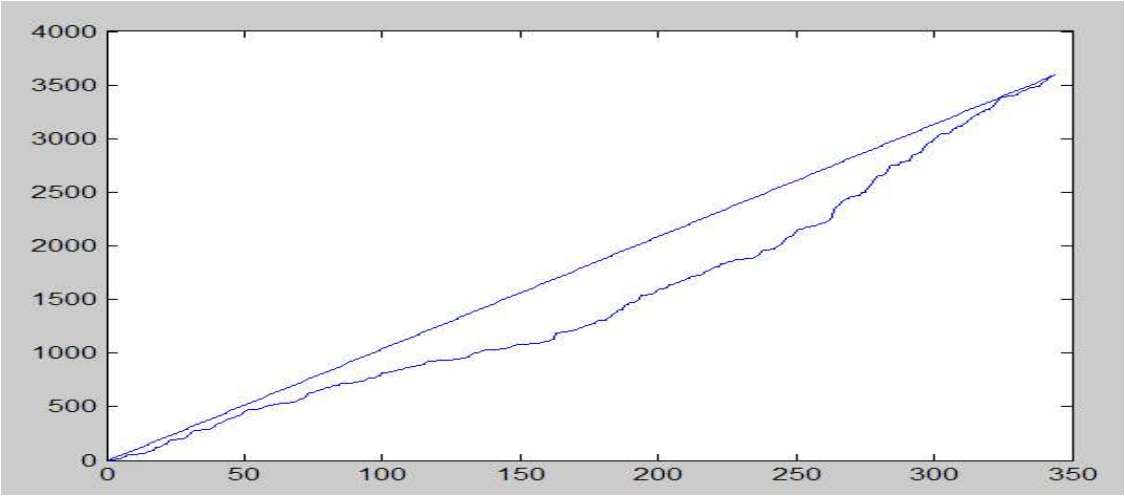
Dave Raines
March 10th, 2006

Probability Distributions of Students in a College Dining Hall

The data for this project was obtained by sitting at the entrance to Food Court, which is the largest dining facility on campus, between the hours of 6 and 7 pm on a Thursday. The time and location were designed to get the largest possible data set. In this hour, 486 students entered, in 343 distinct groups. Data was collected with SnapTimePro, a freeware stopwatch utility. For the wait time distribution, each group was considered as a single individual, with the time of the first member to enter being used as the group time. The group size was also considered, and its distribution analyzed.

Waiting distribution between random, independent events of a specific rate theoretically follows a gamma distribution. The main purpose of this experiment was to decide if the waiting time between students follows a gamma distribution. One problem with this was the changing rate of students entering. As the graph below illustrates, the rate of entry was decreasing throughout the hour, with the area being busier earlier, and getting less busy as the hour progressed.



Horizontal - index entering, Vertical - Time (in seconds), straight line indicates a constant rate.

This variance of rate caused some problems, but was not large enough to completely ruin

the data.

