Conjugate priors and one-parameter inference

- Exact analytical solutions for posterior distributions can be found in special cases.
- Occurs if prior $\pi(\theta)$ is of the same functional form as $\pi(X|\theta)$, when seen as function of θ.
- These are called conjugate priors.

Conjugate priors and one-parameter inference

- First example is Binomial model:

  $P(X|\theta) = \text{Binomial}(N,\theta)$

  Model for sample data X,N.

  θ is e.g. population prevalence, etc.
- Conjugate prior is $\pi(\theta) = \text{Beta}(\alpha,\beta)$
- Note: Beta(1,1)=Uniform(0,1)
- Find out $\pi(\theta|X)$ by simple algebra, starting from Bayes theorem.

Binomial model

- **Posterior density: π(θ | X) = P(X|θ)π(θ)/c**
  - Assuming uniform prior, this is:

$$\pi(\theta \mid x) = \binom{N}{x} \theta^{x} (1-\theta)^{N-x} 1_{\{0<\theta<1\}}(\theta) / c$$

  - Take a look at this as a function of θ, with N, x, and c as fixed constants.
  - What probability density function can be seen? Hint: compare to beta-density.

$$\pi(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Binomial model

- The posterior density of θ can be written, up to a constant term as

$$\pi(\theta \mid N, x) \propto \theta^{x+1-1}(1-\theta)^{N-x+1-1}$$

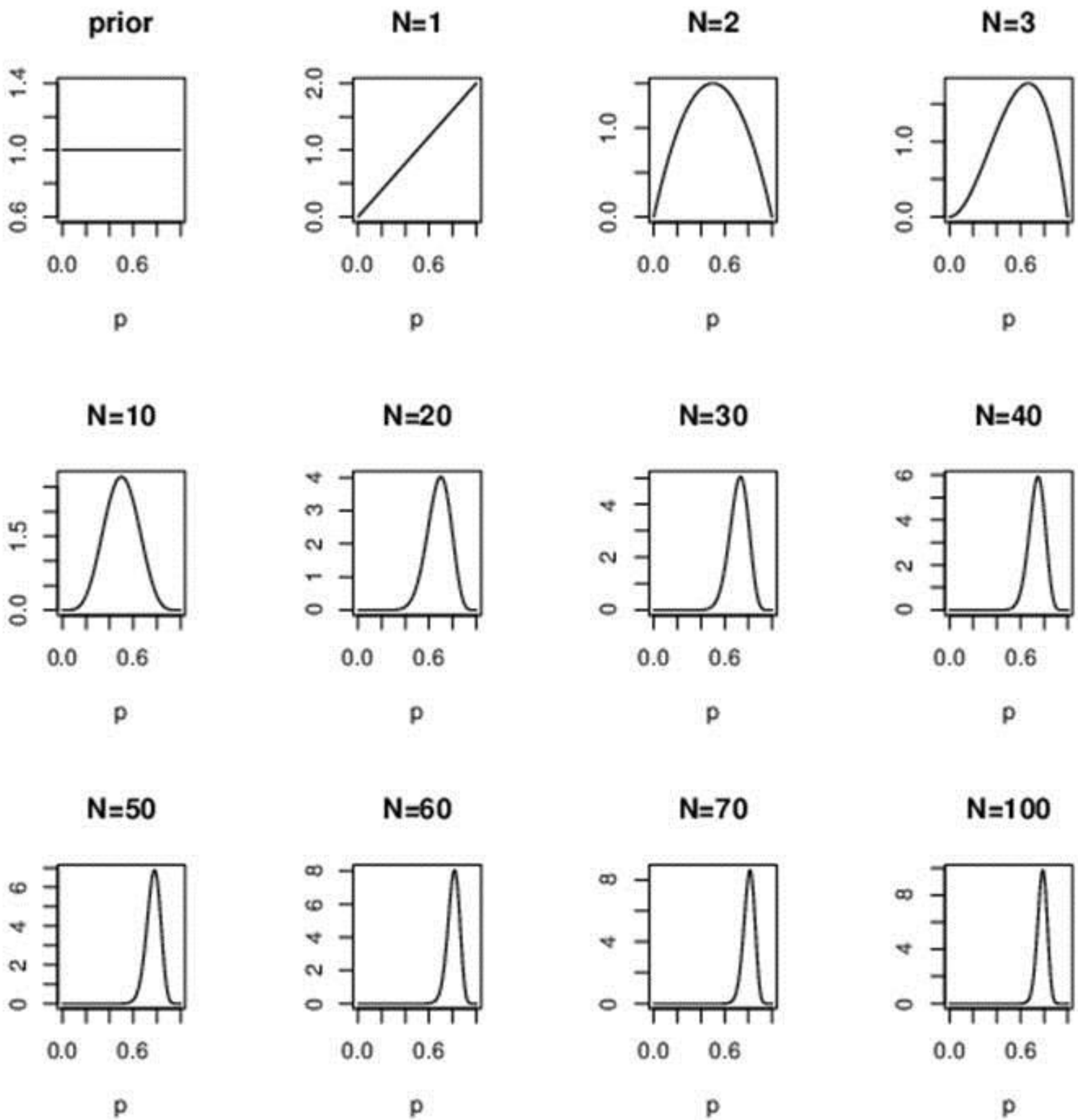
  - Same as beta(x+1,N-x+1)-density.

