Bayesian Inference and Decision Theory

Unit 3: Bayesian Inference with Conjugate Pairs: Single Parameter Models

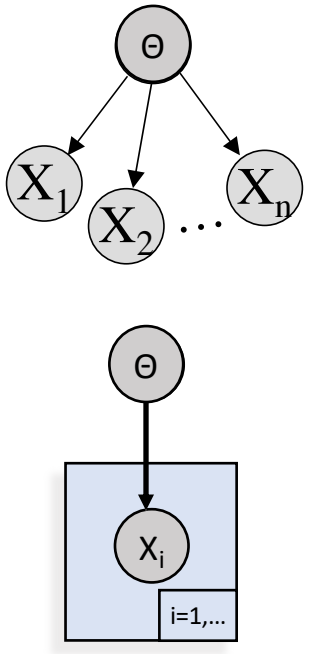
v3.2

Learning Objectives for Unit 3

- Define a conjugate pair of distributions
- Given a sample of observations and a conjugate prior distribution for the parameter(s), find the posterior distribution for the parameter(s), for the following conjugate pairs:
  - Poisson / Gamma
  - Exponential / Inverse-Gamma
  - Binomial / Beta
- Use a triplot to visualize the relationship between prior distribution, likelihood, and posterior distribution
- Define and compare Bayesian credible intervals and frequentist confidence intervals for a parameter
- Find the marginal likelihood for the sufficient statistic for a future sample when the prior and likelihood are from one of the conjugate pairs we have studied
- Evaluate the adequacy of a model in the light of observations
- Define reference priors for use as default priors

Canonical Statistical Inference Problem

- Given: data set of N observations $X_1, \ldots, X_N$
- Assume: X_i are a random (iid) sample from a probability distribution with unknown parameter Θ
  - X_i and Θ can be univariate or multivariate
  - Set of possible values for X and/or Θ can be finite, discrete infinite, or continuous
  - This unit considers univariate, continuous parameters
- Objectives:
  - Draw conclusions about the unknown parameter Θ
  - Predict future X_i
  - Recommend action in a decision problem that depends on Θ and/or future X_i



Bayesian Approach to Canonical Inference Problem

- Canonical inference problem:
  - Use N iid observations $X_1, \ldots, X_N$ drawn from $f(x|\theta)$ to draw inferences about unknown parameter Θ
- Inputs:
  - Prior distribution: gpdf $g(\theta)$ for unknown parameter Θ and
  - Likelihood function: Conditional gpdf $f(\underline{x}|\theta)$ for iid observations $\underline{X}$ given parameter $\Theta=\theta$

$$f(\underline{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_N|\theta) \quad \text{(product of likelihoods)}$$

- Output:
  - Posterior distribution: gpdf $g(\theta|\underline{x})$ for Θ given observations $\underline{X} = \underline{x}$
  - Predictive distribution: gpdf $f(x^{new}|\underline{x})$ for future observations given past observations (with parameter integrated out)

Bayes Rule for Continuous Distributions

- In many statistical models the parameter has a continuous range of possible values
- Bayes rule is the same for continuous as for discrete distributions except that the denominator is an integral rather than a sum:

$$g(\theta|x_1,\ldots,x_n) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)g(\theta)}{\int_\theta f(\underline{x}|\theta)g(\theta)d\theta}$$

- But often we cannot find an exact expression for the posterior distribution because there is no exact expression for the integral
- ***What then?***

Conjugate Pairs of Distributions

- ***Conjugate pairs allow exact computation of posterior distributions***
- A gpdf family $g(\theta|\alpha)$ is *conjugate* to the gpdf family $f(x|\theta)$ if it is *closed under sampling* from $f(x|\theta)$, that is:

  IF Data $X_1, \ldots, X_n$ are a random sample from $f(x|\theta)$ AND prior distribution for unknown parameter Θ is $g(\theta|\alpha)$

  THEN Posterior distribution for parameter Θ is $g(\theta|\alpha^*)$, another member of the conjugate family

- There is a simple updating rule to find α^* from α and the observations

α is called a *hyperparameter*

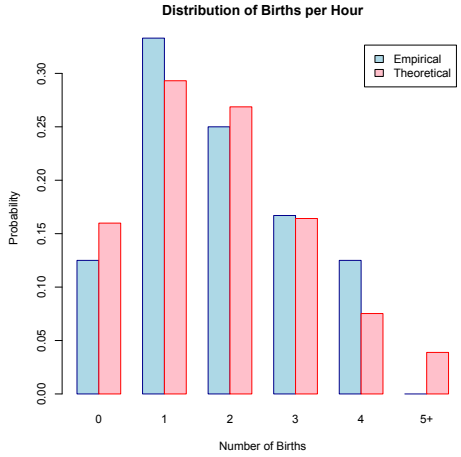
Recall: uppercase (e.g., X_i, Θ) denotes unknown ("random") quantities; lowercase (e.g., x_i, θ) denotes known values

Example: Bayesian Inference for Accident Rate

- A transportation engineer studying traffic accidents models the number of accidents per day as independent Poisson RVs with unknown parameter Λ accidents / day
  - X_i is the number of accidents in the i^{th} day of observation
  - Likelihood for n observations given accident rate $\Lambda = \lambda$: $f(x_1,\ldots,x_n \mid \lambda) = \prod_i \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \left(\prod_i x_i!\right)^{-1} e^{-n\lambda}\lambda^{\sum_i x_i}$
- The engineer wants to use the data to draw inferences about the unknown parameter Λ

