The Exponential Family (continued)

Definition (e.g., Bernardo and Smith, 1994): Given data y_1 (a sample of size 1) and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (marginal) sampling distribution $p(y_1|\theta)$ belongs to the k-**dimensional exponential family** if it can be expressed in the form

$$p(y_1|\theta) = c\, f_1(y_1)\, g_1(\theta)\, \exp\left[\sum_{j=1}^{k} \phi_j(\theta)\, h_j(y_1)\right] \quad (51)$$

for $y_1 \in \mathcal{Y}$ and 0 otherwise; if $\mathcal{Y}$ doesn't depend on θ the family is called **regular**.

The vector $[\phi_1(\theta), \ldots, \phi_k(\theta)]$ in (51) is called the **natural parameterization** of the exponential family.

In this case the **joint distribution** $p(y|\theta)$ of a **sample** $y = (y_1, \ldots, y_n)$ of size n which is conditionally IID from (51) (which also defines, as usual, the **likelihood function** $l(\theta|y)$) will be

$$\begin{aligned} p(y|\theta) &= l(\theta|y) = \prod_{i=1}^{n} p(y_i|\theta) \qquad (52) \\ &= c\left[\prod_{i=1}^{n} f_1(y_i)\right] [g_1(\theta)]^n \exp\left[\sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i)\right]. \end{aligned}$$

The Exponential Family (continued)

This leads to **another way** to define the exponential family: in (52) take $f(y) = \prod_{i=1}^{n} f_1(y_i)$ and $g(\theta) = [g_1(\theta)]^n$ to yield

Definition: Given data $y = (y_1, \ldots, y_n)$ (a conditionally IID sample of size n) and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (joint) sampling distribution $p(y|\theta)$ belongs to the k-**dimensional exponential family** if it can be expressed in the form

$$p(y|\theta) = c\, f(y)\, g(\theta)\, \exp\left[\sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i)\right]. \qquad (53)$$

Either way you can see that $\{\sum_{i=1}^{n} h_1(y_i), \ldots, \sum_{i=1}^{n} h_k(y_i)\}$ is a set of **sufficient** statistics for θ under this sampling model, because the likelihood $l(\theta|y)$ depends on y only through the values of $\{h_1, \ldots, h_k\}$.

Now here's the theorem about the conjugate prior: if the likelihood $l(\theta|y)$ is of the form (53), then in searching for a **conjugate** prior $p(\theta)$—that is, a prior of the same functional form as the likelihood—you can see directly what will work:

$$p(\theta) = c\, g(\theta)^{\tau_0}\, \exp\left[\sum_{j=1}^{k} \phi_j(\theta)\, \tau_j\right], \qquad (54)$$

for some $\tau = (\tau_0, \ldots, \tau_k)$.

With this choice the **posterior** for θ will be

$$p(\theta|y) = c\, g(\theta)^{1+\tau_0}\, \exp\left\{\sum_{j=1}^{k} \phi_j(\theta) \left[\tau_j + \sum_{i=1}^{n} h_j(y)\right]\right\}, \qquad (55)$$

which is indeed of the **same form** (in θ) as (53).

The Exponential Family (continued)

As a first example, with $s = \sum_{i=1}^{n} y_i$, the **Bernoulli/binomial** likelihood in (41) can be written

$$\begin{aligned} l(\theta|y) &= c\,\theta^s(1-\theta)^{n-s} \\ &= c\,(1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^s \\ &= c\,(1-\theta)^n \exp\left[s \log\left(\frac{\theta}{1-\theta}\right)\right], \end{aligned} \tag{56}$$

which shows (a) that this sampling distribution is a member of the **exponential family** with $k = 1$, $g(\theta) = (1-\theta)^n$, $\phi_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ (**NB** the natural parameterization, and the basis of **logistic regression**), and $h_1(y_i) = y_i$, and (b) that $\sum_{i=1}^{n} h_1(y_i) = s$ is sufficient for θ.

