***Problem 1*:**

This problem continues analysis of the automobile traffic data from Assignment 3. Transforming the arrival times to counts of cars in each 15-second interval gives the following table of counts:

Number of Cars	Number of Occurrences
0	3
1	5
2	7
3	3
4	3
5 or more	0

- Assume a Poisson likelihood and a uniform prior distribution for the unknown rate Λ. Find the posterior distribution for Λ.[1]
- Find the mean, standard deviation, median and mode of the posterior distribution.
- Find a 95% symmetric tail area credible interval for Λ.
- Compare your results with the results from Assignment 3.

***Solution*:**

The prior distribution is a Gamma(α,β) distribution with α=1 and β=∞. We observed $\Sigma_i\, x_i$ = 40 cars passing in n = 21 15-second time intervals. Therefore, the posterior distribution is a Gamma($\alpha*$,$\beta*$) distribution with shape $\alpha* = \alpha+\Sigma_i\, x_i = 1+40 = 41$ and scale $\beta* = 1/(\beta^{-1}+n) = 1/(0+21) = 0.0476$. (Here we are using the fact that $1/\infty=0$.)

If I say "find the distribution," and the distribution is a member of a parametric family, you should name the parametric family, and you should give the parameters. If you just plot the posterior density function without saying it is a Gamma distribution, or what the posterior hyperparameters are, then you will not receive full credit.

Remember, we update the shape by adding the prior shape plus the number of events occurring. For this problem, events are cars passing, and there were 40 cars, so this is $\alpha*=\alpha+40$. We update the scale by adding the inverse prior scale and the number of time periods, and then inverting. There were 21 time periods so this is $\beta* = 1/(\beta^{-1}+21)$.

- The posterior mean is $\alpha*\beta* = 41/21 = 1.9524$.
- The posterior standard deviation is $\sqrt{\alpha*(\beta*)^2} = \sqrt{14}/21 = 0.3049$.
- The median of the posterior distribution is 1.9365, the 0.5 percentage point of the Gamma(41, 1/21) distribution.

[1] *Note: a uniform prior distribution $g_u(\lambda)$ would assign equal density to all values of λ. Actually, because integrating any positive constant over the real line yields ∞, there is no such distribution. Still, it is convenient to use a prior distribution that is essentially uniform on the range of values where the likelihood is non-negligible. The uniform distribution is the limit as the scale β tends to ∞ of a Gamma(1, β) distribution.*

- The mode of the posterior distribution is $(\alpha^* - 1)\ \beta^* = 40/21 = 1.9048$.
- The endpoints of the posterior credible interval are the 0.025 and 0.975 quantiles of the Gamma distribution, which we can find in R using the `qgamma` function.
  - 0.025 quantile: 1.4011
  - 0.975 quantile: 2.5937

Note: I am reporting these values to 4 significant figures to allow you to check your calculations. We would rarely report 4 significant figures to a decision maker. I would report a posterior mean of 0.61, a posterior standard deviation of 0.06, a posterior median of 0.61, and a posterior mode of 0.61.

Comparing with Assignment 3, we have the following table:

	Posterior mean	**Posterior SD**	**Posterior mode**	**Posterior median**	**Posterior 95% interval**
Discretized	1.95	0.305	2	2	[1.4, 2.6]
Continuous	1.95	0.305	1.9	1.94	[1.40, 2.59]

These values are almost identical. The slight differences in the median and mode are clearly artifacts of discretization.

After observing 40 cars traveling past this location, our expected value for the rate is about 1.95 cars every 15 seconds (or 7.8 cars per minute), and we are 95% sure that the rate is between about 1.4 and 2.6 cars every 15 seconds.

***Problem 2*:**

Suppose a highway engineer provided the following prior judgments about the rate of traffic on the stretch of highway (before seeing the data).

- The rate is equally likely to be above or below 15 cars per minute (or 3.75 cars every 15 seconds).
- There is a 90% chance that the rate is fewer than 28 cars per minute (or 7 cars every 15 seconds).
- There is a 90% chance that the rate is greater than 6 cars per minute (or 1.5 cars every 15 seconds).