Review: Poisson Distribution

- Poisson distribution models counts of events occurring in a period of time / region of space
- Examples:
  - Number of accidents on a stretch of road at a given time of day and day of week
  - Number of defects in a given time interval for a manufacturing process
  - Number of individuals of a given species in a given spatial region (does not apply to insects that cluster, such as ants or bees)
  - Number of goals in soccer matches
    - http://pena.lt/y/2012/10/29/using-poisson-to-predict-football-matches/
- When to use Poisson distribution:
  - Events in non-overlapping intervals or regions are independent
  - Rate (expected number of events) for small region is proportional to volume of region



Bayesian Inference for Accident Rate: The Poisson / Gamma Conjugate Pair

- A transportation engineer studying traffic accidents models the number of accidents per day as independent Poisson RVs with unknown parameter Λ accidents / day
  - X_i is the number of accidents in the i^{th} day of observation
  - Likelihood for n observations given accident rate Λ: $f(x_1,\ldots,x_n \mid \lambda) = \prod_i \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \left(\prod_i x_i!\right)^{-1} e^{-n\lambda}\lambda^{\sum_i x_i}$
- Engineer models the prior distribution for Λ as a Gamma distribution with shape α and scale β:

$$g(\lambda \mid \alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda/\beta} & \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

  - Parameters α and β of the prior distribution are often called *hyperparameters*
- The engineer has chosen a *conjugate prior* for the Poisson likelihood
  - Gamma family of prior distributions is closed under sampling from the Poisson likelihood

Gamma function

$$\Gamma(y) \equiv \int_0^{\infty} u^{y-1}e^{-u}du$$

Facts:
- $\Gamma(\alpha) = \alpha\Gamma(\alpha-1)$
- If n is an integer then $\Gamma(n) = (n-1)!$

Review: Gamma Distribution

- Density function for Gamma distribution

$$g(\lambda \mid \alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda/\beta} & \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Properties:

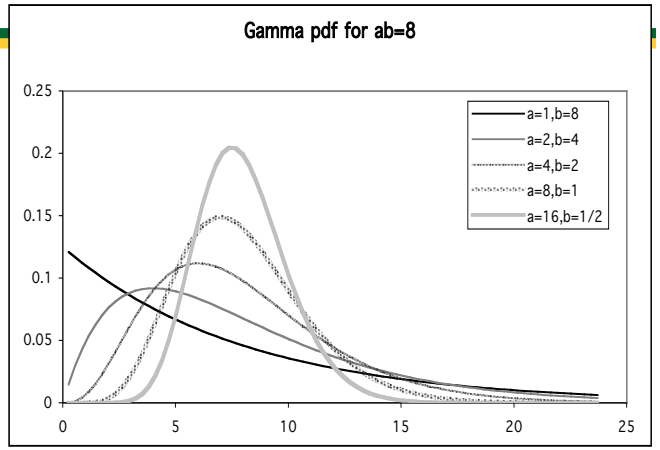
$$\mathrm{E}[\Lambda \mid \alpha,\beta] = \alpha\beta$$

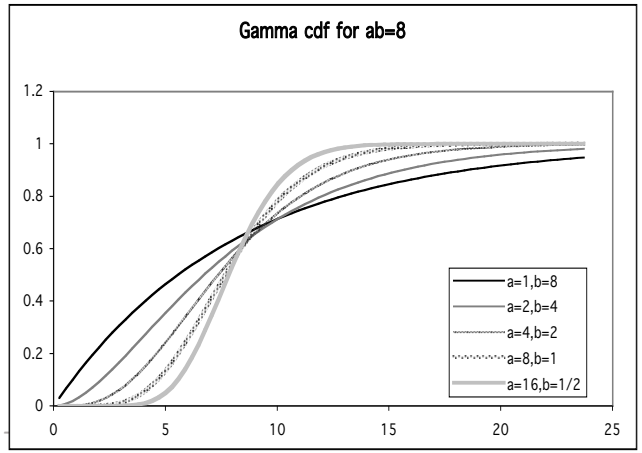
$$\mathrm{Var}[\Lambda \mid \alpha,\beta] = \alpha\beta^2$$

Mode is $(\alpha - 1)\beta$

This is the shape/scale parameterization for the Gamma distribution. The Hoff text and the R default use the shape/rate parameterization, where rate = 1/scale

- Parameter α controls the shape and parameter β controls the scale





Parameterizations for Gamma Distribution

- There are several common parameterizations of the Gamma distribution
- We used the *shape / scale* parameterization:
  - Shape parameter α controls skewness of distribution (for small values distribution is positively skewed; for large values distribution is nearly symmetrical)
  - Scale parameter β controls width of distribution (for small values distribution is more concentrated; for large values distribution is more spread out)
  - Density function: $g(\lambda \mid \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda/\beta}$
- The Hoff book uses the *shape / rate* parameterization:
  - Shape parameter $a = \alpha$ controls skewness of distribution
  - Rate (also called inverse scale) parameter $b = 1/\beta$
  - Density function: $g(\lambda \mid a, b) = \frac{b^{a}}{\Gamma(a)} \lambda^{a-1} e^{-\lambda b}$
- Some authors use the *shape / mean* parameterization:
  - Shape parameter $k = \alpha$ controls skewness of distribution
  - Mean parameter $\mu = k\beta$ controls width of distribution (for small values distribution is more spread out; for large values distribution is more concentrated)
  - Density function: $g(\lambda \mid k, \mu) = \left(\frac{k}{\mu}\right)^{k} \frac{1}{\Gamma(k)} \lambda^{k-1} e^{-\lambda k/\mu}$