Then (54) says that the **conjugate prior** for the Bernoulli/binomial likelihood is

$$\begin{aligned} p(\theta) &= c\,(1-\theta)^{n\tau_0} \exp\left[\tau_1 \log\left(\frac{\theta}{1-\theta}\right)\right] \\ &= c\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \text{Beta}(\alpha,\beta) \end{aligned} \tag{57}$$

for some α and β, as we've already seen is **true**.

2.8 Integer-Valued Outcomes

Case Study: *Hospital length of stay for birth of premature babies.* As a small part of a study I worked on at the Rand Corporation in the late 1980s, we obtained data on a random sample of $n = 14$ women who came to a hospital in Santa Monica, CA, in 1988 to **give birth to premature babies**.

One (integer-valued) outcome of interest was
$y =$ **length of hospital stay** (LOS).

Here's a preliminary look at the data in an excellent **freeware statistical package** called R (see http://www.r-project.org/ for more details and instructions on how to **download** the package).

```
rosalind 77> R

R : Copyright 2001, The R Development Core Team
Version 1.2.1  (2001-01-15)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> y

 [1] 1 2 1 1 1 2 2 4 3 6 2 1 3 0

> sort( y )

 [1] 0 1 1 1 1 1 2 2 2 2 3 3 4 6

> table( y )

0 1 2 3 4 6
1 5 4 2 1 1
```

Poisson Modeling

```
> stem( y, scale = 2 )

  The decimal point is at the |

  0 | 0
  1 | 00000
  2 | 0000
  3 | 00
  4 | 0
  5 |
  6 | 0

> mean( y )

[1] 2.071429

> sd( y )

[1] 1.54244

> q( )

Save workspace image? [y/n/c]: y
rosalind 1777>
```

One possible model for non-negative integer-valued outcomes is the **Poisson distribution**

$$P(Y_i = y_i) = \left\{ \begin{array}{cc} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} & \text{for } y_i = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right\}, \qquad (58)$$

for some $\lambda > 0$.

As usual Maple can be used to work out the **mean** and **variance** of this distribution:

```
rosalind 78> maple

    |\^/|     Maple V Release 5 (University of California, Santa Cruz)
._|\|   |/|_. Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
 \  MAPLE  /  reserved. Maple and Maple V are registered trademarks of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.
```

Poisson Modeling (continued)

```
> assume( lambda > 0 );

> p := ( y, lambda ) -> lambda^y * exp( - lambda ) / y!;

                                                  y
                                             lambda  exp(-lambda)
                    p := (y, lambda) -> ---------------------
                                                 y!

> simplify( sum( p( y, lambda ), y = 0 .. infinity ) );

                                          1

> simplify( sum( y * p( y, lambda ), y = 0 .. infinity ) );

                                       lambda~

> simplify( sum( ( y - lambda )^2 * p( y, lambda ),
    y = 0 .. infinity ) );

                                       lambda~
```

Thus if $Y \sim \text{Poisson}(\lambda), E(Y) = V(Y) = \lambda$, which people sometimes express by saying that the **variance-to-mean ratio** (VTMR) for the Poisson is 1.

R can be used to check informally whether the Poisson is a **good fit** to the LOS data:

```
rosalind 77> R

R : Copyright 2001, The R Development Core Team
Version 1.2.1  (2001-01-15)

> dpois( 0:7, mean( y ) )
[1] 0.126005645 0.261011693 0.270333539 0.186658872 0.096662630
[6] 0.040045947 0.013825386 0.004091186

> print( n <- length( y ) )

[1] 14

> table( y ) / n
          0           1           2           3           4           6
0.07142857 0.35714286 0.28571429 0.14285714 0.07142857 0.07142857
```

Poisson Modeling (continued)

```
> cbind( c( dpois( 0:6, mean( y ) ),
    1 - sum( dpois( 0:6, mean( y ) ) ) ),
    apply( outer( y, 0:7, '==' ), 2, sum ) / n )

             [,1]       [,2]
[1,] 0.126005645 0.07142857
[2,] 0.261011693 0.35714286
[3,] 0.270333539 0.28571429
[4,] 0.186658872 0.14285714
[5,] 0.096662630 0.07142857
[6,] 0.040045947 0.00000000
[7,] 0.013825386 0.07142857
[8,] 0.005456286 0.00000000
```

The second column in the above table records the values of the **Poisson probabilities** for $\lambda = 2.07$, the mean of the y_i, and the third column is the **empirical relative frequencies**; informally the fit is reasonably good.

