***Problem 1a*:**

Times were recorded at which 41 vehicles passed a fixed point on the M1 motorway in Bedfordshire, England on March 23, 1985.[1] The times were subtracted to form 40 intervals between successive cars. These interarrival times, rounded to the nearest second, are:

12, 2, 6, 2, 19, 5, 34, 4, 1, 4, 8, 7, 1, 21, 6, 11, 8, 28, 6, 4, 5, 1, 18, 9, 5, 1, 21, 1, 1, 5, 3, 14, 5, 3, 4, 5, 1, 3, 16, 2

A common model for interarrival times is a random sample from an exponential distribution. Do you think an exponential distribution provides a good model for the interarrival times? Justify your answer.

***Solution*:**

An exponential distribution is a good model to consider for the interarrival times because:

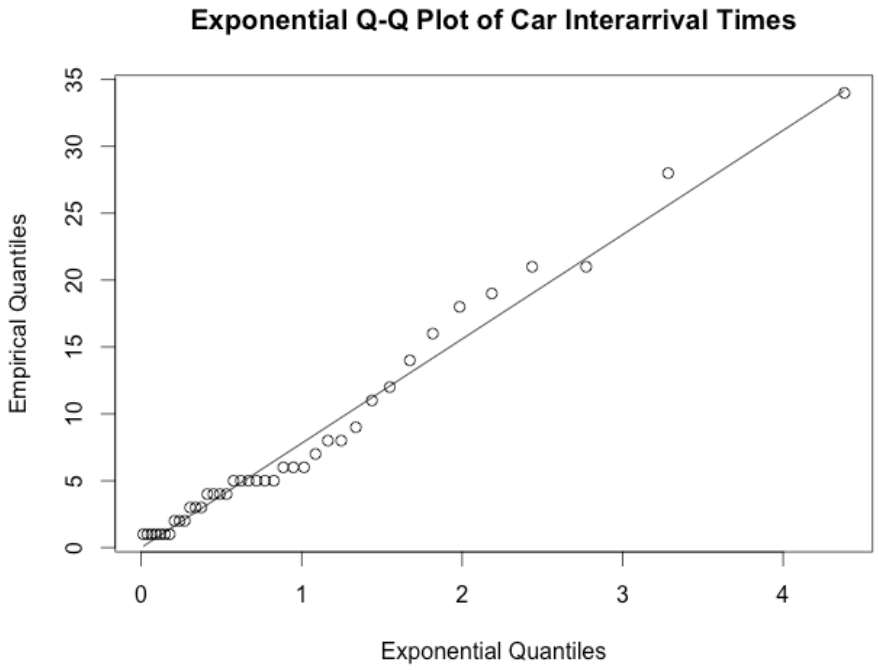
- The data consist of times between events.
- It seems reasonable that the rate of occurrence would be constant over the time period (about 5 minutes) that the sample was taken.
- It seems reasonable that the time until the next car passes would not depend on how long it has been since the last car passed.

But the reasonableness of these assumptions is not sufficient. We need to evaluate whether the exponential model actually fits the observations. A useful tool for this purpose is the quantile-quantile plot. A q-q plot is shown on the next page. It plots quantiles of the car interarrival times against quantiles for a standard (scale = 1) exponential distribution. R code for producing this plot is available for download on Blackboard.

By inspection, the plot looks fairly linear. Notice that there is an anomaly at the lower end of the plot due to all times being rounded to the nearest second. These repeated values create problems for some formal goodness-of-fit tests: for example, the Kolmogorov-Smirnov test cannot be used when there are repeated values. Although a formal test was not required for this problem, we can do a chi-square goodness-of-fit test.[2] To do this, I broke the range of the sample into bins, setting the bin sizes so all bins have equal expected count. Using the rule of thumb that the expected count should be at least 5, I used 8 bins. I set the bin boundaries at the quantiles of the exponential distribution with mean equal to the sample mean of the observations, $\overline{X} = 7.8$. The chi-square test statistic is calculated as:

$$\chi^2 = \sum_i \frac{(Y_i - E_i)^2}{E_i}$$

[1] These data were taken from Hand, et al., *A Handbook of Small Data Sets*, Chapman and Hall, 1994.
[2] A good reference on the chi-square goodness of fit test is here:
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm



where Y_i is the observed count of interarrival times in the i^{th} bin, and E_i is the expected number of interarrival times in the i^{th} bin for a sample of 40 interarrival times. The observed and expected counts are given in the following table. (R code for computing the observed and expected counts is available on Blackboard.)

