MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2008-9

3 Parameters and likelihood

3.1 Introduction

Now, from where did I get the probabilities in table 1? Well, I could have thought of them directly. There are, after all, only four possible outcomes. However, in fact I thought less directly. Especially when the number of possible outcomes starts to increase it becomes helpful to structure our thoughts by introducing *parameters*. We also often make use of ideas such as that all individuals in a group appear, in some way, "the same" to us.

Introduce unknowns T_1, T_2 such that $T_i = 1$ if S_i and $T_i = 0$ otherwise (i.e. if $\bar{S}_i$, where this denotes "not S_i"). Consider a quantity θ such that, *if we knew the value of* θ, then $\Pr(S_1) = \Pr(S_2) = \theta$ and T_1, T_2 are *independent* given θ. We can think of θ as the unknown proportion of animals that have the gene. We can represent this situation in a diagram or *graphical model*. There are various kinds of graphical model but the diagram shown here is a *directed acyclic graph* or DAG. The nodes are connected by directed arrows (called *arcs* or *edges*) and it is impossible to find a path along the directions of the arrows which returns to where you started. The term *influence diagram* is also sometimes used for such graphs. See figure 2.

In this case we would say that, in terms of possessing the gene, the animals are *exchangeable.* That is T_1 and T_2 are exchangeable.

Suppose that, before we observe either animal, we believe that θ can have only two values, 0.1 and 0.4, and we give these probabilities of 2/3 and 1/3 respectively. This is called a *prior* probability distribution for θ. (It is, of course, unrealistic to allow only two values but we will put this right very soon).

	$\theta = 0.1$	$\theta = 0.4$
Prior probs.	$\pi_1 = 2/3$	$\pi_2 = 1/3$
$\Pr(S\|\theta)$	0.1	0.4

We can now easily work out the joint probabilities for θ, T_1, T_2.

$$\Pr(\theta, T_1, T_2) = \Pr(\theta)\Pr(T_1|\theta)\Pr(T_2|\theta)$$

Notice how this relates to the diagram.
We can find the probabilities in the original table by summing over the distribution of θ.

$$\begin{aligned}
\Pr(S_1 \wedge S_2) &= (2/3) \times 0.1^2 + (1/3) \times 0.4^2 = 0.06 \\
\Pr(S_1 \wedge \bar{S}_2) &= (2/3) \times 0.1 \times 0.9 + (1/3) \times 0.4 \times 0.6 = 0.14 \\
\Pr(\bar{S}_1 \wedge S_2) &= (2/3) \times 0.9 \times 0.1 + (1/3) \times 0.6 \times 0.4 = 0.14 \\
\Pr(\bar{S}_1 \wedge \bar{S}_2) &= (2/3) \times 0.9^2 + (1/3) \times 0.6^2 = 0.66 \\
\Pr(S_1) &= (2/3) \times 0.1 + (1/3) \times 0.4 = 0.2
\end{aligned}$$

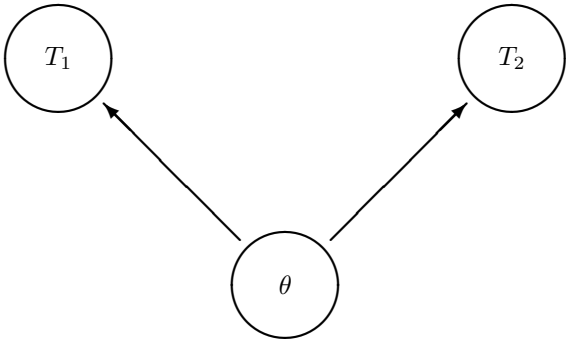

Figure 2: Graphical model for animals example

Now, if we observe the first animal, we can apply Bayes' rule to find the *posterior* probabilities for the values of θ.

$$\begin{aligned} \Pr(\theta = 0.1|S_1) &= \frac{\pi_1 \times 0.1}{\pi_1 \times 0.1 + \pi_2 \times 0.4} \\ &= \frac{(2/3) \times 0.1}{(2/3) \times 0.1 + (1/3) \times 0.4} = \frac{1}{3} \end{aligned}$$

Similarly $\Pr(\theta = 0.4|S_1) = 2/3$. In fact we can deduce the following simple form of Bayes' rule.