  - Generally, if the uniform prior is replaced by beta(α,β)-density, we get beta(x+α,N-x+β).

Binomial model

- The uniform prior corresponds to having two *'pseudo observations'*: one red ball, one white ball, as if that was 'observed' before data.
- The *posterior mean is* (1+X)/(2+N)
  - Generally: $(\alpha+X)/(\alpha+\beta+N)$
  - Can be expressed as: $w\frac{\alpha}{\alpha+\beta}+(1-w)\frac{X}{N}$

    With $w = (\alpha+\beta)/(\alpha+\beta+N)$
  - See what happens if $N \rightarrow \infty$, or if $N \rightarrow 0$.

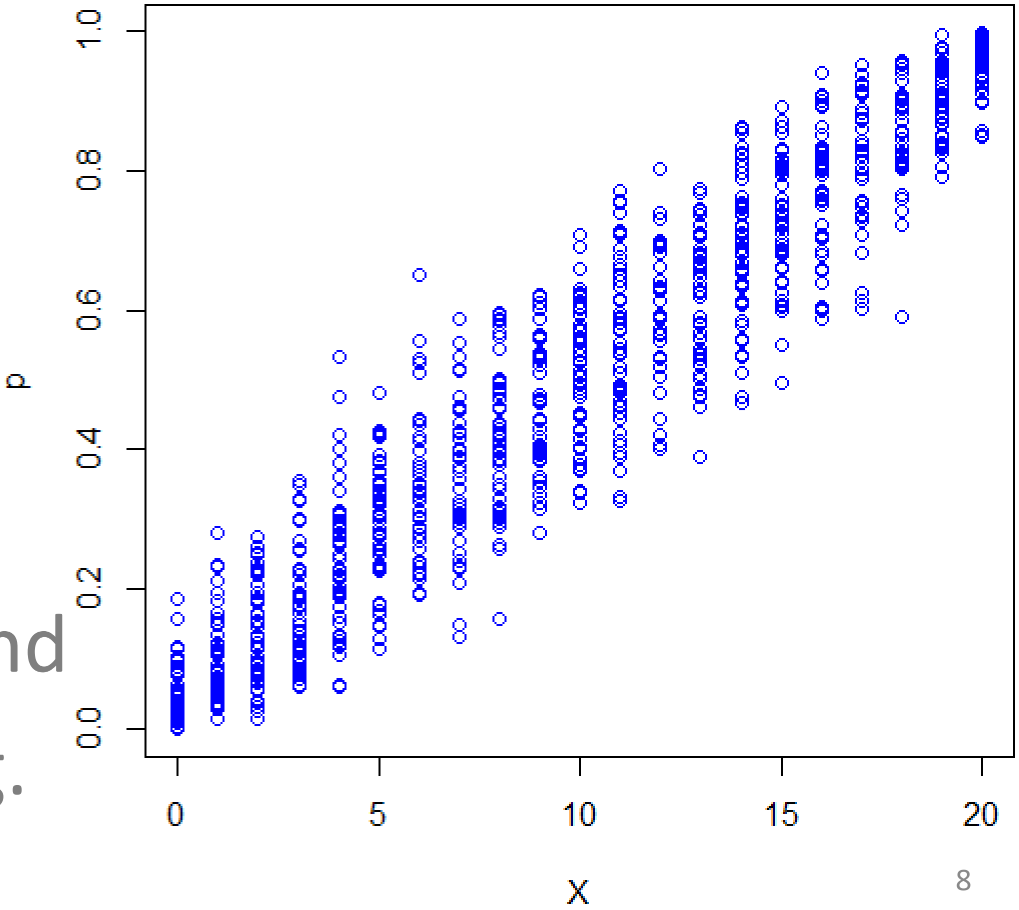

prior
N=1
N=2
N=3
N=10
N=20
N=30
N=40
N=50
N=60
N=70
N=100
p
0.0
0.6

Binomial model

- With any amount of data, we can make inference about θ.
- But, of course, with no data, we are left with the prior density! (which means we have learned nothing).
- But even one data point gives some additional piece of evidence…
- There is no requirement for size of data!

Binomial model

- Simulated sample from the joint distribution $\pi(\theta,X) = P(X|N,\theta)\pi(\theta)$
- Spot $P(X|N,\theta)$ and $\pi(\theta|X)$ in the Fig.



Why conjugate priors?

- Conjugate choice of prior leads to closed form solutions. (Posterior density is in the same family as prior density).
- Can also interpret conjugate prior as ’pseudo data’ or ’prior data’. → The amount of prior evidence easy to compare with amount of real data.

- Only a few conjugate solutions exist!

Likelihood principle

- ***Likelihood principle*: all information provided by data is contained in the likelihood function (uskottavuusfunktio) $L(\theta;data) = P(data|\theta)$.**
- Then, if two data sets lead to the same likelihood function, the inference must be identical.
- Likelihood inference (uskottavuuspäättely) in classical statistics is based on $L(\theta;data)$.

- Bayesian methods also obey likelihood principle:
  - e.g. it does not matter if we decide to make n experiments to observe some $x \sim Bin(n,p)$, or if we decide to continue until x successes, so that $n \sim NegBin$ → for p, the likelihood is same!

Bernoulli and Binomial model