Time Interval	Observed Count	Expected Count	$(X_i-E_i)^2/E_i$
0 – 1.04	7	5	0.8
1.0 – 2.2	3	5	0.8
2.2 – 3.7	3	5	0.8
3.7 – 5.4	10	5	5.0
5.4 – 7.7	4	5	0.2
7.7 – 10.8	3	5	0.8
10.8 – 16.2	4	5	0.2
> 16.2	6	5	0.2

Adding up the last column gives $\chi^2 = 8.8$.

Next, we find the degrees of freedom. The degrees of freedom should be $n - p$ -1, where n is the number of categories and p is the number of parameters estimated. We have 8 categories, and we estimated the mean of the exponential distribution, so the degrees of freedom are 8 - 1 - 1, or 6. Therefore, we compare our test statistic $\chi^2 = 8.8$ against the critical value of the chi-square distribution with 6 degrees of freedom. The 95th percentile for the chi-square distribution with 6 degrees of freedom is 12.6.

The test statistic $\chi^2 = 8.8$ is smaller than the critical value of 12.6 for a chi-square distribution with 6 degrees of freedom. Therefore, we cannot reject the null hypothesis that the observations have an exponential distribution.

If you gave a thoughtful argument for whether the exponential distribution was a good model, and if you used the data in a reasonable way to evaluate the fit of the distribution, you would receive credit for this problem. You did not have to do a goodness-of-fit test.

***Problem 1b*:**

When interarrival times are randomly sampled from an exponential distribution, the counts of events per unit time are a random sample from a Poisson distribution. Using a time unit of 15 seconds, find the number of cars passing in each 15-second block of time after the initial car. (The initial car is used to bound the recording interval, so the total car count in your data set should be 40.) Do you think a Poisson distribution provides a good model for the count data? Justify your answer.

Solution:

A Poisson distribution is a good model to consider for the arrival counts because:

- The data consist of counts of discrete events.
- It seems reasonable that the events would be independent.
- It seems reasonable that the rate of occurrence would remain constant over the 4 minutes that the sample was taken.

Also, if the interarrival times are exponential, then the counts are Poisson. So any argument or against for exponential interarrival times is also an argument for or against Poisson counts, and vice versa. For part a, we found a reasonably linear q-q plot and a chi-square test based on exponential quantiles failed to reject the null hypothesis. This is an argument for both exponential interarrival times and Poisson counts.

For this problem, I will evaluate how well the Poisson model fits the observations.

To investigate whether a Poisson distribution fits these observations, we compare the sample frequencies with Poisson probability mass function. First, we find the counts of cars in 21 blocks of length 15 seconds (ignoring the fact that the 21st block is slightly shorter than 15 seconds). These counts are:

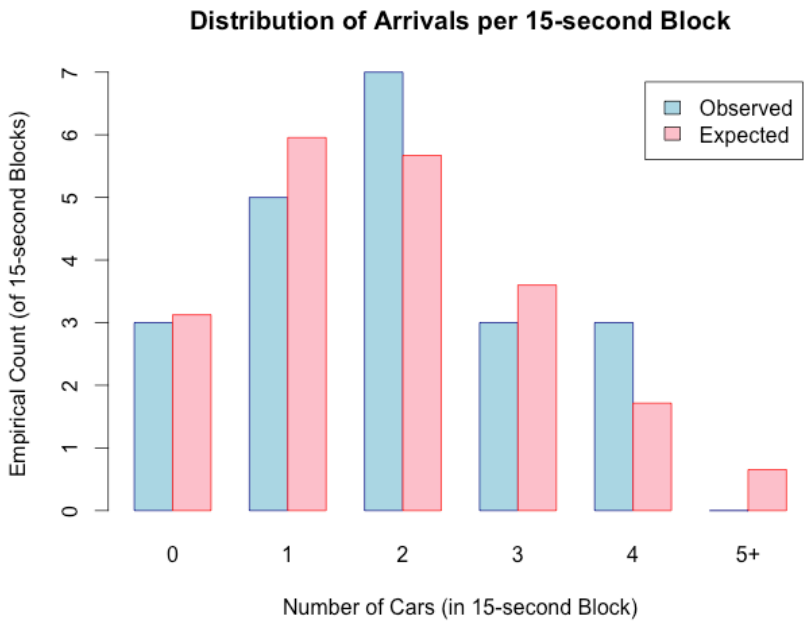
2, 2, 1, 1, 0, 4, 3, 0, 2, 1, 1, 1, 4, 0, 2, 2, 3, 2, 4, 3, 2