The gamma distribution is

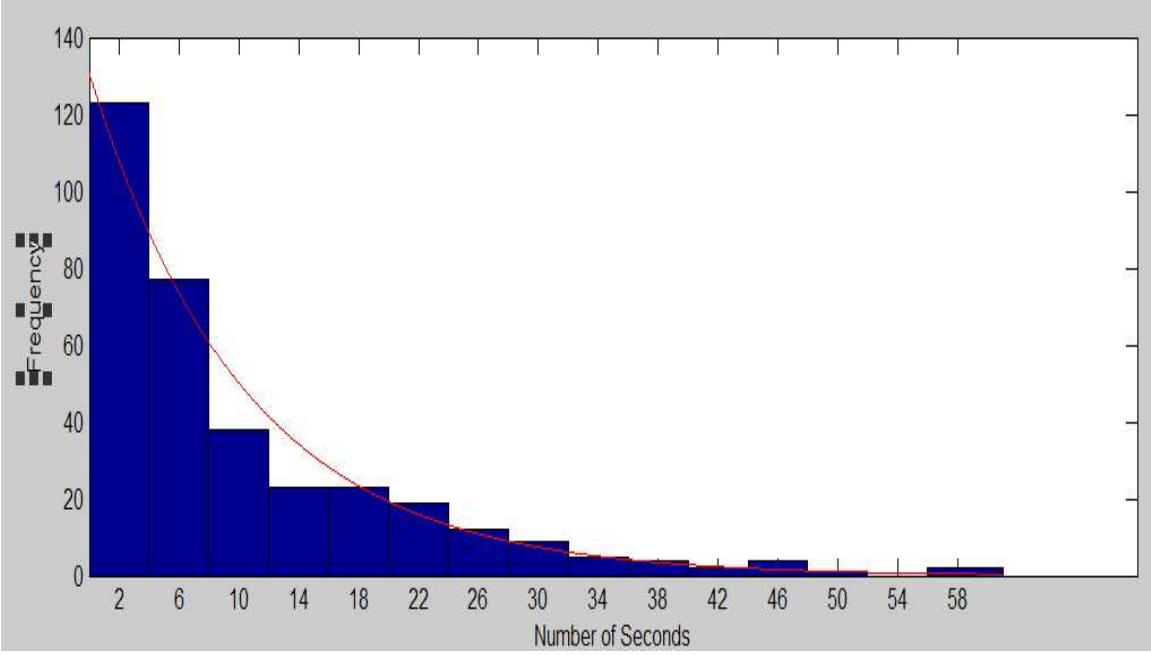
$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)} \text{ for } x > 0$$

With Theta representing the rate of entering, and k being the number being waited for. That is, k=1 would be the distribution of every entrant, and k = 2 would be the distribution of wait time for two entrants. The maximum likelihood estimate of the gamma distribution is

$$\theta = \frac{1}{kN}\sum_{i=1}^{N} x_i$$

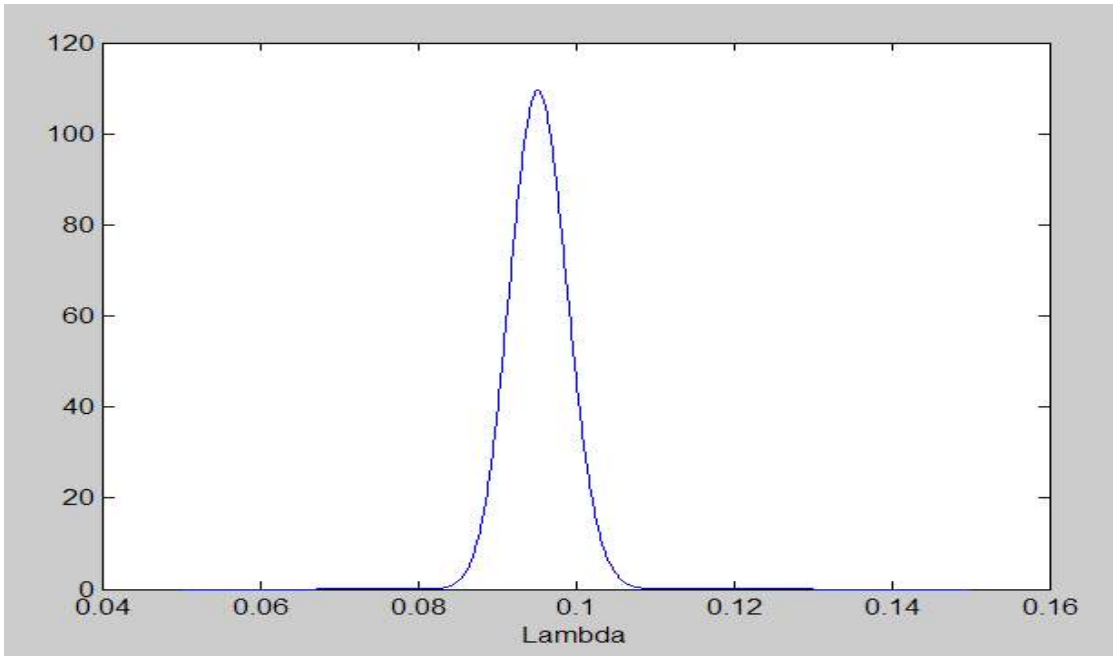
which is essentially just the inverse of the mean wait time, divided by k. It is this maximum likelihood estimate which was used to estimate the parameters for all the models which follow.

