.

The exponential distribution shown in Figure 4 is then easily explained as the limiting case of this process, as the number of intervals n goes to infinity, while the probability p adjusts accordingly such that np is held constant at some value λ (10 per hour in this case). Each instant of time is considered an independent trial, seeing zero or one arrival, independent of all other instants of time.

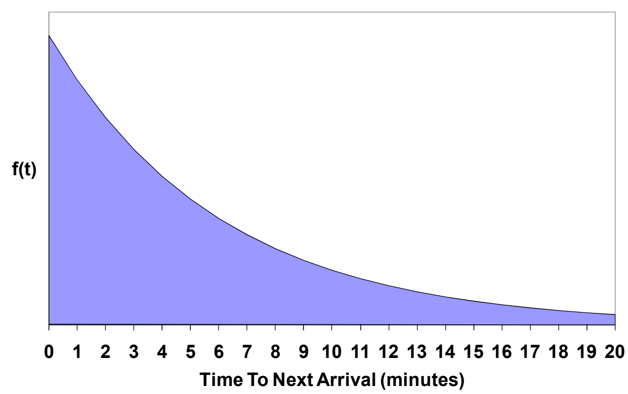

Figure 4: Time until next arrival, modeled as exponential distribution with mean 6 minutes

Understanding the implications of the shape of the exponential distribution is also very instructive. Viewing Figure 4, note the exponential distribution (along with its discrete counterpart, the geometric) has the only shape such that we can truncate the distribution at some time t, take the portion of the distribution to the right of t, rescale it to area of one, and have the same shape we started with.

Figure 5, in contrast, demonstrates an attempt to do this with some other arbitrary distribution. Here, the shaded area represents the shape of the conditional distribution for *remaining* time until the next arrival, given that an arrival has not occurred by some time t. (The shaded area would of course need to be shifted to a lower bound of zero, and re-scaled to an area of one in order to accurately describe this conditional time remaining.)

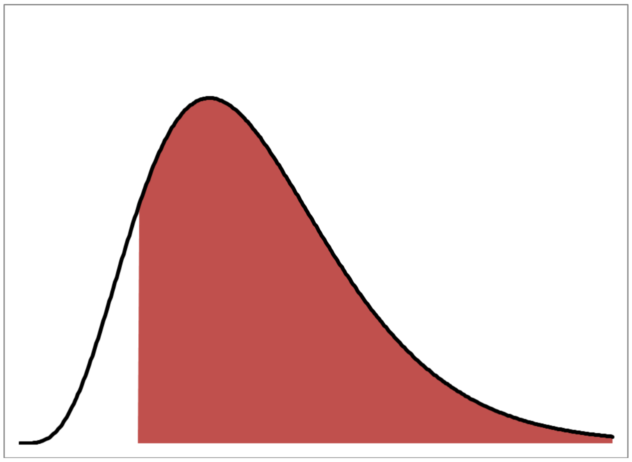

Figure 5: Arbitrary distribution reflecting lack of memory-less property

Note also the implications of the left tail of the arbitrary distribution shown in Figure 5, resulting in mode greater than zero. The small probability of inter-arrival times close to zero, with increasing probability as *t* approaches the mode, implies that one arrival somehow "discourages" or "delays" other arrivals from occurring until some amount of time has passed.

2.2 Why Arrival Processes?

The preceding discussion helps to explain what the memory-less property means, how it can be interpreted (each instant of time is an independent trial), and why the exponential distribution is the only continuous distribution that displays this property. The question remains, however, as to why we would expect any real-life arrival process to follow this model and display this property. The explanation of this was alluded to in the examination of Figures 4 and 5. When customer arrivals come from a large (essentially infinite) "calling population", each potential customer independently determines when to arrive. As my arrival neither hastens nor delays the independent timing of any other person's arrival, then the memoryless property must apply, and we should see exponentially distributed times between arrivals.

3 DATA COLLECTION ASSIGNMENT

Understanding that time between arrivals "should be" exponentially distributed motivates the data collection and distribution fitting exercise. Students are required to collect 30 to 40 observations of time between arrivals to some service facility. (Not entirely coincidentally, the data collection activity occurs during spring break, which can result in some interesting scenarios.) Specific instructions must be provided in order to avoid some common pitfalls that have occurred over the semesters:

- Measure inter-arrival times in seconds, and report them in seconds format. For example, 90 seconds should be reported as 90, not 1:30
- Capture data from a "stationary" arrival process. Collect your data during an interval in which the mean arrival rate is expected to be constant. For example, you will not want to mix inter-arrival times during the busy lunch hour at a fast food restaurant with inter-arrival times collected during a less busy portion of the day.
- Capture "balkers" as part of the arrival process. If possible, try to identify customers who arrive to the system, but choose not to enter the queue if there is one. For example, you may observe a car entering the drive-up area for an ATM machine, but leave due to the line being too long. Use your judgment.
- Measure time between arrival *events*. Treat "batch arrivals" as a single event. For example, if four people get out of a car and enter a facility at the same time, record it as a single arrival event. (Do not record three inter-arrival times of zero for the three customers who follow the first through the door.)
- Measure arrivals to *the system* not arrivals to *the server*. In a queuing situation, measuring the time between customers in queue approaching the server is actually measuring successive service times (if there was always at least one person in line), not inter-arrival times.