- Think of a set of Bernoulli-variables $B_1,\ldots,B_n$ for which B_i = 0 or 1.
- $B_i \perp B_j$ are independent for all i & j, conditionally, given θ = the success probability.
- For each B_i, the Bernoulli probability is thus

$$P(B_i \mid \theta) = \theta^{B_i}(1-\theta)^{1-B_i}$$

- Then, the probability for the whole data, conditionally on θ is

$$P(B_1,\ldots,B_n \mid \theta) = \prod_{i=1}^{n} P(B_i \mid \theta) = \prod_{i=1}^{n} \theta^{B_i}(1-\theta)^{1-B_i} = \theta^{X}(1-\theta)^{n-X}$$

- So that $X = \Sigma(B_i) \sim \mathrm{Bin}(n,\theta)$.

Bernoulli and Binomial model

- X is called *sufficient statistics*. (tyhjentävä tunnusluku).
- For a given value of X, the inference on θ should be the same because the likelihood function $L(\theta)=P(\text{data}|\theta)$ is the same, regardless of the permutation of the B_i.
- Then, also the posterior of θ is the same under Binomial or Bernoulli data, (as long as the prior remains the same too).

Binomial model & priors

- Uniform prior U(0,1) for θ was 'uninformative'. In what sense?
- What if we study the density of θ^2 or $\log(\theta)$, assuming $\theta \sim U(0,1)$?
- Jeffreys' prior is uninformative in the sense that it is *transformation invariant*:

$$\pi(\theta) \propto J(\theta)^{1/2} \quad \text{with} \quad J(\theta) = E[(\frac{d \log(P(X \mid \theta))}{d\theta})^2 \mid \theta]$$

Binomial model & priors

- J(θ) is known as 'Fisher information for θ'
- With Jeffreys' prior for θ we get, for any one-to-one smooth transformation ϕ=h(θ) that:

Transformation of variables rule

Jeffreys'

$$\pi(\phi) = \pi(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{E[(\frac{d\log(L)}{d\theta})^2(\frac{d\theta}{d\phi})^2]}$$

$$= \sqrt{E[(\frac{d\log(L)}{d\phi})^2]} = \sqrt{J(\phi)}$$ where L = P(X|parameter)

Binomial model & priors

- For the binomial model, Jeffreys’ prior is Beta(1/2,1/2).
- But in general:
  - Jeffreys’ prior can lead to improper densities (integral is infinite).
  - Difficult to generalize into higher dimensions.
  - Violates likelihood principle which states that inferences should be the same when the likelihood function is the same.