"Posterior $\propto$ Prior $\times$ Likelihood"

Here the "likelihood" is $\Pr(S_1|\theta)$.
Now the information propagates through θ to T_2 and the probability of S_2 becomes $(1/3) \times 0.1 + (2/3) \times 0.4 = 0.3$.

3.2 A slightly less simple example

We keep the same example except that this time there are twenty animals. We can use the same diagram as above except that there are twenty nodes $T_1, \ldots, T_{20}$.

What happens if we observe 3 animals with the gene out of twenty observed?

The likelihood is proportional to $\theta^3(1-\theta)^{17}$. The likelihood function, scaled so that its maximum value is 1.0, is shown in figure 3. The maximum of this function occurs at $\theta = 3/20$.

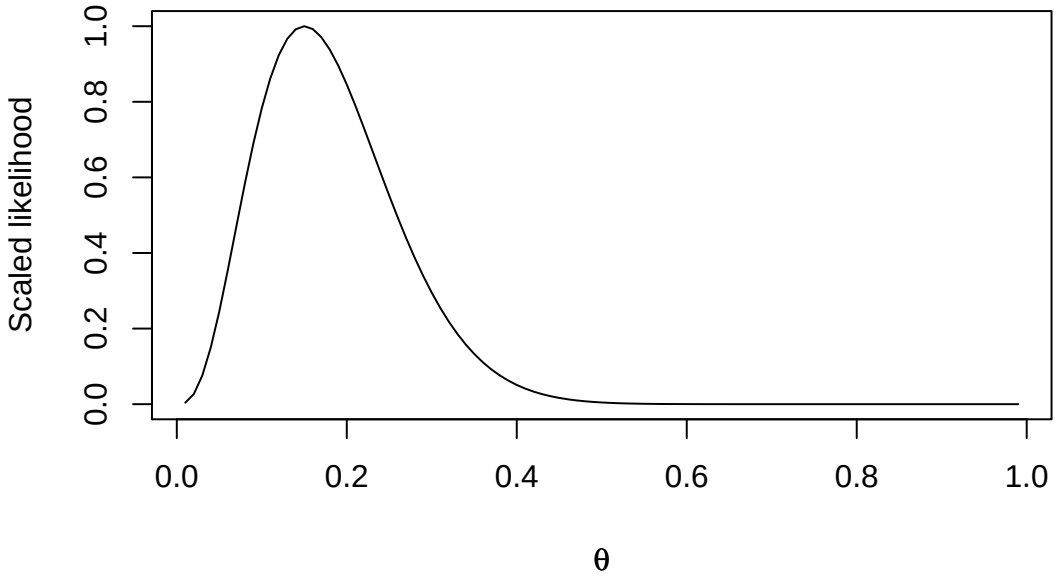

Figure 3: Scaled likelihood in the animals' gene example, with 3 animals out of twenty observed with the gene.

The posterior probabilities are proportional to:

$$\begin{aligned}(2/3) \times 0.1^3 0.9^{17} &= 1.1118 \times 10^{-4} \\ (1/3) \times 0.4^3 0.6^{17} &= 3.61 \times 10^{-6}\end{aligned}$$

So the posterior probability that $\theta = 0.1$ is

$$\Pr(\theta = 0.1|D) = \frac{1.1118}{1.1118 + 0.0361} = 0.969$$

We can now find predictive probabilities for future animals. E.g. the probability that two new animals both have the gene:
$0.969 \times 0.1^2 + 0.031 \times 0.4^2 = 0.01465$.
In effect we are adding two extra nodes, T_{21}, T_{22} to our diagram but not observing these.

3.3 The continuous form of Bayes' theorem

So far, in our animals' gene example, our prior probability distribution for θ, the proportion of animals with the gene, has been rather unrealistic. We are not likely, in practice, to believe that there are only two possible values. It would be more realistic to suppose that any value between 0 and 1 was possible. This requires a continuous prior distribution for θ. How does this affect the application of Bayes' theorem?