Using Bayes Rule to Combine Prior Distribution With Observations

- Multiply likelihood times prior to obtain joint gpdf for $(\underline{X},\Lambda)$:
  - Likelihood: $f(\underline{x} \mid \lambda) = \left(\prod_i x_i!\right)^{-1} e^{-n\lambda} \lambda^{\sum_i x_i}$
  - Prior density: $g(\lambda \mid \alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda/\beta}$
  - Joint gpdf for $(\underline{X},\Lambda)$ – prior times likelihood:

$$f(\underline{x} \mid \lambda) g(\lambda \mid \alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\prod_i x_i!\right)^{-1} \lambda^{\alpha+\sum_i x_i - 1} e^{-(\beta^{-1}+n)\lambda} = \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\prod_i x_i!\right) \lambda^{\alpha^*-1} e^{-\lambda/\beta^*}$$

$\sum X_i$ is a sufficient statistic for Λ

- Define $\alpha^* = \alpha + \sum_i x_i$ and $\beta^* = \frac{1}{\beta^{-1}+n} = \frac{\beta}{1+n\beta}$
- The joint gpdf for $(\underline{X},\Lambda)$ is proportional to the density function for a gamma distribution with parameters α^* and β^*

Posterior Distribution for Λ

- Likelihood for $\underline{x}$ given λ : $f(\underline{x}\,|\,\lambda) = \left(\prod_i x_i!\right)^{-1} e^{-n\lambda}\lambda^{\sum_i x_i}$
- Prior density for λ: $g(\lambda\,|\,\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda/\beta}$
- Joint gpdf for $(\underline{X},\Lambda)$: $f(\underline{x}\,|\,\lambda)g(\lambda\,|\,\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}\left(\prod_i x_i!\right)^{-1}\lambda^{\alpha+\sum_i x_i-1}e^{-(\beta^{-1}+n)\lambda}$

$$= \frac{1}{\beta^\alpha\Gamma(\alpha)}\left(\prod_i x_i!\right)^{-1}\lambda^{\alpha^*-1}e^{-\lambda/\beta^*}$$

$$\alpha^* = \alpha + \sum_i x_i$$

$$\beta^* = \frac{1}{\beta^{-1}+n} = \frac{\beta}{1+n\beta}$$

$\sum X_i$ is a sufficient statistic for Λ

- Posterior gpdf for Λ given $\underline{X}$:

$$g(\lambda\,|\,\underline{x},\alpha,\beta) = \frac{f(\underline{x}|\lambda)g(\lambda|\alpha,\beta)}{\int_{\lambda\geq 0} f(\underline{x}|\lambda)g(\lambda|\alpha,\beta)d\lambda} = \frac{\frac{1}{\beta^\alpha\Gamma(\alpha)}(\prod_i x_i!)^{-1}\lambda^{\alpha^*-1}e^{-\lambda/\beta^*}}{\int_{\lambda\geq 0}\frac{1}{\beta^\alpha\Gamma(\alpha)}(\prod_i x_i!)^{-1}\lambda^{\alpha^*-1}e^{-\lambda/\beta^*}d\lambda}$$

$$= \frac{\lambda^{\alpha^*-1}e^{-\lambda/\beta^*}}{\int_{\lambda\geq 0}\lambda^{\alpha^*-1}e^{-\lambda/\beta^*}d\lambda} = \frac{1}{(\beta^*)^{\alpha^*}\Gamma(\alpha^*)}\lambda^{\alpha^*-1}e^{-\lambda/\beta^*}$$

Summary: Poisson-Gamma Conjugate Pair

- The Poisson and Gamma families of distributions are a *conjugate pair*:

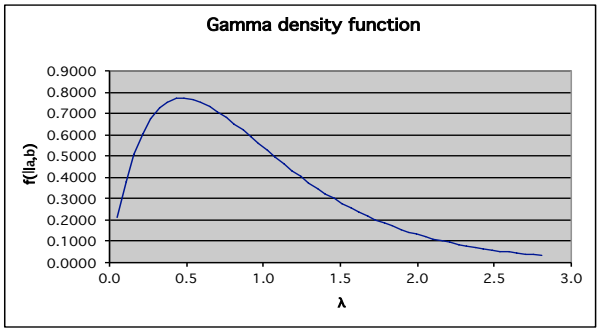
  IF Observations $X_1, \ldots, X_n$ are a random sample from Poisson(Λ) and prior distribution for Λ is Gamma(α, β)

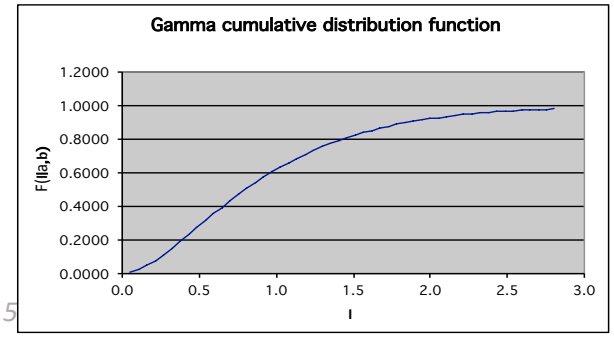
  THEN Posterior distribution for Λ is Gamma(α^*, β^*), another member of the conjugate family, where $\alpha^* = \alpha + \sum_i x_i$ and $\beta^* = (\beta^{-1} + n)^{-1}$