Find a Gamma prior distribution that matches these judgments as well as possible. What are the parameters of this Gamma distribution? How well does it match the engineer's judgments? Comment on whether you think it is reasonable to use this Gamma distribution as a prior distribution for the Poisson rate parameter.

***Solution*:**

Our objective is to find a Gamma distribution that fits these values as well as possible. That is, we want to find parameters α and β such that:

- The 10th percentile of a Gamma(α, β) is approximately 1.5
- The 50th percentile of a Gamma(α, β) is approximately 3.75
- The 90th percentile of a Gamma(α, β) is approximately 7

The most straightforward way to do this is to try some values of α and β and look for values that fit these quantiles as well as possible. We can get a rough idea of the range of the parameters by recognizing that we are looking for a distribution with center somewhere in the neighborhood of 4. So we can try some different values of α, and set β to $4/\alpha$, and then home in from there. Increasing β will make the distribution more spread out. Increasing α will make the distribution more symmetrical.

In the R file provided on Blackboard, I created a brute force algorithm that works as follows:

- Loop through many values of α. I started at 0.5 and went through 7, with 200 equally spaced values in between. (If you do this and discover the
- Loop through many values of β. I started at 0.5 and went through 10, with 200 equally spaced values in between.
- For each α and β, find the Gamma quantiles using
  `qgamma(c(0.1,0.5,0.9), alpha, beta)`
- Compare these with the expert quantiles. I computed the sum of squared differences for an overall measure of how close they are – values closer to zero mean a better fit.
- Choose the α and β that have the smallest sum of squared differences.

Doing this, I got values $\alpha = 3.34$ and $\beta = 1.22$. (The choice or range to search for α and β is a judgment call. If you follow this method and the best value is at the extreme of the range you chose, then extend the range and try again to look for better values outside the range.)

We can also do this using optimization. This is easy in Excel. The left-hand panel of the figure below is a screenshot of my starting point. There are cells for the initial α and β (I initialized both to 1). I listed the probability points to evaluate, entered the data given by the expert, and used the GAMMA.INV function to calculate the quantiles. Then I calculated the squared differences between theoretical and expert-provided, and added them up. For $\alpha = 1$ and $\beta = 1$, the sum of squared differences was 33.36.

After setting this up, I invoked Excel Solver. If you have it installed, you invoke it from the Tools menu. I set the objective to cell D10, chose "Min", and set the changing variable cells to \$B\$3:\$B\$4 (the α and β values). The right-hand panel shows the final

values. Excel found values $\alpha = 3.32$ and $\beta = 1.22$, for a sum of squared differences of 0.018. These values are very close to those from the brute-force method.

	A	B	C	D	E
1	Homework 4 - Fit Gamma distribution to expert quantiles				
2					
3	alpha	1			
4	beta	1			
5					
6	Prob	Ex Quantile	Th Quantile	sq.diff	
7	0.1	1.5	0.10536052	1.94501929	
8	0.5	3.75	0.69314718	9.34434916	
9	0.9	7	2.30258509	22.0657068	
10			sum	33.3550753	
11					

	A	B	C	D	E
1	Homework 4 - Fit Gamma distribution to expert quantiles				
2					
3	alpha	3.31832432			
4	beta	1.22053921			
5					
6	Prob	Ex Quantile	Th Quantile	sq.diff	
7	0.1	1.5	1.58734343	0.00762888	
8	0.5	3.75	3.65137206	0.00972747	
9	0.9	7	7.03149525	0.00099195	
10			sum	0.0183483	
11					