If our prior pdf for θ is $f^{(0)}(\theta)$ and our likelihood (That is the conditional probability of the data given θ) is $L(\theta)$ then our posterior pdf for θ (the conditional pdf of θ given the data) is

$$f^{(1)}(\theta) = \frac{f^{(0)}(\theta)L(\theta)}{\int f^{(0)}(\theta)L(\theta)\ d\theta}. \tag{1}$$

This is the form of Bayes' theorem which is used when the unknown is continuous. We see that, once again

$$\text{Posterior} \ \propto \ \text{Prior} \ \times \ \text{Likelihood}$$

where, this time, "Prior" and "Posterior" are probability density functions.

Proof: Suppose we divide the range of our unknown θ into short intervals of equal length, $\delta\theta$. Thus interval i is $[\theta_i,\ \theta_{i+1})$ and $\theta_{i+1} = \theta + \delta\theta$. Suppose that our prior probability that $\theta_i \leq \theta < \theta_{i+1}$ is $p_i^{(0)} = F^{(0)}(\theta_{i+1}) - F^{(0)}(\theta_i)$ where $F^{(0)}(\theta)$ is the prior distribution function of θ. Let the likelihood, evaluated at $\theta = \theta_i$, be $L(\theta_i)$. Then, by Bayes' theorem, the posterior probability that $\theta_i \leq \theta < \theta_{i+1}$ is given approximately by

$$p_i^{(1)} \approx \frac{p_i^{(0)}L(\theta_i)}{\sum_j p_j^{(0)}L(\theta_j)}$$

where the sum in the denominator is taken over all of the intervals.

Now, if $f^{(0)}(\theta)$ and $f^{(1)}(\theta)$ are respectively the prior and posterior probability density functions, then $p_i^{(0)} \approx f^{(0)}(\theta_i)\delta\theta$ and $p_i^{(1)} \approx f^{(1)}(\theta_i)\delta\theta$. So

$$f^{(1)}(\theta_i)\delta\theta \approx \frac{f^{(0)}(\theta_i)L(\theta_i)\delta\theta}{\sum_j f^{(0)}(\theta_j)L(\theta_j)\delta\theta}$$

and, dividing both sides by $\delta\theta$,

$$f^{(1)}(\theta_i) \approx \frac{f^{(0)}(\theta_i)L(\theta_i)}{\sum_j f^{(0)}(\theta_j)L(\theta_j)\delta\theta}.$$

Now consider what happens if we increase the number of intervals and let $\delta\theta \to 0$. Informally it is easy to see that, in the limit, we obtain (1).

3.4 The example with a continuous prior distribution

More realistically we might have a continuous prior probability distribution for θ.

Prior p.d.f. $f^{(0)}(\theta)$.

Posterior p.d.f.:

$$\frac{f^{(0)}(\theta)\theta^3(1-\theta)^{17}}{\int_0^1 f^{(0)}(\theta)\theta^3(1-\theta)^{17}.d\theta}$$

The integration is straightforward in this case. In more complicated cases, especially when there are many unknowns, the integration has been the main computational problem. In the past this was a major obstacle to the use of Bayesian inference. In recent years huge progress has been made and we can now routinely handle complicated problems with very large numbers of unknowns.

In fact the calculation is particularly simple if we use a *conjugate* prior distribution. This simply means a prior distribution which "matches" the likelihood in the sense that the posterior distribution belongs to the same family as the prior distribution. In this example the conjugate family is the family of beta distributions. The prior p.d.f. is proportional to

$$\theta^{a-1}(1-\theta)^{b-1}$$

where we specify the values of a and b.
The posterior p.d.f. is then proportional to

$$\theta^{a+3-1}(1-\theta)^{b+17-1}.$$

The *prior mean* is

$$\frac{a}{a+b}$$

and the *posterior mean* is

$$\frac{a+3}{a+b+20}.$$

We might, of course, feel that our prior beliefs are not represented by a conjugate distribution. In this case we could use a different distribution and employ numerical methods of integration. We might however be able to use a *mixture* of conjugate prior distributions to approximate the shape we want. The posterior is then a mixture of the conjugate posterior distributions, although the weights are changed. We will consider this in more detail in a later lecture.