The first model is the estimate when k=1, the distribution of wait time between individuals. This subset of the gamma distribution, known as the exponential distribution, is simply exponential decay, normalized as a pdf. The graph below shows a histogram of the actual data in blue, and prediction line from the exponential distribution in red. Due to the continuous nature of time, putting together a histogram is necessarily a subjective exercise. For this reason, a goodness-of-fit test would not be of value, but it is clear from the graph that the distribution is a reasonably good fit to the data.



Wait time, red line is gamma distribution with k=1, theta is mle of .0951

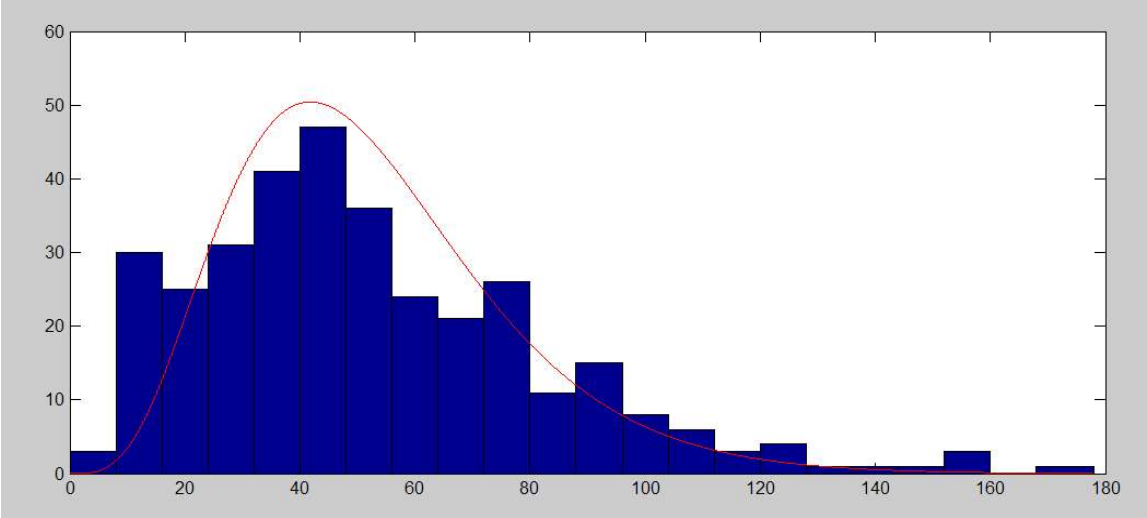
Due to the large data set, the rate, theta, can be given with a high degree of certainty, the graph below is the Bayesian posterior on the rate. The 95% Bayesian confidence interval, that is, the values between which 95% of the probability of the rate lies, is .088 and .1022, a width of less than two hundredths of a second.