Binomial model & priors

- Also: Haldane's prior $\pi(\theta) \propto \theta^{-1} (1-\theta)^{-1}$ is uninformative. (≈ "beta(0,0)")
    - (How? Think of 'pseudo data'…)
    - But is **improper**.
- *Can a prior be improper density?*
    - **Yes, but!** - the likelihood needs to be such that the posterior still integrates to one.
    - With Haldane's prior, this works only when the binomial data X is either >0 or <N. (but we could not know X in advance…)

Binomial model & priors

- For the binomial model $P(X|\theta)$, when computing the posterior $\pi(\theta|X)$, we have at least 3 different uninformative priors:

- $\pi(\theta)=U(0,1)=Beta(1,1)$ Bayes-Laplace
- $\pi(\theta)=Beta(1/2,1/2)$ Jeffreys'
- $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ Haldane's

- Each of them is uninformative in different ways!
- Unique definition for **uninformative** does not exist.

Binomial model & priors

- example: estimate the mortality

THIRD DEATH

"The expanded warning came as Yosemite announced that a third person had died of the disease (Hantavirus) and the number of confirmed cases rose to eight, all of them among U.S. visitors to the park."

Ok, it's a small data,

but we try:

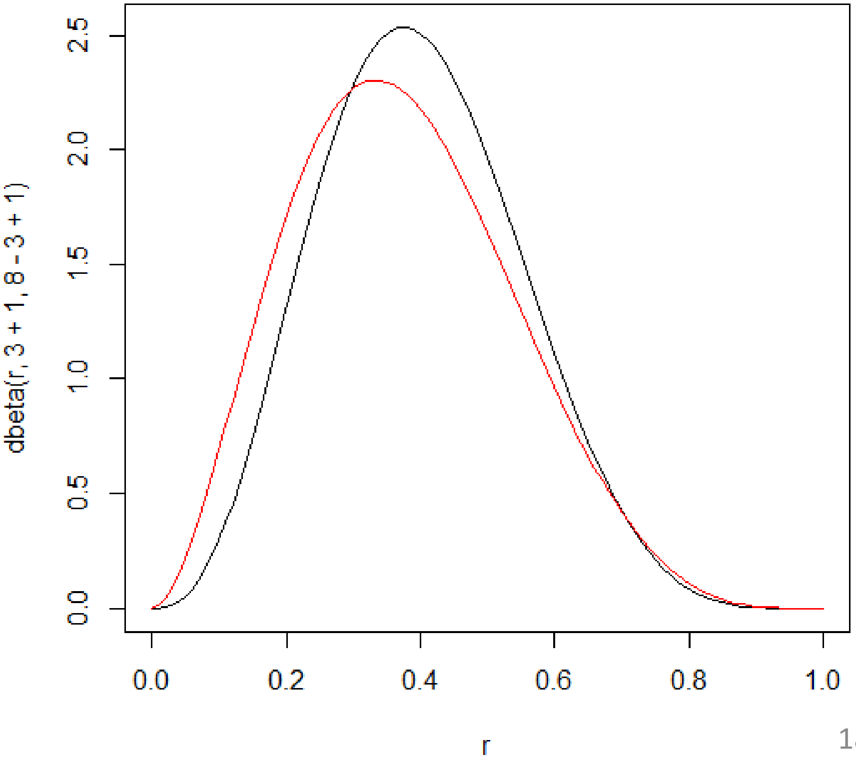
with uniform prior:

$\pi(r \mid \text{data})=\text{beta}(3+1,8-3+1)$.

Try also other priors.

Posterior with Haldane's in red →

"Since 1993, when the virus first was identified, the average death rate is 36 percent, according to the CDC"



Binomial model & N?