4 Bayes' Rule

4.1 Theory

In general we can use Bayes' rule to change our *prior probability distribution,* which expresses our beliefs about parameters before we see the data, to a *posterior probability distribution* representing beliefs about the parameters given the data.

Suppose we have a prior probability density function for a vector θ of parameters, $f_\theta(\theta)$. Suppose the p.d.f. for a vector $\underline{Y}$ of observations *given* $\underline{\theta}$ is $f_{Y|\theta}(\underline{y} \mid \underline{\theta})$. This latter p.d.f. is treated as a function of $\underline{\theta}$ once $\underline{y}$ is observed and is called the *likelihood.*

Then our *posterior p.d.f.* is

$$f_{\theta|y}(\underline{\theta} \mid \underline{y}) = \frac{f_\theta(\underline{\theta}) f_{Y|\theta}(\underline{y} \mid \underline{\theta})}{\int_\theta f_\theta(\underline{\theta}) f_{Y|\theta}(\underline{y} \mid \underline{\theta}).d\theta}.$$

We can think of this as

$$\Pr(\underline{\theta} \mid \underline{y}) = \frac{\Pr(\underline{\theta}) \Pr(\underline{y} \mid \underline{\theta})}{\Pr(\underline{y})}.$$

Often it is sufficient to write

$$f_{\theta|y}(\underline{\theta} \mid \underline{y}) \propto f_\theta(\underline{\theta}) f_{Y|\theta}(\underline{y} \mid \underline{\theta}).$$

That is **posterior $\propto$ prior $\times$ likelihood.**

When $\underline{Y}$ or $\underline{\theta}$ is discrete rather than continuous, the probability density functions are replaced by probabilities as appropriate.

4.2 Example: The rate of a Poisson process

We wish to model the occurrence of events in time as a Poisson process with rate λ. (This is a model for the times when events occur if the events occur "at random." The rate is the average number per unit time).

Suppose our prior probability density function for λ is proportional to

$$\lambda^{\alpha-1} e^{-\beta\lambda}.$$

This means that our prior distribution for λ is a gamma distribution. The probability density function for a gamma(α, β) distribution is

$$f(x) = \begin{cases} 0 & (x < 0) \\ \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & (x \geq 0) \end{cases}$$

where $\Gamma(y)$ is the *gamma function* which has the property that $\Gamma(y) = (y-1)\Gamma(y-1)$ and, if n is a positive integer, $\Gamma(n) = (n-1)!$.

We observe the process for τ time units. Events occur at times $t_1, t_2, \ldots, t_n$. Writing $t_0 = 0$, the likelihood is

$$L = \prod_{i=1}^{n} \{\lambda e^{-\lambda(t_i - t_{i-1})}\} e^{-\lambda(\tau - t_n)}.$$

The last term is for the probability that no events occur in (t_n, τ).

We can simplify L to

$$L = \lambda^n e^{-\lambda\tau}.$$

This is effectively the probability of n events in $(0, \tau)$, i.e.

$$(\lambda\tau)^n e^{-\lambda\tau}/n! \propto \lambda^n e^{-\lambda\tau}.$$

Hence the posterior p.d.f. is proportional to

```
12.40   3, 6, 9,15,24,28,30          12.50   4,30,34,47
12.41   7,12,14,16,21,24,30,50      12.51   0, 4,18,21,52,57,59
12.42   9,22,28,46,53               12.52   4, 9,29,30,31,38,59
12.43  22,25,35,38,58               12.53   2, 6,31,53
12.44   2, 5, 8,10,14,17,27,30,45   12.54   7, 9,13,24,39,47
12.45   3,46                        12.55  28,46,49,59
12.46  13,42,51                     12.56   6, 9,14,35,41,46
12.47   0, 9,11,18,23,26,39,51      12.57   8,10,15,22,25,49,52,59
12.48  35,39,55                     12.58  31,34,53,55,56
12.49   8,10,19,20,33,45,56,58      12.59   2,31,34,38,54,59
```