- Conjugate pairs simplify Bayesian inference
  - Posterior distribution can be found exactly
  - There is a simple updating rule to find the parameters of the posterior distribution from parameters of the prior distribution and sufficient statistic

Transmission Error Example Revisited: Prior Distribution

- Number of transmission errors per hour is distributed as Poisson distribution with unknown parameter Λ
- Data on previous system established error rate as 1.6 errors per hour
- New system design goal is 0.8 errors per hour
- Expert gives us the following information:
  - Chance of meeting design goal is 50% (Median of prior distribution is 0.8)
  - Chance that new system is worse than old is 15% (85^{th} percentile of prior distribution is 1.6)
- Fit Gamma distribution to these judgments:
  - Parameters shape α=2 and scale β=0.48
  - Expected value is 0.95; standard deviation is 0.68
  - Verify other quantiles and shape of pdf with expert





Transmission Error Example Revisited: Posterior Distribution

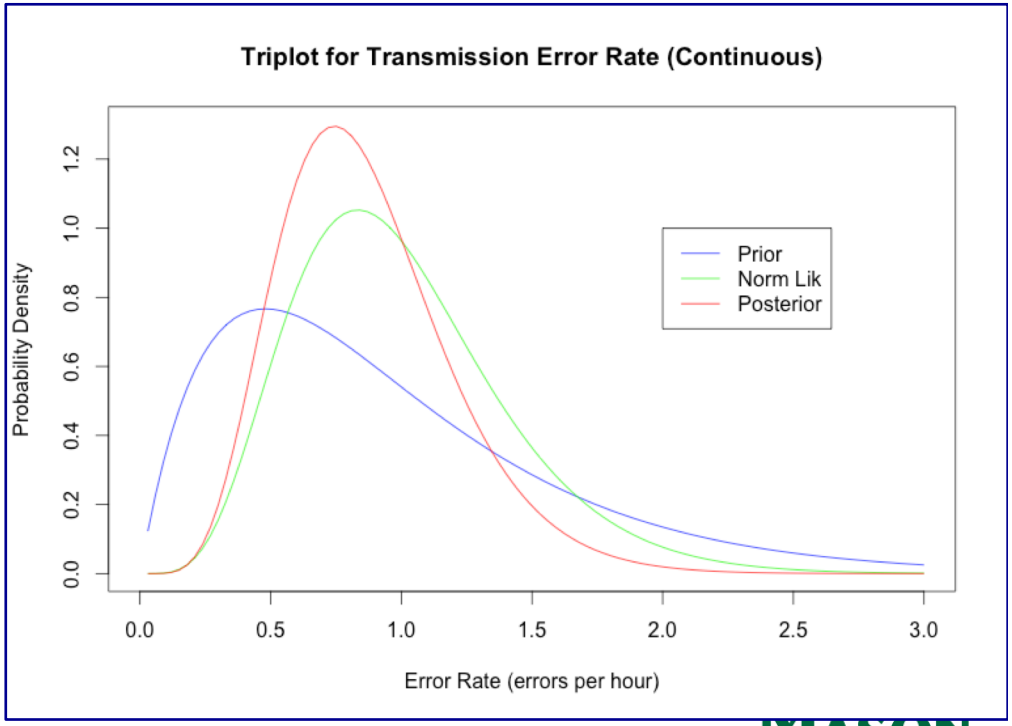
- The data: 6 one-hour observation periods with 1,0,1,2,1,0 errors
- Posterior distribution of Λ given data is a Gamma distribution with:

  $\alpha^* = 2 + 5 = 7$ and $\beta^* = \frac{1}{\frac{1}{0.48} + 6} = 0.124$

- Conjugate updating gives an exact expression for the posterior distribution
  - Mean = $\alpha^*\beta^* = 0.87$
  - Standard deviation = $\sqrt{\alpha^*(\beta^*)^2} = 0.33$

Triplot: A Tool for Visualizing Bayesian Updating

- Visual tool for examining Bayesian belief dynamics
- Plot prior distribution, normalized likelihood, and posterior distribution
- Normalized likelihood:
  - Posterior distribution we would obtain if all values had equal prior density
  - To calculate, divide likelihood by integral over λ
  - Normalized likelihood is also a gamma distribution



Finding the Normalized Likelihood

- The normalized likelihood is:
  - proportional to the likelihood function;
  - the posterior distribution from using a uniform prior;*
- If the likelihood function has a conjugate family, the normalized likelihood is a member of the conjugate prior family
- We find the normalized likelihood by choosing a member of the conjugate family proportional to the likelihood function
- For the Poisson/Gamma family
  - Likelihood: $f\left(\underline{x}\middle|\lambda\right) = \left(\prod_i x_i!\right)^{-1} e^{-n\lambda} \lambda^{\sum_i x_i}$
  - Gamma density: $g(\lambda|\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda/\beta}$
  - Matching terms involving λ, we find that the likelihood function is proportional to a Gamma density with $\alpha - 1 = \sum_i x_i$ and $1/\beta = n$
  - Therefore, the normalized likelihood for a random sample from a Poisson distribution is a Gamma distribution with $\alpha = \sum_i x_i + 1$ and $\beta = 1/n$