- In previous slides, N was fixed (known). We can also think situations where θ is known , X is known, but N is unknown.
- Exercise: solve P(N | θ,X) = P(X | N,θ)P(N)/c with suitable choice of prior.
  - Try e.g. discrete uniform over a range of values.
  - Try e.g. $P(N) \propto 1/N$
- Bayes generally: compute probabilities of any unknowns, given the knowns & prior & likelihood (model).

Exponential model

- Applicable for event times, concentrations, positive measurements,…

$$\pi(X \mid \theta) = \theta e^{-\theta X}$$

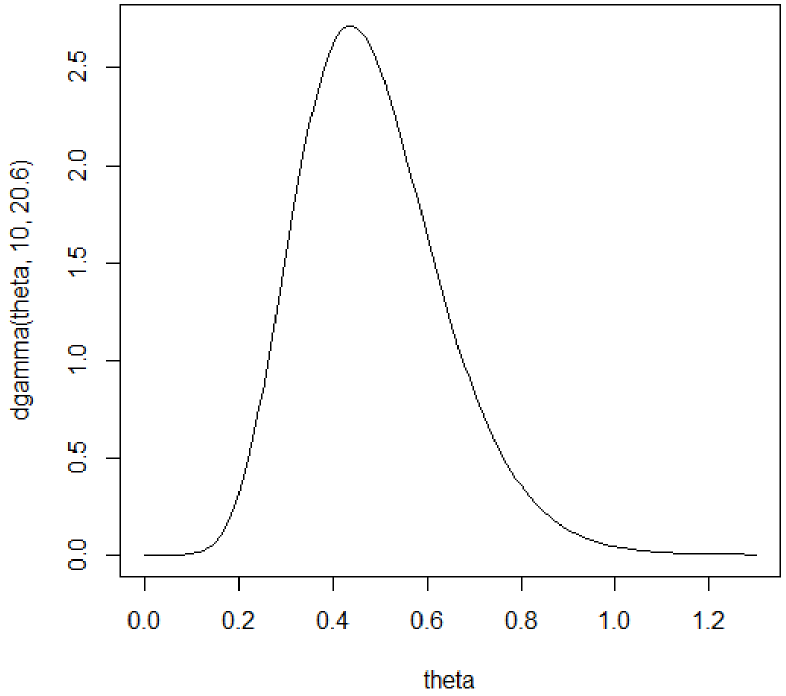
- Mean E(X) = $1/\theta$
- Aim to get $\pi(\theta|X)$, or $\pi(\theta \mid X_1,\ldots,X_N)$.
- Conjugate prior Gamma(α,β)
- Posterior: Gamma($\alpha+1,\beta+X$) or Gamma($\alpha+N,\beta+X_1+\ldots+X_N$).

Exponential model

- Posterior mean of θ is $(\alpha+N)/(\beta+X_1+\ldots+X_N)$
- What happens if $N \rightarrow \infty$, or $N \rightarrow 0$?
- Uninformative prior $(\alpha,\beta) \rightarrow (0,0)$
- Subjective & Objective Bayes approach:
  - Prior could be based on existing knowledge (→ expert knowledge elicitation or literature or previous data → informative gamma-prior)
  - Without using previous knowledge → use uninformative gamma-prior
  - As long as it's gamma-prior, exact solutions.

Exponential model

- Example: life times of 10 light bulbs were T = 4.1, 0.8, 2.0, 1.5, 5.0, 0.7, 0.1, 4.2, 0.4, 1.8 years. Estimate the failure rate? (true=0.5)
- $T_i \sim \exp(\theta)$
- Uninformative prior gives $\pi(\theta|T)$ = gamma(10,20.6).
- Could also parameterize with $1/\theta$ and use inverse-gamma prior.



Exponential model

- Some observations may be censored, so we only know that $T_i < c_i$, or $T_i > c_i$
- The probability for the whole data is then of the form ('full likelihood'):
- $P(\text{data} \mid \theta) =$

$$\Pi \pi(T_i \mid \theta) \, \Pi \, P(T_i < c_i \mid \theta) \, \Pi \, P(T_i > c_i \mid \theta)$$

- *For this we need cumulative probability functions, but Bayes theorem still applies, just more complicated.*

Poisson model

- Widely applicable model for counts x=0,1,2,3,… For example: disease cases, accidents, faults, births, deaths over a time, or within an area, etc…

- λ = E(X) $$P(X \mid \lambda) = \frac{\lambda^X}{X!} e^{-\lambda}$$
- Also: constant intensity in a Poisson process: E(X in time T) = λT

- With single observation X, aim to get: $\pi(\lambda|X) = P(X|\lambda)\pi(\lambda)/c$

Poisson model

- Conjugate prior? Gamma-density:

$$\pi(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

- Then:

$$\pi(\lambda \mid X) = \frac{\lambda^{X}}{X!} e^{-\lambda} \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} / c$$

- Simplify expression, what density you see? (up to a normalizing constant).