Table 2: Times of arrival of motor vehicles

3	3	3	6	9	4	2	37	5	2	2	5	3	6	20	19	13	6	18	7
29	3	10	3	20	4	3	3	2	4	3	10	3	15	18	43	27	29	9	9
9	2	7	5	3	13	12	44	4	16	13	2	9	1	13	12	11	2	6	26
4	13	13	4	14	3	31	5	2	5	5	20	1	1	7	21	3	4	25	22
14	2	4	11	15	8	41	18	3	10	7	3	5	21	6	5	22	2	5	7
3	24	3	7	32	3	19	2	1	6	29	3	4	16	5					

Table 3: Inter arrival times of motor vehicles

$$\lambda^{\alpha+n-1}e^{-(\beta+\tau)\lambda}.$$

This is clearly a gamma distribution so the posterior p.d.f. is

$$\frac{(\beta+\tau)^{\alpha+n}\lambda^{\alpha+n-1}e^{-(\beta+\tau)\lambda}}{\Gamma(\alpha+n)}.$$

The posterior mean is $(\alpha+n)/(\beta+\tau)$ and the posterior variance is $(\alpha+n)/(\beta+\tau)^2$.

4.3 Conjugate and non-conjugate priors

We are not, of course, restricted to using a prior of the form

$$\lambda^{\alpha-1}e^{-\beta\lambda}.$$

This particular form works out neatly because it is the *conjugate* form of prior for this likelihood. If our beliefs can not be adequately represented by a conjugate prior we may resort to numerical evaluation of the posterior distribution. The normalising constant, to make the posterior p.d.f. integrate to 1, may be found, if necessary, by numerical integration. In complicated models numerical approaches are commonly used. Monte Carlo integration has become particularly popular since about 1990.

Another approach is to form the prior density as a *mixture* of conjugate prior densities with different parameters. This keeps the calculation relatively straightforward while allowing considerable flexibility.

4.4 Chester Road traffic

At one time I worked in an office overlooking Chester Road in Sunderland. The times of arrivals of motor vehicles passing a point in Chester Road going East, from 12.40 till 13.00 on Wednesday 30th September 1987 are given in Table 2.

These can be converted to time intervals between vehicles. These intervals, in seconds, are given in Table 3. (The first value is the time till the first arrival).

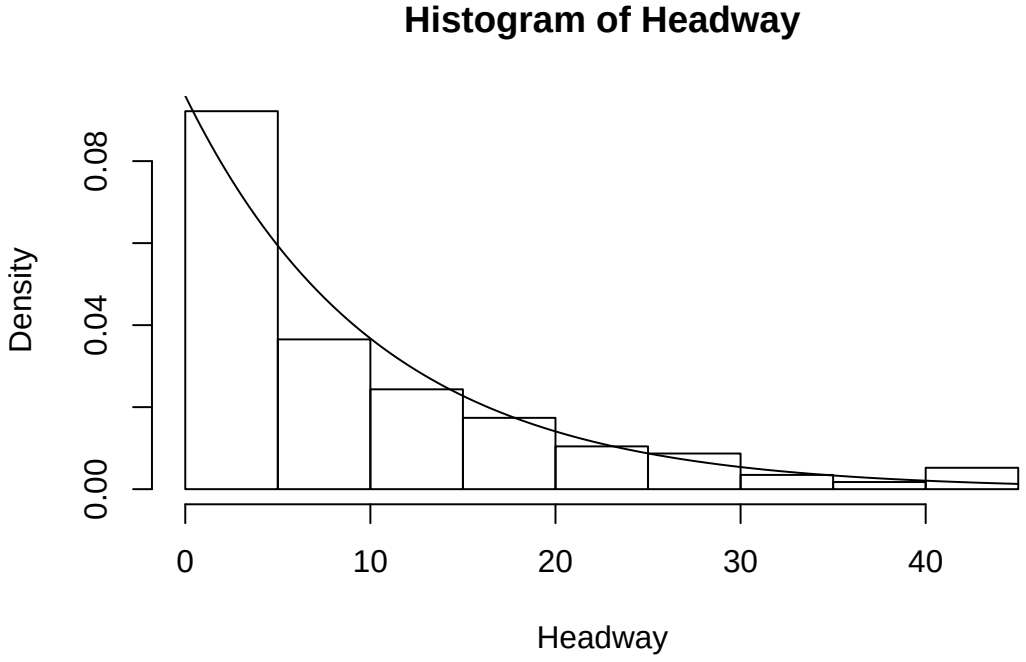

Figure 4: Inter arrival times (115 observations) with negative exponential p.d.f.