**A uniform distribution on the positive real numbers integrates to infinity, so is not a proper probability density function.*

Transmission Errors: Prior to Posterior

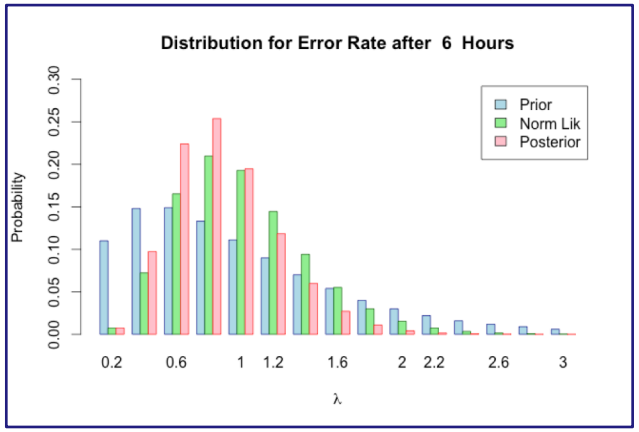
	Prior	Posterior	Normalized Likelihood
median	.81	.83	.95
85th percentile	1.62	1.20	1.42
95th percentile	2.28	1.47	1.75
mean	.96	.87	1.00
standard deviation	.68	.33	.41
P(meet design goal)	.50	.47	.35
P(worse than old)	.15	.03	.08
mode	.48	.74	.83

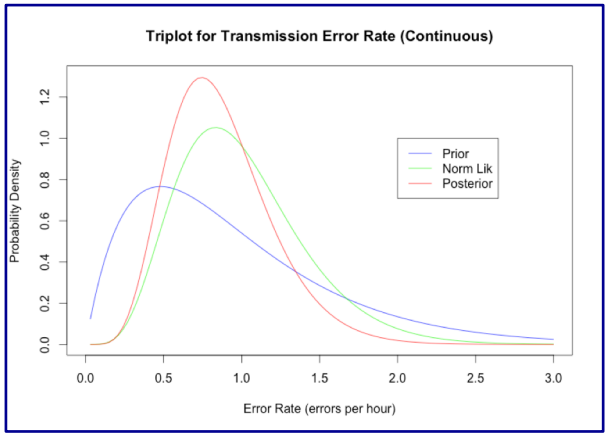
- *Posterior distribution is more concentrated than prior and normalized likelihood*
- *Posterior distribution "shrinks" mode of likelihood toward mode of prior*

Comparison: Approximate and Exact

	Approximate	Exact
• Posterior Mean	0.87	0.87
• Posterior Standard Deviation	0.33	0.33
• Probability of Meeting Goal	0.58	0.47
• Probability of *Exactly* Meeting Goal	0.25	0.00
• Probability New System is Better	0.96	0.97

Discrete approximation from Unit 2





Continuous Model from Unit 3

Discrete Approximation to Continuous Density Function

- Area of rectangle centered at λ is density at midpoint times rectangle width 0.2

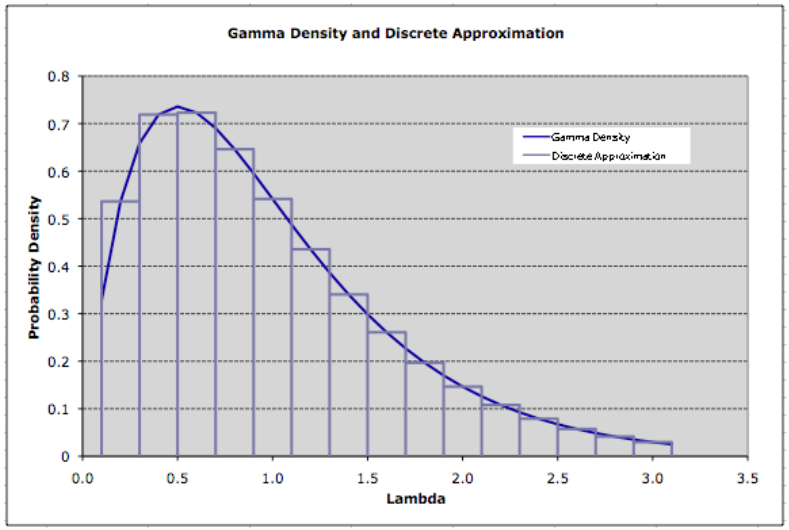
$$A(\lambda_i) = g(\lambda_i|\alpha,\beta) * 0.2$$

- Sum of rectangle areas is approximately equal to 1 (the integral of the density function):

$$\sum_i A(\lambda_i) = \sum_i g(\lambda_i|\alpha,\beta) * 0.2 \approx \int_\lambda g(\lambda|\alpha,\beta)d\lambda$$

- Discrete approximation at x is height of rectangle divided by sum of rectangle areas

$$g_{approx}(\lambda_i) = \frac{g(\lambda_i|\alpha,\beta)}{\sum_i g(\lambda_i|\alpha,\beta)}$$