We calculate the sample mean by summing these observations (40 total cars passing) and dividing by the number of intervals (21), to obtain an estimated arrival rate of about 1.9 cars per 15 second time interval.

Next, we use the Poisson probability mass function f(x|Λ=1.9) = $f(x \mid \Lambda = 1.9) = e^{-1.9}(1.9)^x / x!$ to estimate the probability of each number of cars in a single interval, and multiply by 21 to find the expected number of times each value occurs (Note: for the 5 or more entry, we use 1-F(4|Λ=1.9), or 1 minus the Poisson cdf at 4. The probability of more than 5 cars in 15 seconds is very small.)

Number of Cars	Sample Occurrences	Expected Occurrences
0	3	3.12
1	5	5.95
2	7	5.67
3	3	3.60
4	3	1.71
5 or more	0	0.93

Examining the numbers in this table shows fairly good agreement. We can do a visual inspection by drawing a bar chart, shown below. (R code for making this chart is posted on Blackboard.) The bar chart shows good agreement between the empirical and the Poisson probabilities.



We can also perform a Pearson chi-square test (this is not required).[3] The test statistic is calculate as:

$$\chi^2 = \sum_i \frac{(Y_i - E_i)^2}{E_i}$$

[3] A good explanation of using the chi-square test to evaluate goodness of fit of the Poisson distribution is given at http://courses.wcupa.edu/rbove/Berenson/10th%20ed%20CD-ROM%20topics/section12_5.pdf

where Y_i is the observed count for i cars passing, and E_i is the expected number of instances of i cars passing in a sample of size 21.

The chi-square test should not be applied when counts are very small. A common (and conservative) rule of thumb is to avoid using the chi-square test if any cell has an expected count less than 5. Other less conservative rules have been proposed.[4] We will combine the last two categories to increase the expected count to 2.37.

To find the expected counts in this last category, we assign it probability 1 minus the cdf at x=3, and multiply by 21. The observed and expected counts are:

Number of Cars	Count	Expected Count	$(Y_i-E_i)^2/E_i$
0	3	3.12	0.0051
1	5	5.95	0.1530
2	7	5.67	0.3116
3	3	3.60	0.1002
4 or more	3	2.65	0.0467

Adding up the last column gives $\chi^2 = 0.616$.

Next, we find the degrees of freedom, $n - p$ -1, where n is the number of categories and p is the number of parameters estimated. We have 5 categories, and we estimated the mean of the Poisson distribution, so the degrees of freedom are 5 - 1 - 1, or 3. Therefore, we compare our test statistic $\chi^2 = 0.616$ against the critical value of the chi-square distribution with 3 degrees of freedom. The 95th percentile for the chi-square distribution with 3 degrees of freedom is 7.8.

The test statistic $\chi^2 = 0.616$ is much smaller than the critical value of 7.8 for a chisquare distribution with 3 degrees of freedom. Therefore, we cannot reject the null hypothesis that the observations have a chi-square distribution.

Some students do a Q-Q plot of the Poisson quantiles against the data quantiles. However, because there are only 4 distinct values in the sample, the plot shows most of the dots on top of each other (some authors suggest "jittering" each point a tiny random amount to avoid this problem). For discrete distributions with very few values, I find the comparison of empirical and theoretical frequencies to be much more informative than a q-q plot.