Another **informal check** comes from the fact that the sample mean and variance are 2.07 and $1.542^2 \doteq 2.38$, which are reasonably close.

Exchangeability. As with the AMI mortality case study, before the data arrive I recognize that my uncertainty about the Y_i is exchangeable, and you would expect from a generalization of the binary-outcomes version of de Finetti's Theorem that the structure of a **plausible Bayesian model** for the data would then be

$$\begin{array}{rcll} \theta & \sim & p(\theta) & \text{(prior)} \\ (Y_i|\theta) & \overset{\text{IID}}{\sim} & F(\theta) & \text{(likelihood),} \end{array} \tag{59}$$

where θ is some parameter (vector) and $F(\theta)$ is some **parametric family of distributions** on the non-negative integers indexed by θ.

Poisson Modeling (continued)

Thus, in view of the preliminary examination of the data above, a **plausible Bayesian model** for these data is

$$\begin{aligned} \lambda &\sim p(\lambda) && \text{(prior)} \\ (Y_i|\lambda) &\overset{\text{IID}}{\sim} \text{Poisson}(\lambda) && \text{(likelihood)}, \end{aligned} \tag{60}$$

where λ is a **positive real number**.

NB (1) This approach to model-building involves a form of **cheating**, because we've **used the data twice**: once to choose the model, and again to draw conclusions conditional on the chosen model.

The result in general can be a failure to **assess** and **propagate model uncertainty** (Draper 1995).

(2) **Frequentist** modeling often employs this **same kind of cheating** in specifying the likelihood function.

(3) There are two Bayesian ways out of this dilemma: **cross-validation** and **Bayesian non-parametric/semi-parametric** methods.

The **latter** is beyond the scope of this course; I'll give examples of the **former** later.

To get more practice with Bayesian calculations I'm going to **ignore the model uncertainty problem for now** and pretend that somehow we knew that the Poisson was a good choice.

The likelihood function in model (60) is

$$\begin{aligned} l(\lambda|y) &= c\, p_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n|\lambda) \\ &= c \prod_{i=1}^{n} p_{Y_i}(y_i|\lambda) \\ &= c \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \\ &= c\, \lambda^s\, e^{-n\lambda}, \end{aligned} \tag{61}$$

The Conjugate Prior

where $y = (y_1, \ldots, y_n)$ and $s = \sum_{i=1}^n y_i$; here $\left(\prod_{i=1}^n y_i!\right)^{-1}$ can be **absorbed** into the generic positive c because it doesn't involve λ.

Thus (as was true in the Bernoulli model) $s = \sum_{i=1}^n y_i$ is **sufficient** for λ in the Poisson model, and we can write $l(\lambda|s)$ instead of $l(\lambda|y)$ if we want.

If a **conjugate** prior $p(\lambda)$ for λ exists it must be such that the product $p(\lambda)\, l(\lambda|s)$ has the same mathematical form as $p(\lambda)$.

Examination of (61) reveals that the same trick works here as with Bernoulli data, namely taking the **prior to be of the same form as the likelihood**:

$$p(\lambda) = c\, \lambda^{\alpha-1} e^{-\beta\lambda} \tag{62}$$

for some $\alpha > 0, \beta > 0$—this is the **Gamma** distribution $\lambda \sim \Gamma(\alpha, \beta)$ for $\lambda > 0$ (see Gelman et al. Appendix A).