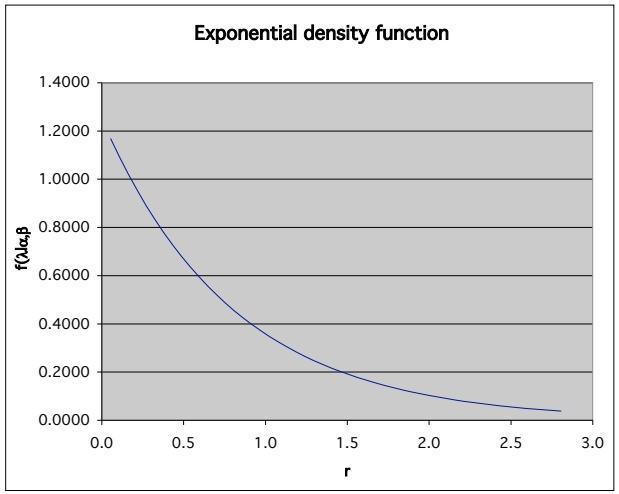
Summary: Transmission Errors

- We used Bayesian conjugate updating to find the posterior distribution for the transmission error rate
- Conjugate pairs simplify Bayesian inference by giving a method for calculating an exact posterior distribution
  - But the likelihood must be a good model for the observations and the prior distribution must reflect the prior knowledge
- A gamma prior distribution is conjugate to the Poisson likelihood
- The discrete analysis in Unit 2 for the transmission errors example is an approximation to the exact analysis from this unit

Example: Exponential / Inverse-Gamma Conjugate Pair

- A transportation engineer studying traffic accidents models the time between accidents as independent exponential RVs with unknown mean parameter Θ
  - X_i is the time between the i-1st and the ith accident
  - g.p.d.f for N observations given θ:

$$f(\underline{x}|\theta) = \prod_i \frac{1}{\theta} e^{-x_i/\theta} = \theta^{-n} \exp\left\{-\sum_i x_i/\theta\right\}$$



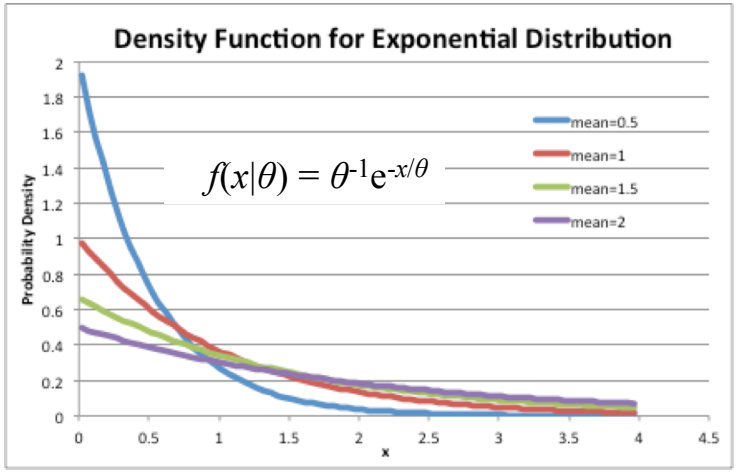
- Engineer models the prior distribution for the mean Θ as an Inverse-Gamma distribution with shape α and inverse-scale β

$$g(\theta \mid \alpha, \beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-1/(\theta\beta)} & \text{if } \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

- The engineer has chosen a prior distribution from a *conjugate family* to the Exponential distribution

Exponential Distribution – Overview

- Exponential distribution is a model for time intervals between events
- Time until next event does not depend on
  - When the last event occurred
  - How many events have occurred previously
  - What time it is now
- Exponential distribution is the only continuous memoryless distribution



Exponential Distribution – Details

- Exponential distribution with mean parameter θ:
  - Density function $f(x|\theta) = \theta^{-1}e^{-x/\theta}$
  - $E[X \mid \theta] = \theta$
  - Exponential(θ) distribution is a Gamma distribution with shape 1 and scale θ
  - Exponential distribution has *memoryless property*:
    - $P(X \geq q+r \mid X \geq q, \theta) = P(X \geq r \mid \theta)$
    - (If time between accidents is exponentially distributed, the probability of an accident in the next hour is independent of the time since the last accident)
  - If the number of events in 1 unit of time has the Poisson distribution with rate parameter λ, then the time between events has the exponential distribution with mean $\theta = \lambda^{-1}$
- Joint distribution for n independent observations $X_1, \ldots, X_n$ from exponential distribution with mean θ:
  - $f(\underline{x}|\theta) = \theta^{-n} \exp\left\{-\sum_i x_i / \theta\right\}$
  - $\sum X_i$ is sufficient for θ
  - $\sum X_i$ has Gamma distribution with shape n and scale θ
- The conjugate prior for the mean parameter Θ of the exponential distribution is the *inverse-Gamma* distribution

Inverse-Gamma Distribution

- If Λ has a Gamma(α,β) distribution then $\Theta=\Lambda^{-1}$ has a distribution with density function

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

*Do you know how to derive this density function?**

- Mean and Variance of Θ:

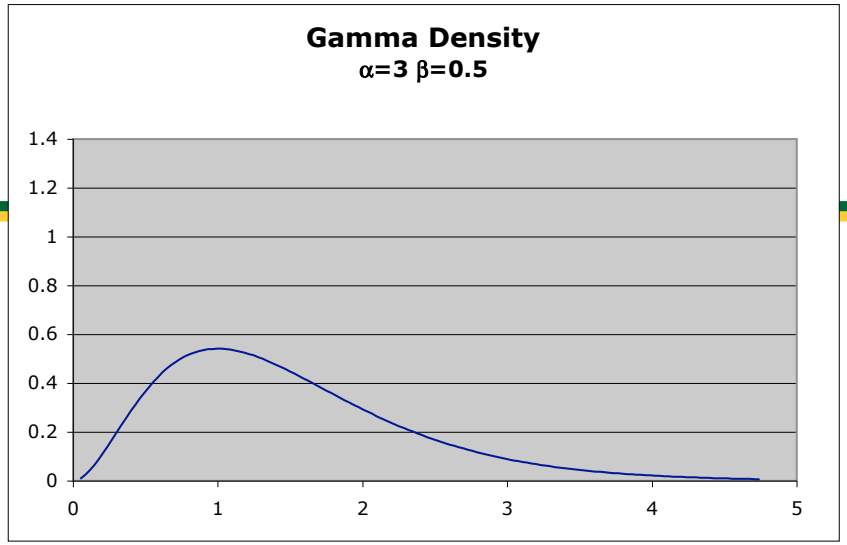
$$E[\Theta \mid \alpha,\beta] = \frac{1}{\beta(\alpha-1)} \quad \text{if } \alpha > 1$$

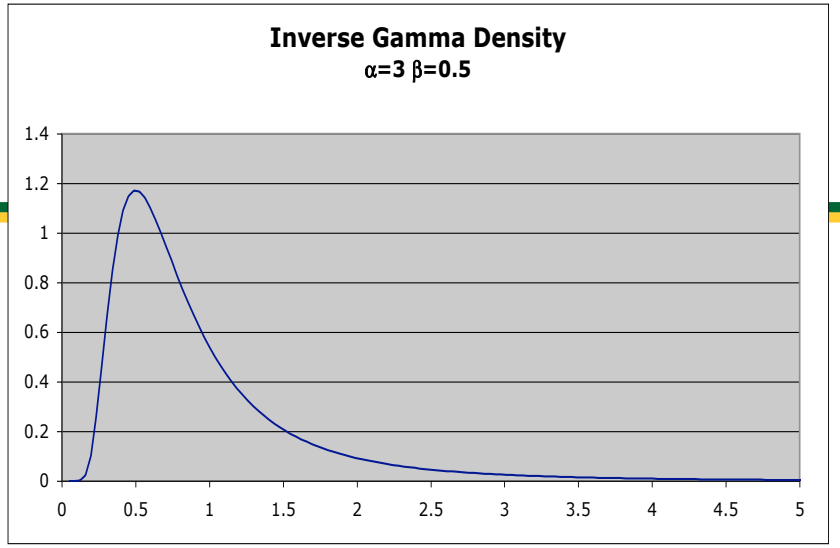
$$V[\Theta \mid \alpha,\beta] = \frac{1}{\beta^2(\alpha-1)^2(\alpha-2)} \quad \text{if } \alpha > 2$$

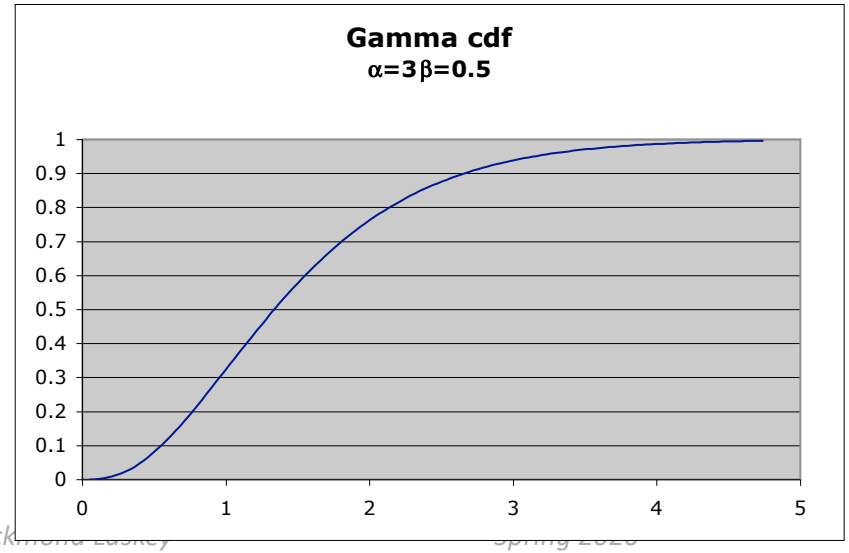
If the mean number of events per time period has a Gamma(α, β) distribution, then the mean time between events has an Inverse-Gamma(α, β) distribution