We can also compare the sample mean and sample variance of the observations, recalling that the mean and variance of the Poisson distribution are the same. The sample mean of the observations is 1.90 and the sample variance is 1.59. The standard deviation of the sample mean is approximately equal to the sample standard deviation divided by the square root of the sample size, or $\sqrt{1.59/21} = 0.275$. Therefore, the mean is roughly $1.90 \pm 0.275 \times 2$, or lies within the interval [1.35,

[4] http://www.basic.northwestern.edu/statguidefiles/gfdist_ass_viol.html#Small%20expected%20cell%20counts

2.46]. Therefore, the observations are consistent with the hypothesis that the mean and standard deviation are equal.

***Problem 1c*:**

Assume that Λ, the rate parameter of the Poisson distribution (and the inverse of the mean of the exponential distribution), has a discrete uniform prior distribution on 20 equally spaced values between (0.2, 0.4, ..., 3.8, 4.0) cars per 15-second interval. Find the posterior distribution after observing the first 10 observations of car counts in 15 second intervals. Find the posterior mean, standard deviation, median and 95th percentile of Λ given the first 10 observations.

Solution.

To find the posterior distribution, we calculate the Poisson likelihood at each of the 20 lambda values, and multiply by the prior probability of 1/20. Then we divide each of these values by their sum. The likelihood is the product of the likelihoods for the observations: 2 observations of 0, 3 observations of 1, 3 observations of 2, 1 observation of 3, and 1 observation of 4. The formula is:

$$p(\lambda|X_1,\ldots,X_{10}) = \frac{\frac{1}{20} f(0|\lambda)^2 f(1|\lambda)^3 f(2|\lambda)^3 f(3|\lambda) f(4|\lambda)}{\sum_{i=1}^{20} \frac{1}{20} f(0|\lambda_i)^2 f(1|\lambda_i)^3 f(2|\lambda_i)^3 f(3|\lambda_i) f(4|\lambda_i)}$$

Note that the prior $g(\lambda) = 1/20$ factors out of both the numerator and denominator, so it is unnecessary to include it. In fact, when the prior distribution is uniform, the posterior distribution is the same as the normalized likelihood.

A plot of the posterior pmf is shown on the next page.

R code is provided on Blackboard for calculating the posterior pmf, mean, and standard deviation. The posterior mean and standard deviation, rounded to 3 decimal places, are:

- The posterior mean is $E[\lambda \mid \underline{X}] = \sum_i \lambda_i p(\lambda_i) = 1.700.$
- The posterior standard deviation is $\sqrt{\sum_i (\lambda_i - E[\lambda \mid \underline{X}])^2 \, p(\lambda_i)} = 0.412.$

To find the median, 0.95 quantile, a symmetric tail area 95% credible interval, and the mode:

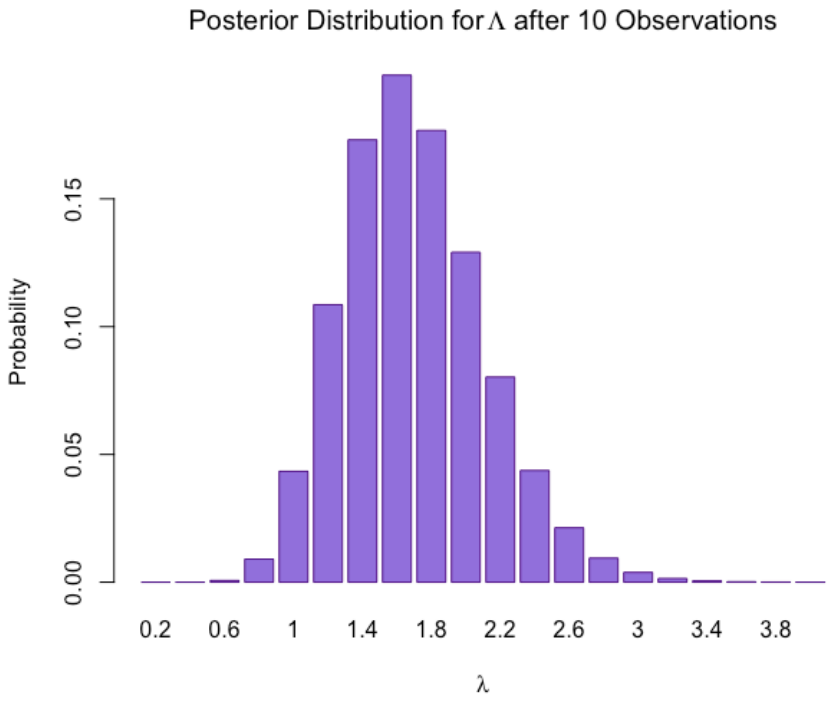
- The median of the posterior distribution is a value $\lambda_{0.5}$ such that $P(\Lambda \geq \lambda_{0.5} \mid \underline{X}) \geq ½$, and $P(\Lambda \leq \lambda \mid \underline{X}) < ½$ for all $\lambda < \lambda_{0.5}$. We calculate the cdf of the posterior distribution $F(\lambda \mid \underline{X})$ (see R code posted on Blackboard) and find that F(0.1.4) = 0.0.355 and F(1.6) = 0.533. Therefore, the median is 1.6, because at least 50% of the probability is less than or equal to this value and at least 50% of the probability is greater than or equal to this value.