A histogram of these data is shown in Figure 4 together with a plot of the p.d.f. of a negative exponential distribution, the parameter of which was chosen so that the distribution had the same mean as the observed data. The fit appears satisfactory.

No evidence of nonzero correlations between successive inter-arrival times was found, either by plotting t_i against t_{i-1} or by estimating the correlation coefficients. Similarly a plot of t_i against i showed no obvious pattern.

Suppose the prior expectation for the rate of arrival was, say, 1 vehicle every 5 seconds, i.e. $\lambda_0 = 0.2$. Suppose we were almost certain that the rate would be less than 1 vehicle every 2 seconds. Let us say that $\Pr(\lambda < 0.5) = 0.99$. Suppose we use the conjugate (gamma) prior. Then $\alpha/\beta = 0.2$.

We require

$$\int_0^{0.5} f(\lambda).d\lambda = 0.99$$

where $f(\lambda)$ is the p.d.f. of a gamma$(\alpha, \alpha/0.2)$ distribution. We can evaluate this integral in R.

First we set up a vector of values for α and a corresponding vector for β. Then we evaluate the integral for each pair of values.

```
> alpha<-seq(1,10)
> beta<-alpha/0.2
> prob<-pgamma(0.5,alpha,beta)
> d<-data.frame(alpha,beta,prob)
> d
   alpha beta      prob
1      1    5 0.9179150
2      2   10 0.9595723
3      3   15 0.9797433
4      4   20 0.9896639
5      5   25 0.9946545
```

```
6      6   30 0.9972076
7      7   35 0.9985300
8      8   40 0.9992214
9      9   45 0.9995856
10    10   50 0.9997785
```

We find that we get approximately what we require with $\alpha = 4,\ \beta = 20$.

Now $\tau = 1200$ and $n = 115$.
Thus the posterior p.d.f. is

$$\frac{1220^{119}}{118!}\lambda^{118}e^{-1220\lambda}.$$

The prior mean was 0.2. The posterior mean is

$$\frac{119}{1220} = 0.0975.$$

The prior variance was 0.01. The posterior variance is

$$\frac{119}{1220^2} = 0.00008.$$

We can evaluate the posterior pdf and distribution function using R. The posterior standard deviation is

$$\frac{\sqrt{119}}{1220} = 0.00894.$$

At a guess the important part of the posterior distribution is likely to be within ± 3 standard deviations of the posterior mean. That is roughly 0.07 to 0.13 so we create an array of λ values from 0.070 to 0.130 in steps of 0.001. We can check that this covers the important part of the distribution by looking at the distribution function. (In this case we find that very little of the probability is outside the range).

```
> lambda<-seq(0.07,0.13,0.001)
> pdf<-dgamma(lambda,119,1220)
> cdf<-pgamma(lambda,119,1220)
> plot(lambda,pdf,type="l")
> plot(lambda,cdf,type="l")
```

Figure 5 shows the prior density the (scaled) likelihood and the posterior density. We see that the posterior density is only slightly different from the likelihood, the difference being due to the effect of the prior distribution.