As usual Maple can work out the **normalizing constant**:

```
rosalind 80> maple

    |\^/|     Maple V Release 5 (University of California, Santa Cruz)
._|\|   |/|_. Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
 \  MAPLE  /  reserved. Maple and Maple V are registered trademarks of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.

> assume( lambda > 0, alpha > 0, beta > 0 );

> p1 := ( lambda, alpha, beta ) -> lambda^( alpha - 1 ) *
    exp( - beta * lambda );

                                          (alpha - 1)
     p1 := (lambda, alpha, beta) -> lambda            exp(-beta lambda)

> simplify( integrate( p1( lambda, alpha, beta ),
    lambda = 0 .. infinity ) );

                          (-alpha~)
                    beta~          GAMMA(alpha~)
```

The Gamma Distribution

Thus $c^{-1} = \beta^{-\alpha}\Gamma(\alpha)$ and the
proper definition of the Gamma distribution is

$$\text{If } \lambda \sim \Gamma(\alpha, \beta) \text{ then } p(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \qquad (63)$$

for $\alpha > 0, \beta > 0$.

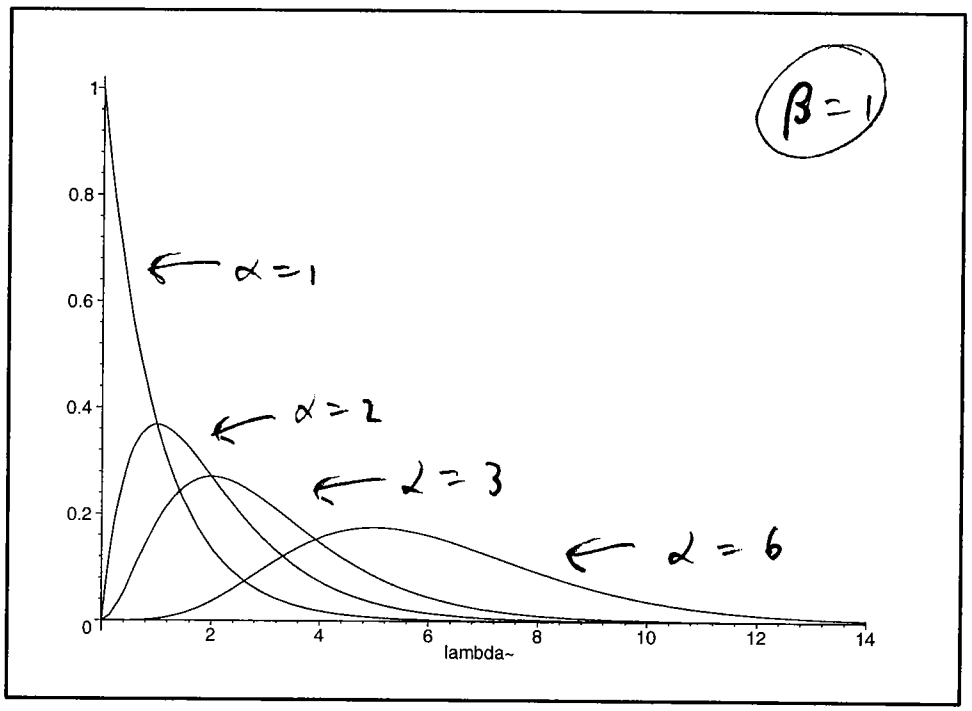
As usual Maple can also be used to explore the behavior of this family of distributions **as a function of its inputs** α and β:

```
> p := ( lambda, alpha, beta ) -> beta^alpha * lambda^( alpha - 1 ) *
    exp( - beta * lambda ) / GAMMA( alpha );

                                    alpha       (alpha - 1)
                                beta      lambda             exp(-beta lambda)
p := (lambda, alpha, beta) -> ----------------------------------------------
                                              GAMMA(alpha)

> plotsetup( x11 );

> plot( { p( lambda, 1, 1 ), p( lambda, 2, 1 ), p( lambda, 3, 1 ),
    p( lambda, 6, 1 ) }, lambda = 0 .. 14, color = black );
```