- $P(\Lambda \leq 0.8|\underline{X}) = 0.010$ and $P(\Lambda \leq 1|\underline{X}) = 0.053$. Therefore, the 2.5th percentile is $\lambda = 1.0$, because 1 is the minimum value of Λ for which the cdf is larger than 0.025.
- $P(\Lambda \leq 2.2|\underline{X}) = 0.919$ and $P(\Lambda \leq 2.4|\underline{X}) = 0.963$. Therefore, the 95th percentile is $\lambda = 2.4$, because 2.4 is the minimum value of Λ for which the cdf is larger than 0.95.
- $P(\Lambda \leq 2.4|\underline{X}) = 0.963$ and $P(\Lambda \leq 2.6|\underline{X}) = 0.984$. Therefore, the 97.5th percentile is $\lambda = 2.6$.
- The interval [1.0, 2.6] excludes probability 0.010 for values between 0 and 1.2, and probability 0.016 for values between 2.8 and 4. Therefore the interval [1.4, 2.6] excludes probability 0.010 + 0.016 =.025 (discrepancy in 3rd decimal place is due to roundoff). Therefore, [1.4, 2.6] is a 1-.025 = 97.5% credible interval for Λ.
- The mode of the posterior distribution is the most probable value, which occurs at $\lambda = 1.6$.

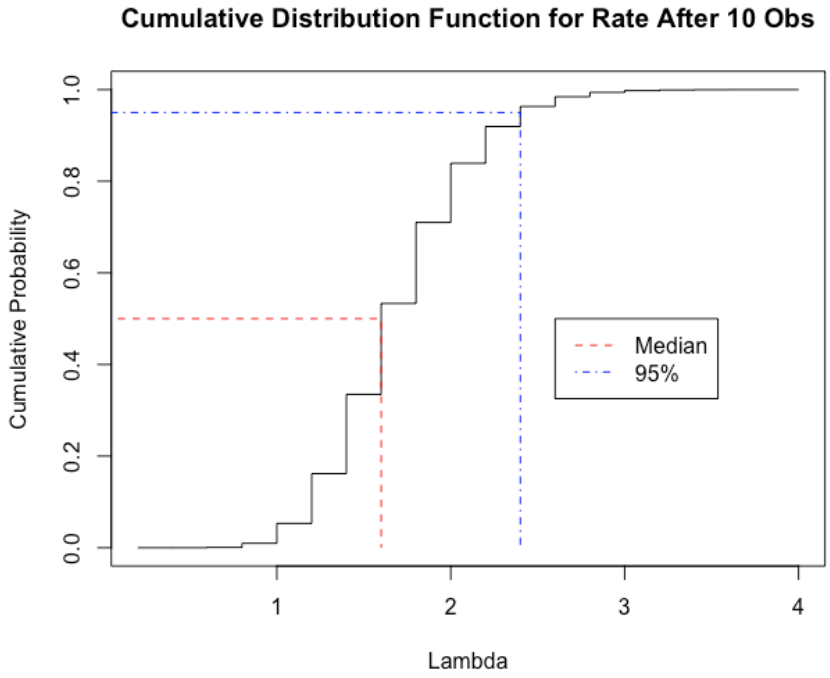
The accompanying R code finds these quantiles directly and also using the `qdiscrete` function in the e1971 package.