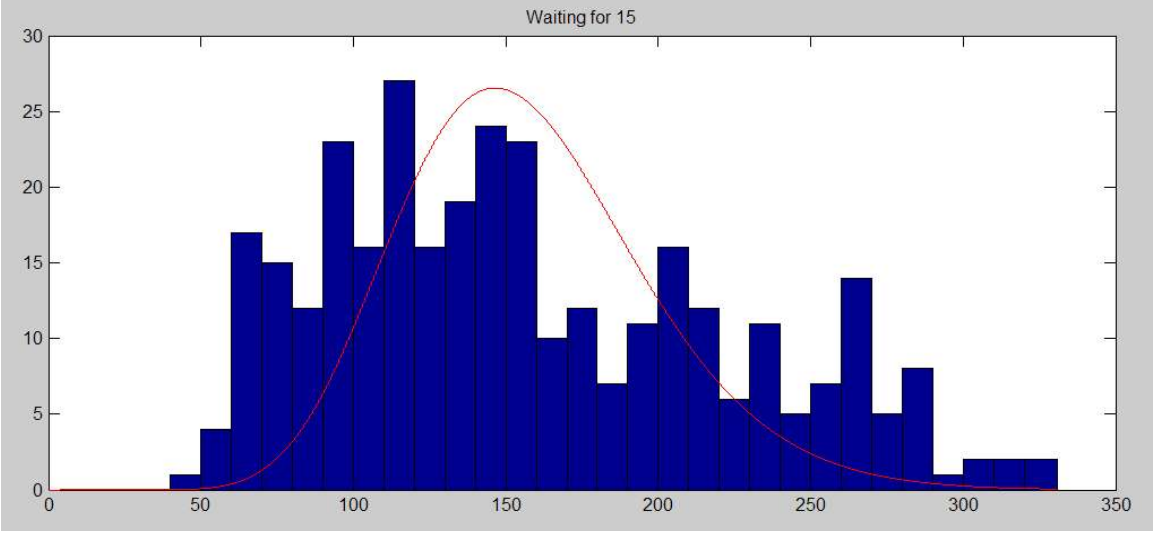
The Bayesian Posterior of the Rate of entry.

As shown by the graphs on the previous page, the approximation works very well for the

case of k=1, despite the inconsistent rate throughout the hour. The differences are hidden by the exponential pdf, which has smaller apparent difference than the gamma distribution for higher levels of k. As seen in the graphs below, for higher levels of k, the data follows the gamma distribution less closely. The small differences in rate are compounded by the larger number of individuals, and there are many observations well outside the predicted gamma.



Waiting Distribution for 5 People, Modeled as Gamma Distribution with k=5.



Waiting Distribution for 15 people, Modeled as Gamma Distribution with k=15

As shown in these graphs, even for k=5, the data follows the model reasonably

well, though certainly not perfectly. For k =15, however, a significant portion of the data lies well outside the model range. It seems intuitively reasonable, however, since even a short period with a faster (or slower) rate would lead to large differences in values when multiplied by 15, while these differences wouldn't appear as significant when multiplied by 5 (or 1).

The size of the groups was also examined, and found to be best modeled by a positive poisson distribution. The positive poisson is the same as the poisson, but is used when values of 0 can not be observed (or are skewed for some reason). Since obviously groups of size zero can't be observed, the positive poisson is a natural choice.

The positive poisson distribution is simply the poisson distribution, divided by zero figure from the poisson:

$$f(y;\lambda) = \begin{cases} \dfrac{e^{-\lambda}\,\lambda^{y}}{y!\,(1-e^{-\lambda})} & y = 1, 2, \ldots \text{ and } \lambda > 0, \\ 0, & y = 0. \end{cases}$$

It's maximum likelihood estimator is

$$\hat{\mu}_i = \frac{\hat{\lambda}_i}{1 - e^{-\hat{\lambda}_i}}$$

In this case, 486 people were divided into 343 groups, so our sample mean was 486/343 = 1.416, which solved numerically gives lambda = .7427. The following chart lists the actual group sizes, followed by the value predicted by a positive poisson with lambda of .7427.

Group size	actual size	predicted value
1	240	231.2
2	74	85.8
3	21	21.26
4	6	3.925
5	1	.9898
6	1	.116

A goodness of fit test gives us a chi-squared value of 9.79, which is less than the 95% confidence level of rejection, 11.07, so we cannot reject the positive poisson distribution as a model of group size.

It is a strange mental exercise to theorize why the group size of college diners should follow a poisson distribution. The best way I have thought of to view it is to assign every person a real number value, which represents any number of factors about who they are, where they are, who they know, and what they're doing for dinner. These peoples values are then rounded to the nearest integer, and anyone landing on the same integer has a close enough situation that they will walk into food court with them. In our example the 486 people are assigned numbers between 0 and 655 (486/lambda). This would give us the desired poisson distribution, but whether this accurately portrays (although simplified) how the distribution comes about in reality would require more research.