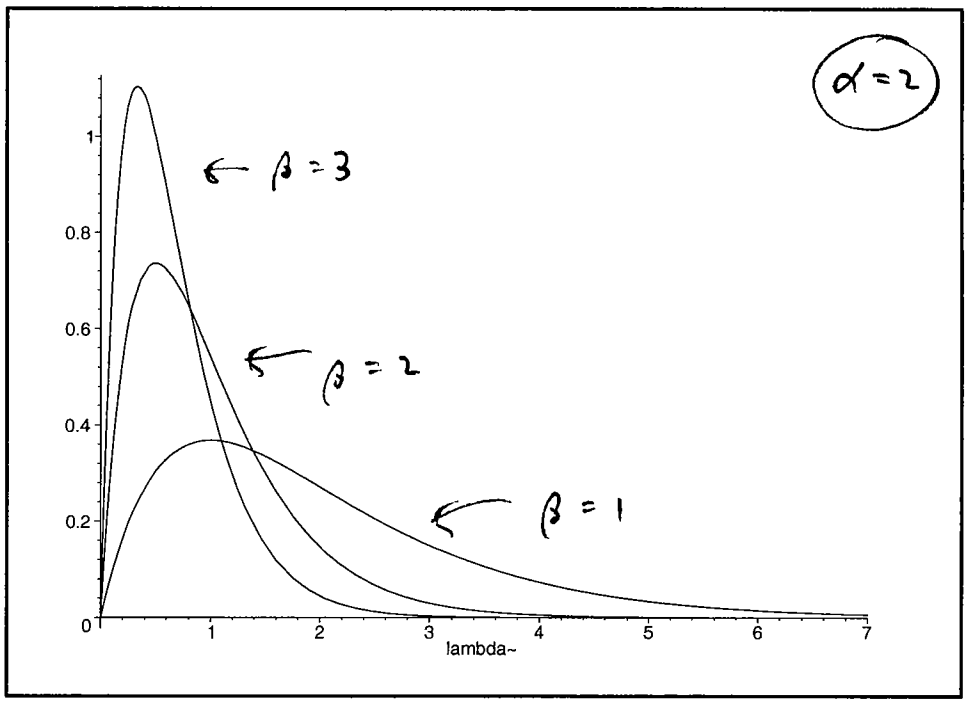
α evidently controls the **shape** of the Gamma family.

Gamma Distribution (continued)

When $\alpha = 1$ the Gamma distributions have a special form which you'll probably recognize—they're the **exponential** distributions $\mathcal{E}(\beta)$: for $\beta > 0$

$$\text{If } \lambda \sim \mathcal{E}(\beta) \text{ then } p(\lambda) = \left\{ \begin{array}{cc} \beta e^{-\beta\lambda} & \text{for } \lambda > 0 \\ 0 & \text{otherwise} \end{array} \right\}. \qquad (64)$$

```
> plot( { p( lambda, 2, 1 ), p( lambda, 2, 2 ), p( lambda, 2, 3 ) },
    lambda = 0 .. 7, color = black );
```



In the Gamma family the parameter β controls the **spread** or **scale** of the distribution.

Definition Given a random quantity y whose density $p(y|\sigma)$ depends on a parameter $\sigma > 0$, if it's possible to express $p(y|\sigma)$ in the form $\frac{1}{\sigma} f(\frac{y}{\sigma})$, where $f(\cdot)$ is a function which does not depend on y or σ, then σ is called a **scale** parameter for the parametric family p.

Gamma Distribution (continued)

Letting $f(t) = e^{-t}$ and taking $\sigma = \frac{1}{\beta}$, you can see that the Gamma family can be expressed in this way, so $\frac{1}{\beta}$ is a **scale parameter** for the Gamma distribution.

As usual Maple can also work out the **mean** and **variance** of this family:

```
> simplify( integrate( p( lambda, alpha, beta ),
    lambda = 0 .. infinity ) );
```

$$1$$

```
> simplify( integrate( lambda * p( lambda, alpha, beta ),
    lambda = 0 .. infinity ) );
```

$$\frac{\mathrm{alpha\tilde{}}}{\mathrm{beta\tilde{}}}$$

```
> simplify( integrate( ( lambda - alpha / beta )^2 *
    p( lambda, alpha, beta ), lambda = 0 .. infinity ) );
```

$$\frac{\mathrm{alpha\tilde{}}}{\mathrm{beta\tilde{}}^2}$$

Thus if $\lambda \sim \Gamma(\alpha, \beta)$ then $E(\lambda) = \frac{\alpha}{\beta}$ and $V(\lambda) = \frac{\alpha}{\beta^2}$.