To summarize:

- Posterior mean of Λ is 1.60
- Posterior standard deviation of Λ is 0.412
- Posterior median of Λ is 1.6.
- Posterior 95th percentile of Λ is 2.4.



The plot below shows the cumulative distribution function. The median is shown as a red dashed line; the 95th percentile is shown as a blue dotted/dashed line.



***Problem 1d*:**

Using the posterior distribution from part c as the prior distribution, find the new posterior distribution after observing the remaining observations. Find the posterior mean, standard deviation, median and 95th percentile of Λ given all observations.

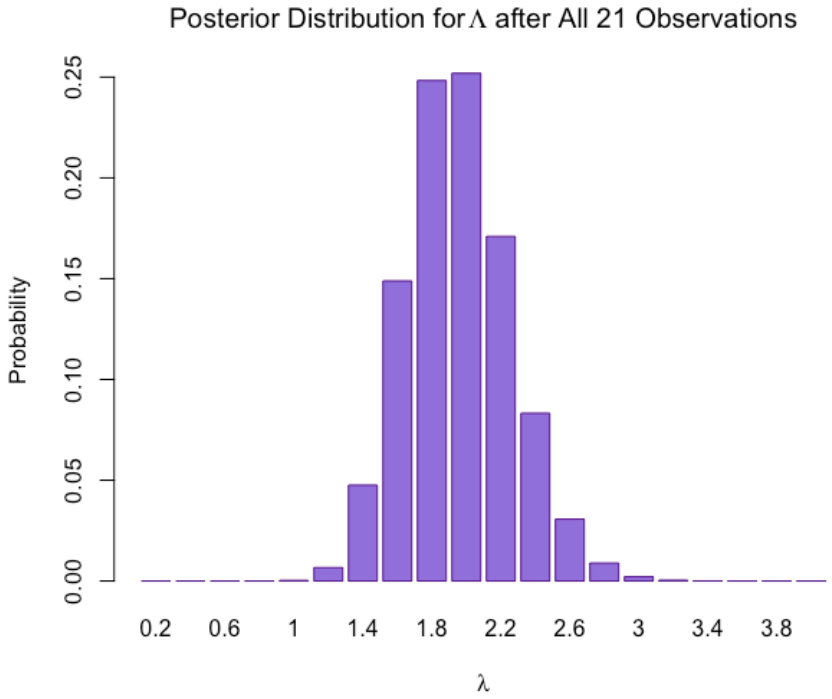
Solution.

To find the posterior distribution, we calculate the Poisson likelihood at each of the 20 lambda values, and multiply by the prior probability from part c. This time, we have 21 observations: 1 value of 0, 2 values of 1, 4 values of 2, 2 values of 3, and 2 values of 4. After finding this product, we divide each of these values by their sum. The formula is:

$$p(\lambda|X_1,\dots,X_{21}) = \frac{p(\lambda|X_1,\dots,X_{10})f(0|\lambda)f(1|\lambda)^2f(2|\lambda)^4f(4|\lambda)^2f(3|\lambda)^2}{\sum_{i=1}^{20}(\lambda|X_1,\dots,X_{10})f(0|\lambda)f(1|\lambda)^2f(2|\lambda)^4f(4|\lambda)^2f(3|\lambda)^2}$$

This time, the prior pmf does not factor out of numerator and denominator. ***You should verify that we will get the same result if we use the posterior distribution from part c as our prior distribution and process observations 11-21, or if we start with a uniform prior distribution and process all 21 observations.***

A plot of the posterior pmf is shown here. Comparing with the result of part c, we can see that this distribution is more concentrated. The probabilities for extreme values (1 or less, 3 or greater) are smaller for this plot; the probability of values near the mode is higher for this plot.