Poisson model

- Posterior density is Gamma($X+\alpha$,$1+\beta$).
- Posterior mean is $(X+\alpha)/(1+\beta)$

- Can be written as weighted sum of ’data mean’ X and ’prior mean’ α/β.

$$\frac{1}{1+\beta}X+\frac{\beta}{1+\beta}\frac{\alpha}{\beta}$$

Poisson model

- With a set of observations: $X_1,\ldots,X_N$:

$$P(X_1,\ldots,X_N \mid \lambda) = \prod_{i=1}^{N} \frac{\lambda^{X_i}}{X_i!} e^{-\lambda}$$

- And with the Gamma(α,β)-prior we get: Gamma($X_1+\ldots+X_N+\alpha$,$N+\beta$).

- Posterior mean $\frac{1}{N+\beta}\sum_{i=1}^{N} X_i + \frac{\beta}{N+\beta}\frac{\alpha}{\beta}$

- What happens if $N \to \infty$, or $N \to 0$?

Poisson model

- Uninformative Gamma-prior: in the limit $(\alpha,\beta) \rightarrow (0,0)$, so posterior is then Gamma($X_1+\ldots+X_N$,N). Alternatively, could use improper flat prior $\pi(\lambda) = U(0,\infty)$ so that posterior is proportional to likelihood.
- Alternatively, use informative prior: e.g. based on expert opinion from which we could elicitate prior mean and variance $E(\lambda)= \alpha/\beta$ and $V(\lambda)= \alpha/\beta^2$ for solving prior parameters α,β.

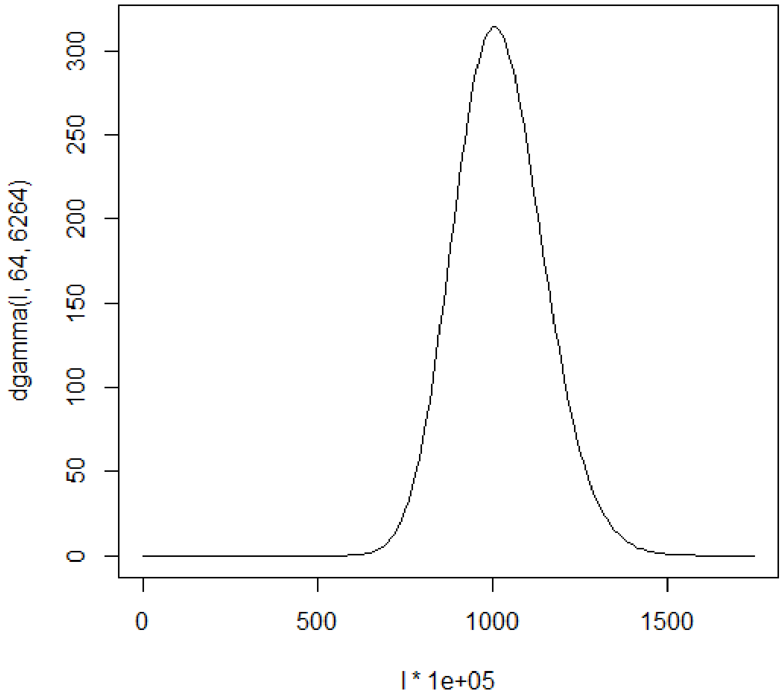
- Compare the conjugate analysis with Binomial model. Note similarities.

Poisson model in epidemiology

- **Parameterize with exposure**
  - epidemiological problems: rate of cases per year, or per 100,000 persons per year.
  - Model: $X_i \sim \text{Poisson}(\lambda E_i)$
  - E_i is **exposure**, e.g. population of the i^{th} city (in a year).
  - λ is common disease incidence (unknown).
  - X_i is observed number of cases in i^{th} city.
  - Aim to get posterior density of λ.