Conjugate updating is now **straightforward**: with $y = (y_1, \ldots, y_n)$ and $s = \sum_{i=1}^{n} y_i$, by Bayes' Theorem

$$\begin{aligned} p(\lambda|y) &= c\,p(\lambda)\,l(\lambda|y) \\ &= c\left(c\,\lambda^{\alpha-1}\,e^{-\beta\lambda}\right)\left(c\,\lambda^{s}\,e^{-n\lambda}\right) \\ &= c\,\lambda^{(\alpha+s)-1}\,e^{-(\beta+n)\lambda}, \end{aligned} \tag{65}$$

and the **resulting distribution** is just $\Gamma(\alpha + s, \beta + n)$.

Conjugate Poisson Analysis

This can be **summarized** as follows:

$$\left\{ \begin{array}{c} (\lambda|\alpha,\beta) \sim \Gamma(\alpha,\beta) \\ (Y_i|\lambda) \overset{\text{IID}}{\sim} \text{Poisson}(\lambda), \\ i = 1,\ldots,n \end{array} \right\} \rightarrow (\lambda|s) \sim \Gamma(\alpha^*,\beta^*), \qquad (66)$$

where $(\alpha^*,\beta^*) = (\alpha+s,\beta+n)$ and $s = \sum_{i=1}^n y_i$ is a **sufficient statistic** for λ in this model.

The posterior mean of λ here is evidently $\frac{\alpha^*}{\beta^*} = \frac{\alpha+s}{\beta+n}$, and the prior and data means are $\frac{\alpha}{\beta}$ and $\bar{y} = \frac{s}{n}$, so (as was the case in the Bernoulli model) the posterior mean can be written as a **weighted average** of the prior and data means:

$$\frac{\alpha+s}{\beta+n} = \left(\frac{\beta}{\beta+n}\right)\left(\frac{\alpha}{\beta}\right) + \left(\frac{n}{\beta+n}\right)\left(\frac{s}{n}\right). \qquad (67)$$

Thus the **prior sample size** n_0 in this model is just β (which makes sense given that $\frac{1}{\beta}$ is the scale parameter for the Gamma distribution), and the prior acts like a **dataset** consisting of β observations with mean $\frac{\alpha}{\beta}$.

LOS data analysis. Suppose that, before the current data set is scheduled to arrive, I know **little** about the mean length of hospital stay of women giving birth to premature babies.

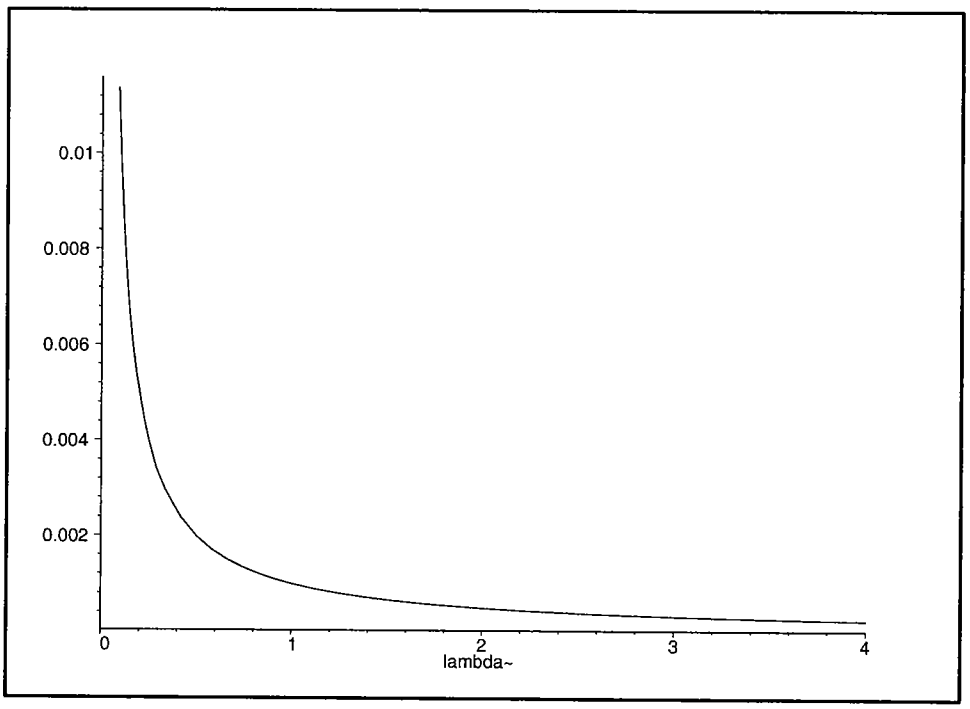
Then for my prior on λ I'd like to specify a member of the $\Gamma(\alpha,\beta)$ family which is relatively **flat in the region in which the likelihood function is appreciable**.