R code is provided on Blackboard for calculating the posterior pmf, mean, and standard deviation. The posterior mean and standard deviation, rounded to 2 decimal places, are:

- The posterior mean is $E[\lambda \mid \underline{X}] = \sum_i \lambda_i p(\lambda_i) = 1.95.$
- The posterior standard deviation is $\sqrt{\sum_i (\lambda_i - E[\lambda \mid \underline{X}])^2 \, p(\lambda_i)} = 0.305.$

To find the median, 0.95 quantile, a symmetric tail area 95% credible interval, and the mode:

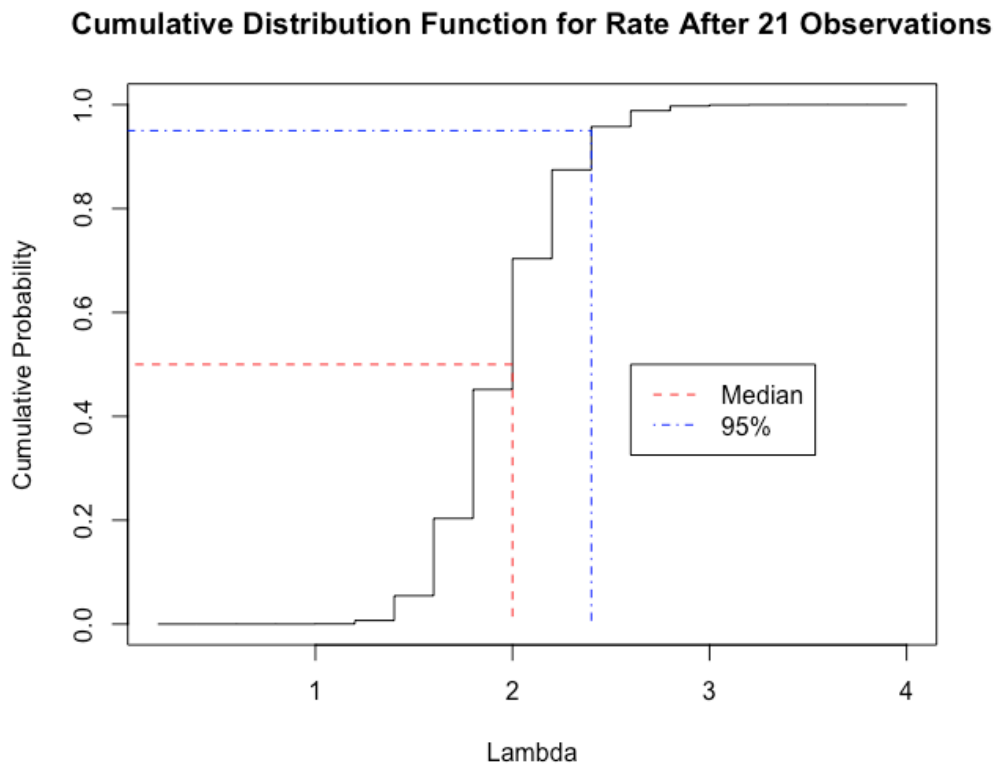
- The median of the posterior distribution is a value $\lambda_{0.5}$ such that $P(\Lambda \geq \lambda_{0.5} \mid \underline{X}) \geq ½$, and $P(\Lambda \leq \lambda \mid \underline{X}) < ½$ for all $\lambda < \lambda_{0.5}$. We calculate the cdf of the posterior distribution $F(\lambda \mid \underline{X})$ (see R code posted on Blackboard) and find that F(0.1.8) = 0.0.452 and F(2.0) = 0.704. Therefore, the median is 2.0, because at least 50% of the probability is less than or equal to this value and at least 50% of the probability is greater than or equal to this value.
- $P(\Lambda \leq 1.2 \mid \underline{X}) = 0.007$ and $P(\Lambda \leq 1.4 \mid \underline{X}) = 0.055$. Therefore, the 2.5th percentile is $\lambda = 1.4$, because 1.4 is the minimum value of Λ for which the cdf is larger than 0.05.