We can also use the distribution function to see, for example, that, in the posterior distribution,

$$\Pr(0.09 < \lambda < 0.11) = 0.91440 - 0.20176 = 0.71264.$$

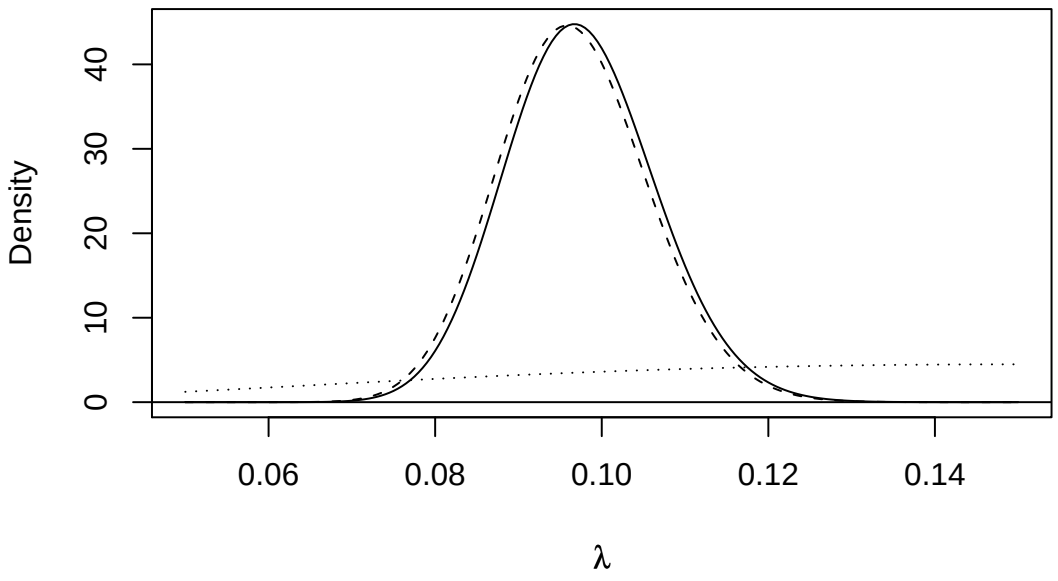

Figure 5: Chester road traffic arrival rate. Dots: prior pdf, Dashes: scaled likelihood, Solid line: posterior pdf.

4.5 Probability intervals

Probability interval A probability interval is an interval for an unknown quantity which contains a given amount of probability. Probability intervals can, of course, be *prior* probability intervals, based on the prior distribution, or *posterior* probability intervals, based on the posterior distribution. (Later we will see that we can also have *predictive* probability intervals).

Example: There is an example above of a probability interval for λ in the Chester Road example.

Symmetric probability interval A *symmetric* probability interval for an unknown quantity θ is an interval $t_1 < \theta < t_2$ which has the property that $\Pr(\theta < t_1) = \Pr(\theta > t_2)$. For example, a 95% symmetric probability interval (t_1, t_2) for θ would have the property that $\Pr(\theta < t_1) = 0.025,\ \Pr(t_1 < \theta < t_2) = 0.95,\ \Pr(\theta > t_2) = 0.025$.

Example: In the Chester Road example above, the posterior distribution for λ is gamma$(119, 1220)$. We can calculate a 95% symmetric posterior probability interval for λ as follows, using R.

```
> qgamma(0.025,119,1220)
[1] 0.08080462
> qgamma(0.975,119,1220)
[1] 0.1158291
```

Thus our interval is $0.0808 < \lambda < 0.1158$.

Highest probability density (hpd) interval A *hpd* or "highest probability density" interval for θ is a probability interval for θ with the property that no point outside the interval has a probability density greater than any point inside the interval. In other words, the interval captures the "most likely" values of the unknown.

It is easy to see that, of all possible $100\alpha\%$ probability intervals for θ, the hpd interval is the shortest.

It is also easy to see that, if the pdf of θ is $f(\theta)$ and the $100\alpha\%$ hpd interval is (t_1, t_2), then t_1, t_2 satisfy the following two properties.

$$\begin{aligned} \int_{t_1}^{t_2} f(\theta)\, d\theta &= \alpha \\ f(t_1) &= f(t_2) \end{aligned}$$

The first of these is true for any $100\alpha\%$ probability interval. The two together make the interval a hpd interval.

If the density $f(\theta)$ is unimodal and symmetric then the symmetric interval and the hpd interval coincide. Otherwise they do not.

Example: Finding a hpd interval in a non-symmetric distribution is not straightforward and we will leave the question of how to do it to a later lecture. A 95% posterior hpd interval for λ in the Chester Road example is $0.0803 < \lambda < 0.1153$.