The $\Gamma(\epsilon, \epsilon)$ Prior

A **convenient** and **fairly all-purpose default choice** of this type is $\Gamma(\epsilon, \epsilon)$ for some small ϵ like 0.001.

When used as a prior this distribution has **prior sample size** ϵ; it also has mean 1, but that usually doesn't matter when ϵ is **tiny**.

```
> plot( p( lambda, 0.001, 0.001 ), lambda = 0 .. 4, color = black );
```



With the LOS data $s = 29$ and $n = 14$, so the **likelihood** for λ is like a $\Gamma(30, 14)$ density, which has mean $\frac{30}{14} \doteq 2.14$ and

$$\text{SD } \sqrt{\frac{30}{14^2}} \doteq 0.39.$$

Thus by the **Empirical Rule** the likelihood is appreciable in the range $(\text{mean} \pm 3\,\text{SD}) \doteq (2.14 \pm 1.17) \doteq (1.0, 3.3)$, and you can see from the plot above that the prior is indeed **relatively flat** in this region.

From the **Bayesian updating** in (66), with a $\Gamma(0.001, 0.001)$ prior the **posterior** is $\Gamma(29.001, 14.001)$.

LOS Data Analysis

It's useful, in summarizing the **updating** from prior through likelihood to posterior, to make a table that records measures of **center** and **spread** at each point along the way.

For example, the $\Gamma(0.001, 0.001)$ **prior**, when regarded (as usual) as a **density** for λ, has mean 1.000 and SD $\sqrt{1000} \doteq 31.6$ (i.e., informally, as far as we're concerned, before the data arrive λ could be **anywhere between 0 and (say) 100**).

And the $\Gamma(29.001, 14.001)$ **posterior** has mean $\frac{29.001}{14.001} \doteq 2.071$ and SD $\sqrt{\frac{29.001}{14.001^2}} \doteq 0.385$, so after the data have arrived we know **quite a bit more than before**.

There are two main ways to summarize the **likelihood**—Fisher's approach based on **maximizing** it, and the Bayesian approach based on regarding it as a density and **integrating** it—and it's instructive to compute them both and **compare**.

The **likelihood-integrating** approach treats the $\Gamma(30, 14)$ likelihood as a density for λ, with mean $\frac{30}{14} \doteq 2.143$ and SD $\sqrt{\frac{30}{14^2}} \doteq 0.391$.

As for the **likelihood-maximizing** approach, from (61) the log likelihood function is

$$ll(\lambda|y) = ll(\lambda|s) = \log\left(c\,\lambda^s e^{-n\lambda}\right) = c + s\log\lambda - n\lambda, \qquad (68)$$

and this is **maximized** as usual (check that it's the max) by setting the **derivative** equal to 0 and solving:

$$\frac{\partial}{\partial\lambda} ll(\lambda|s) = \frac{s}{\lambda} - n = 0 \quad \text{iff} \quad \lambda = \hat{\lambda}_{\text{MLE}} = \frac{s}{n} = \bar{y}. \qquad (69)$$