After the data have been collected, students use distribution fitting software to investigate the shape of the resulting histogram. With a few exceptions, the data displays the distinctive "one tailed" shape of the exponential distribution, often times after some experimentation with the number of histogram cells used to view the data. This provides an excellent opportunity to demonstrate to students a situation where "the theory works", building on the earlier discussion of the unique implications of the shape of the distribution.

4 GOODNESS OF FIT MEASURES

One complication, (from the students' viewpoint), is that the distribution fitting software quite often produces some other distribution as "best fitting". This provides an opportunity to review some basic principles of distribution fitting, and sampling in general. For example, "goodness of fit" tests conducted within a hypothesis testing framework can be used to reject some distributions from consideration, but cannot be expected to definitively identify the "correct" distribution. As with any hypothesis test, "failing to reject" the null hypothesis (where the null hypothesis is that the hypothesized distribution is the correct one) does not confirm the null hypothesis to be true. This concept is reinforced by demonstrating that any number of hypothesized distributions can generate the "fail to reject" value of the test statistic in use.

Beyond the immediate implications of the hypothesis testing framework, students should realize that the minimizing a goodness of fit measure cannot be expected to reliably identify the "correct" distribution. This may become an issue, as without a proper understanding of the variability associated with goodness of fit results, students may conclude the sampling exercise has "disproved" the memory-less theory for their sampling scenario.

To assist with this issue, I have conducted a sampling experiment in which samples of n = 30 observations are generated from an exponential distribution. The "best fitting" distribution is identified for each sample using distribution fitting software. Table 1 displays the results of 100 iterations of this process, using the Palisade *BestFit* Excel add-in (www.palisade.com). In order to provide a reasonable test, test distributions were limited to those bounded below at zero, and unbounded above. Default histogram parameters were used for the Chi-Square statistic. The table displays the number of times each distribution was identified as "best fitting" by each goodness of fit measure (Chi-Square, Anderson-Darling, Kolmogorov-Smirinov) listed. Cases where the Weibull, Gamma, or Erlang distributions are counted as "best fit" represent cases where the fit was distinct from the exponential fit. (All three of those distributions have the ability to represent the exponential distribution as a special case, but can take on other shapes as well. This gives them a distinct advantage in being able to adapt to random variation in samples produced by an exponential distribution.)

The intent here is not to compare the effectiveness of alternative goodness of fit measures. The intent, rather, is to provide a compelling class demonstration that failure to "fit best" by no means rules out any particular distribution as the "correct" distribution. This helps students understand that their own data may reflect an exponential distribution (and hence the memoryless property), despite the fact that goodness-of-fit measure may seem to be indicating otherwise. Indeed, given a strong prior rationale for believing the memoryless property should hold, a "decent" fit for the exponential distribution, even if not the "best" fit, should be taken as confirming rather than contradictory evidence.

Table 1: Number of times a distribution was reported as "best fitting" under alternative goodness of fit measures. 100 samples of size n = 30 each generated from exponential distribution with mean = 10.

	Goodness of Fit Statistic		
Distribution	χ^2	**A-D**	**K-S**
Weibull*	16	39	31
Exponential	25	12	13
Log Logistic	11	9	22
Pearson (5 or 6)	7	18	10
Gamma*	13	11	8
Log Normal	17	6	4
Pareto 2	5	4	9
Inverse Gauss	3	1	2
Erlang*	2	0	1
Chi-Square	1	0	0
Sum	100	100	100

Table 2 below describes the results of a similar process, under the squared error criterion used in the Rockwell Software (Arena) *Input Analyzer* (www.arenasimulation.com). Here I allowed for bounded distributions, including the beta distribution which due to its flexible nature was able to fit a number of samples quite well.

Table 2: Number of times a distribution was reported as "best fitting" under the squared error goodness of fit measure. 100 samples of size n = 30 each generated from an exponential distribution with mean = 10.

Distribution	**Number of Times "Best Fitting"**
Log Normal	28
Beta	20
Weibull*	17
Exponential	15
Gamma*	13
Normal	6
Erlang*	1
Sum	100

Finally, we often get into a discussion of situations that would distort the memoryless nature of the arrival process, resulting in a true deviation from the exponential distribution. The classic example is the case of an establishment in close proximity to a traffic light. To the extent that the traffic light impacts the arrival process, making it somewhat "less random", we see a departure from exponential. Mixing high-volume and low volume arrival processes into a single dataset (in violation of one of the instructions provided above) will result in a "hyper-exponential" distribution, as the unduly large number of small inter-arrival times captured during the high volume phase is graphed against the longer arrival intervals captured during the low volume phase.

5 SUMMARY

We rarely encounter real-life sampling situations in which the correct form of distribution is “known”. In the case of an appropriately specified customer arrival process, however, this is indeed the case as the memory-less property must apply. As such, this provides a dual opportunity for a real-life data collection and distribution fitting exercise. The first opportunity is to confirm the theory that allows us to “know” the correct distribution as exponential. The second is to demonstrate to students the danger of over-reliance on goodness of fit measures, overlooking the implications of random variation in sampling data and the impact on the measures themselves.

AUTHOR BIOGRAPHY

TIMOTHY S. VAUGHAN is Professor of Operations Management at University of Wisconsin- Eau Claire, and is Chair of the Department of Management and Marketing. He received his Ph.D. in Business Administration from the University of Iowa in 1988. His research interests include production and inventory management, production scheduling, and statistical quality control methods. His e-mail address is <vaughats@uwec.edu> and his Web address is <http://www.uwec.edu/cob/info/faculty_staff/vaughan.htm>.