Poisson model in epidemiology

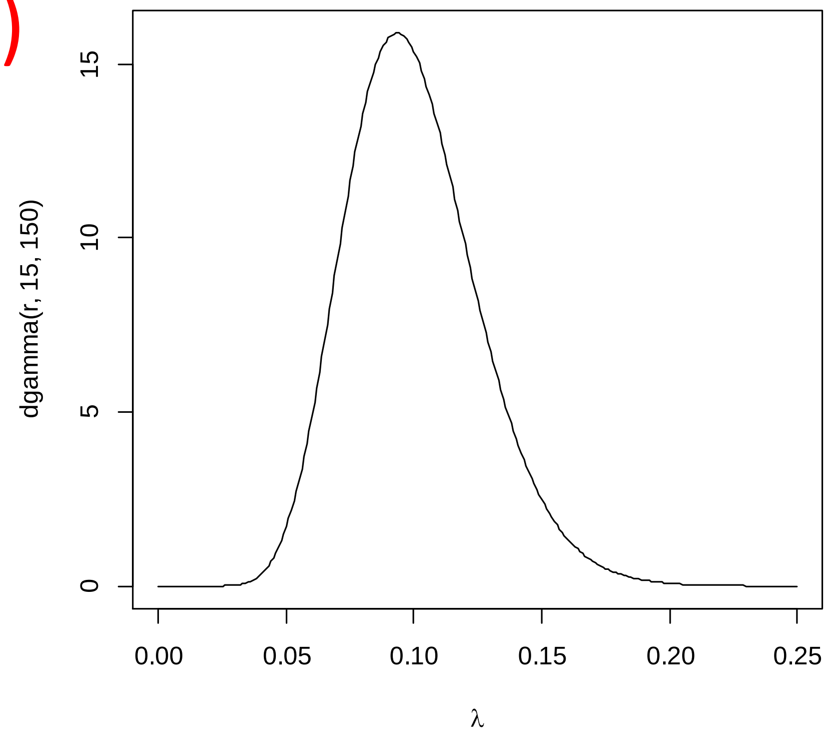
- Example: 64 lung cancer cases in 1968-1971 in Fredericia, Denmark, population 6264. Estimate incidence per 100,000?
- $\pi(\lambda | X,E)$
  $= \text{gamma}(\alpha+X, \beta+E)$
- With uninformative prior, X=64,E=6264, we get gamma(64,6264),
  (→plot: $10^5\ \lambda$)



Poisson model in microbiology

- Similar: λ = bacteria concentrations /g? Observed counts X: 5/100g, 10/50g
- $\pi(\lambda | X,E)$

  $= \text{gamma}(\alpha+\Sigma X_i, \beta+\Sigma E_i)$
- With uninformative prior, we get posterior: gamma(15,150)



Some examples of conjugate priors

Data model $\pi(x\|\theta)$	Prior of parameter $\pi(\theta)$	Posterior of parameter $\pi(\theta\|x)$
$x \sim$ Binomial(n, θ)	$\theta \sim$ Beta(a,b)	$\theta \sim$ Beta$(x+a,n-x+b)$
$x_i \sim$ Poisson(θ)	$\theta \sim$ Gamma(a,b)	$\theta \sim$ Gamma$(\Sigma x_i+a,n+b)$
$x_i \sim$ Exponential(θ)	$\theta \sim$ Gamma(a,b)	$\theta \sim$ Gamma$(n+a, \Sigma x_i+b)$
$x_i \sim N(\theta,1/\tau)$	$\theta \sim N(\theta_0,1/\tau_0)$	$\theta \sim N((\tau_0/(\tau_0+n\tau))\theta_0 + (n\tau/(\tau_0+n\tau))\ \bar{y}\ , 1/(\tau_0+n\tau))$
$x_i \sim N(\mu,1/\theta)$	$\theta \sim$ Gamma(a,b)	$\theta \sim$ Gamma$(a+n/2,b+n[s^2+(\bar{y}-\mu)^2]/2)$ $s^2 = n^{-1}\,\Sigma(y_i - \bar{y})^2$

(These examples for one-parameter inference).