- $P(\Lambda \leq 2.2|\ \underline{X}) = 0.875$ and $P(\Lambda \leq 2.4|\ \underline{X}) = 0.958$. Therefore, the 95th percentile is $\lambda = 2.4$, because 2.4 is the minimum value of Λ for which the cdf is larger than 0.95.
- $P(\Lambda \leq 2.4|\ \underline{X}) = 0.958$ and $P(\Lambda \leq 2.6|\ \underline{X}) = 0.989$. Therefore, the 97.5th percentile is $\lambda = 2.6$.
- The interval [1.4, 2.6] excludes probability 0.007 for values between 0 and 1.2, and probability 0.012 for values between 2.8 and 4. Therefore the interval [1.4, 2.6] is a 98.2% credible interval for Λ.
- The mode of the posterior distribution is the most probable value, which occurs at $\lambda = 2$.

The accompanying R code finds these quantiles directly and also using the `qdiscrete` function in the e1971 package.
To summarize:

- Posterior mean of Λ is 1.95
- Posterior standard deviation of Λ is 0.305
- Posterior median of Λ is 2.
- Posterior 95th percentile of Λ is 2.4.

The plot below shows the cumulative distribution function. The median is shown as a red dashed line; the 95th percentile is shown as a blue dotted/dashed line.



***Problem 1e*:**

Find the predictive distribution for the number of cars in the next 15-second interval. Find the predictive probability that 0, 1, 2, 3, 4, and more than 4 cars will pass the point in the next 15 seconds.

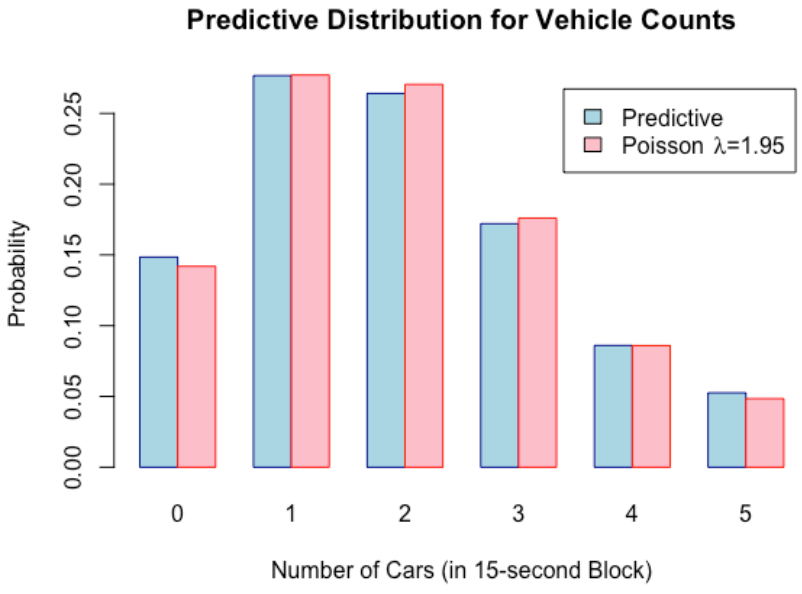
Solution.

To find the predictive probability of k cars in the next 15 seconds, we compute the Poisson probability of k cars given each value of λ, multiply by the posterior probability of λ, and sum for all values of λ. R code for this is posted on Blackboard.

The predictive probabilities are:

Number of Events	Probability
0	0.148
1	0.277
2	0.264
3	0.172
4	0.086
5+	0.053

The predictive distribution is plotted below. For comparison, a Poisson distribution with λ=1.95 (the posterior expected value) is plotted also. Note that the distributions are nearly identical, although the Bayesian predictive distribution is ever so slightly more spread out (larger probabilities for extreme values, smaller probabilities for central values).



Discussion: From this analysis, we see that a Poisson distribution provides a good model for counts of cars passing by this point on the roadway. The rate is approximately 2 cars per 15 seconds. There is some uncertainty about this rate. There is a 95% chance the rate is less than or equal to 2.4, and there is about a 98% probability that the rate lies between 1.4 and 2.6 cars per second. As we gather more data, we get more information about the rate, narrowing the range of uncertainty.