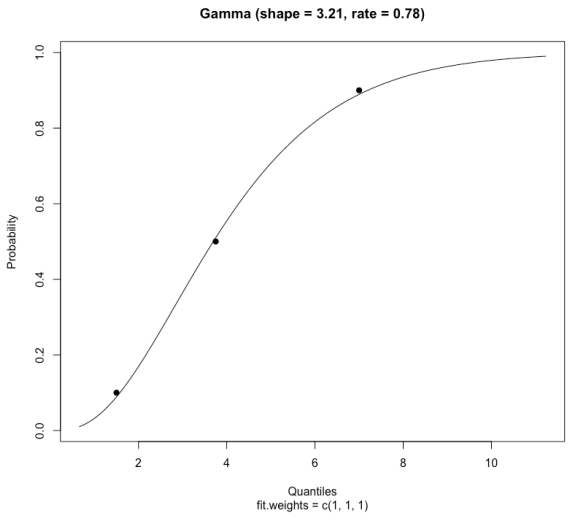
We can also do the optimization in R. There is a built-in function called `optim`, which takes 2 arguments: initial parameter values and a function to be optimized. I created a function that calculates the sum of squared differences and passed it to the optimizing function. By default, `optim` does minimization, which is what we want for this problem, and there is a control parameter that allows it to do maximization. Using `optim` gives the same results as Excel Solver.

The optimal parameter values provide a pretty good fit to the expert's judgments, so it seems quite reasonable to use the Gamma(3.3, 1.2) distribution as a prior distribution for the Poisson rate parameter.

Finally, we can use the `get.gamma.par` function in the `rriskdistribution` package (this function actually calls `optim`). Doing this, we enter:

```
get.gamma.par(c(0.1,0.5,0.9), c(1.5, 3.75, 7))
```

The values returned are shape 3.2 and rate 0.788, or scale 1.28. They are pretty close to the values I found by the other methods. The difference is due to the discrepance measure used by `get.gamma.par`. An advantage of using this approach is that it also plots the theoretical and expert-provided quantiles for a visualization of how well the distribution fits.



This problem does not have a single correct answer, but your shape and scale values should have been fairly close to the ones I found.

***Problem 3*:**

Repeat Problem 1 using the prior distribution from Problem 2. Compare your results with Problem 1.

***Solution*:**

The prior distribution is a Gamma(α,β) distribution with α=3.3 and β=1.2. We observed $\Sigma_i\ x_i$ = 40 cars passing in n = 21 15-second time intervals. Therefore, the posterior distribution is a Gamma($\alpha*$,$\beta*$) distribution with shape $\alpha* = \alpha+\Sigma_i\ x_i$ = 3.3+40 = 43.3 and scale $\beta* = 1/(\beta^{-1}+n) = 1/(1/2.1+21) = 0.046$.

- The posterior mean is $\alpha*\beta* = 43.3 \times 0.046 = 1.99$.
- The posterior standard deviation is $\sqrt{\alpha*(\beta*)^2} = \sqrt{43.3} \times 0.046 = 0.302$.
- The median of the posterior distribution is 1.97, the 0.5 percentage point of the Gamma(41, 1/201) distribution.
- The mode of the posterior distribution is $(\alpha* - 1)\ \beta* = (43.3-1)/21.0001 = 1.94$.
- The endpoints of the posterior credible interval are the 0.025 and 0.975 quantiles of the Gamma distribution, which we can find in R using the `qgamma` function.
  - 0.025 quantile: 1.44
  - 0.975 quantile: 2.62

A comparison of the prior distribution and the two posterior distributions is given in the table below. Note that because we used a uniform prior distribution in Problem 1, the posterior distribution is the same as the normalized likelihood.

	Mean	**SD**	**Mode**	**Median**	**95% Interval**
Prior	4.06	2.22	2.85	3.67	[0.94, 9.44]
Norm.Lik	1.95	0.305	1.90	1.94	[1.40, 2.59]
Posterior	1.99	0.302	1.94	1.97	[1.44, 2.62]

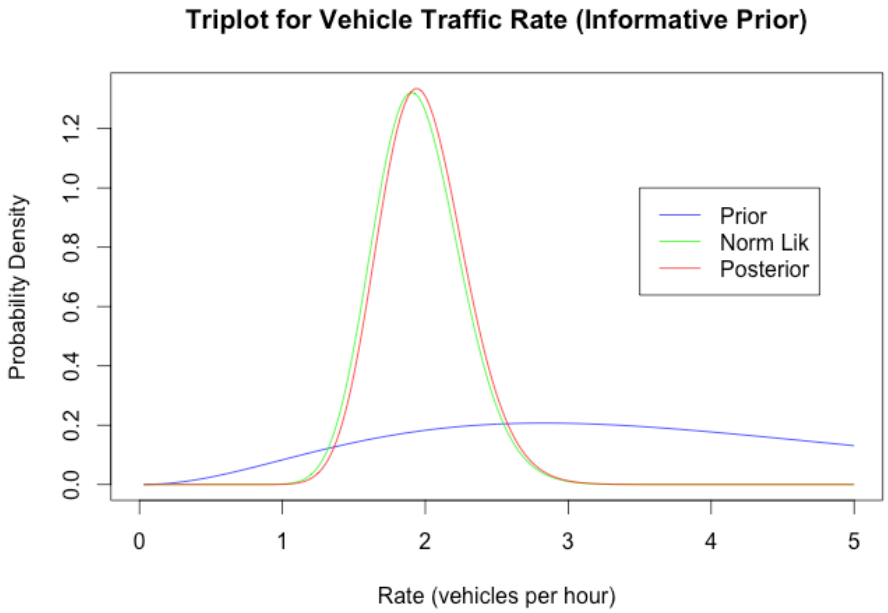
We can see immediately that the normalized likelihood and the posterior distribution from Problem 3 are very similar, whereas the prior distribution is very spread out. The posterior distribution shifts the normalized likelihood very slightly toward higher values, because the prior distribution favors larger values than the likelihood. The posterior distribution is also slightly more concentrated (slightly smaller standard deviation and slightly narrower credible interval) than the normalized likelihood, reflecting information contributed by the prior distribution.

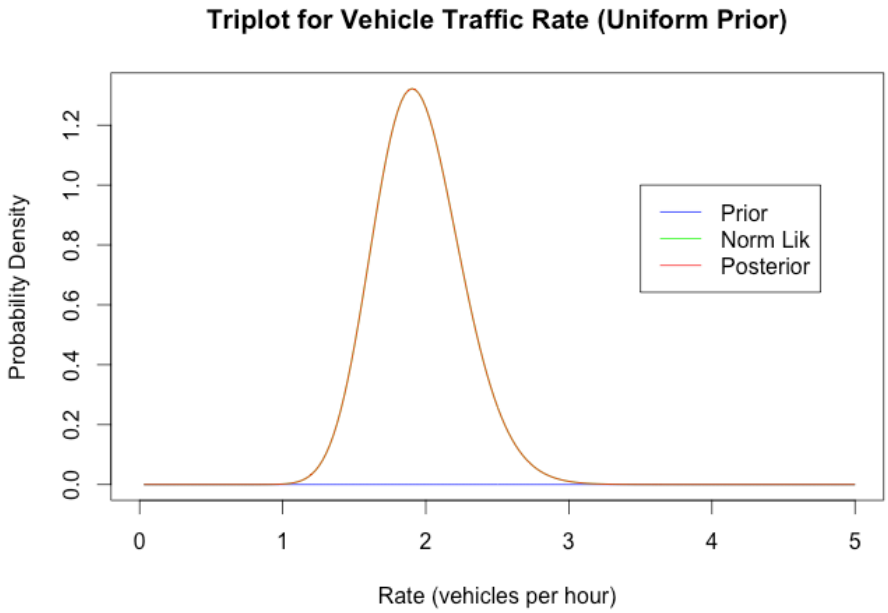
***Problem 4*:**

Make a triplot of the prior distribution, normalized likelihood and posterior distribution for Problem 3. Discuss the plot.

***Solution*:**

The plot is shown below. For comparison purposes, I also include a triplot for Problem 1, although the problem did not ask for it. Notice that for Problem 1, the posterior and the normalized likelihood are essentially identical. This is because we used a nearly uniform prior distribution.





We see from the second plot that the posterior distribution and the normalized likelihood are essentially identical. This is because when we use an essentially uniform prior, the likelihood is proportional to the posterior distribution, and so dividing either by its integral gives the same result.

Looking at the triplot with the informative prior distribution, we see that the prior distribution is fairly flat in the region of highest likelihood, but weakly favors larger rates. This pushes the posterior distribution slightly to the right of the normalized likelihood, although because of the nearly flat prior, there is not much difference between the prior and the normalized likelihood.

Note: for a continuous distribution, you should not use a bar plot. You should do a line plot like this one.