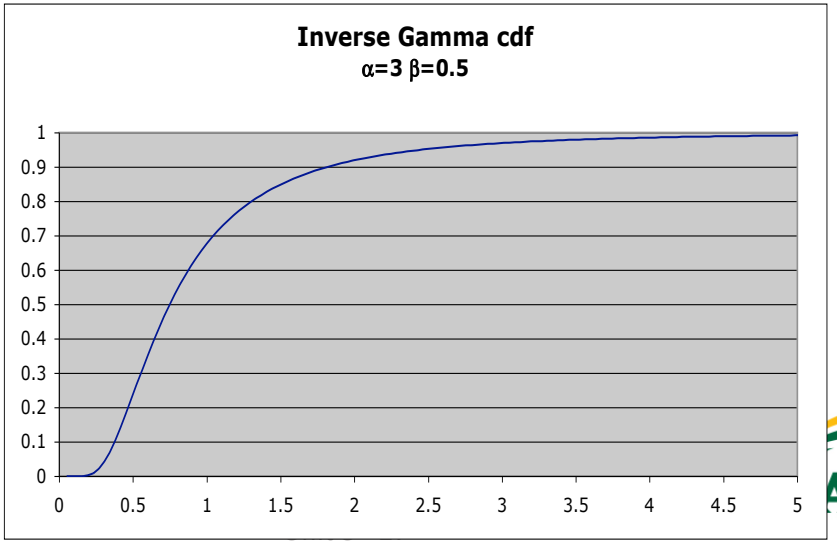
* *See* http://www.math.uah.edu/stat/dist/Transformations.html *for information on transformations of random variables*

- To find percentiles of Θ:
  - If λ_p is *p*th percentile of Λ, then $\theta_{100-p} = 1/(\lambda_p)$ is (100-*p*)th percentile of Θ
  - Example:
    - 0.551 is 10% point of Gamma(3,0.5) distribution $P(\Lambda \le 0.551) = 0.1$
    - 1/0.551 = 1.82 is 90% point of Inverse-Gamma(3,0.5) distribution $P(\Lambda^{-1} \ge 1.82) = 0.1$
- R functions `[d,p,q,r]invgamma` in `invgamma` package
  - `qgamma(p,shape=a,scale=b)` is equal to `1/qinvgamma(1-p,shape=a,scale=b)`
  - (we use the term inverse-scale because β does not satisfy the definition of a scale parameter)









Exponential / Inverse-Gamma Conjugate Pair

- $\underline{X}$ is a random sample from an exponential distribution with unknown mean Θ:

$$f(\underline{x} \mid \theta) = \theta^{-n} e^{-\sum_i x_i / \theta}$$

- Conjugate prior distribution for Θ is Inverse-Gamma(α,β):

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

- Joint density for (X, Θ):

$$f(\underline{x} \mid \theta) g(\theta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \theta^{-(\alpha+n+1)} e^{-\frac{1}{\theta}\left(\frac{1}{\beta} + \sum_i x_i\right)}$$

- Posterior distribution for Θ given $\underline{X}$ is Inverse-Gamma(α^*, β^*)
  - $\alpha^* = \alpha + n$
  - $\beta^* = \frac{1}{\beta^{-1} + \sum x_i} = \frac{\beta}{1 + \beta \sum x_i}$
  - $g(\theta \mid \underline{x}, \alpha, \beta) = g(\theta \mid \alpha^*, \beta^*) = \frac{1}{(\beta^*)^{\alpha^*}\Gamma(a^*)} \theta^{-(\alpha^*+1)} e^{-1/(\theta\beta^*)}$
- Inverse-Gamma family is conjugate to the exponential distribution with mean parameter

Exponential / Inverse-Gamma Conjugate Pair

- The exponential and inverse-gamma families of distributions are a *conjugate pair*:

  IF Observations $X_1, \ldots, X_n$ are a random sample from exponential(Θ) and prior distribution for Θ is inverse-gamma(α, β)

  THEN Posterior distribution for Θ is inverse-gamma(α^*, β^*), another member of the conjugate family, where $\alpha^* = \alpha + n$ and $\beta^* = (\beta^{-1} + \sum_i x_i)^{-1}$

Poisson and Exponential Distributions: The Relationship

Times between events are independent and identically distributed exponential random variables with mean θ

IF AND ONLY IF

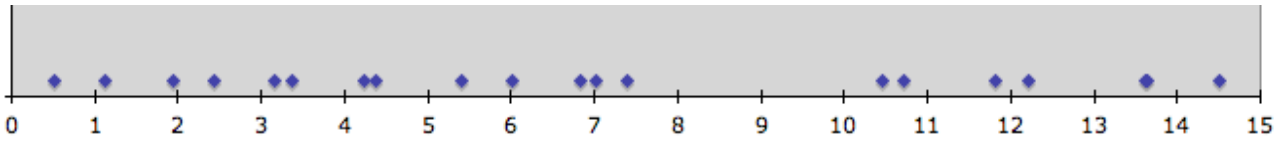
Counts of events per unit of time are independent and identically distributed Poisson random variables with mean $\lambda = \frac{1}{\theta}$

- θ is the mean time between events
- λ is the mean number of events per unit time

Illustration: Counts and Inter-event Times

- A process is observed and coded as both inter-event times and counts:
- Time data:
  - Observation times: 0.53, 1.12, 1.95, 2.45, 3.17, 3.39, 4.25, 4.38, 5.41, 6.01, 6.84, 7.04, 7.40, 10.48, 10.74, 11.85, 12.23, 13.64, 13.67, 14.53
  - Inter-event times: 0.53*, 0.59, 0.83, 0.50, 0.72, 0.22, 0.86, 0.13, 1.03, 0.60, 0.83, 0.20, 0.36, 3.08, 0.26, 1.11, 0.38, 1.41, 0.03, 0.86
  - Inter-event times have exponential distribution
  - 20 events occurred in a total time of 14.53 time units
- Count data:
  - Data: 1, 2, 1, 2, 2, 1, 2, 2, 0, 0, 2, 1, 1, 2, 1
  - Counts have Poisson distribution
  - Process was watched for 15 time units and 20 events were observed

**Because of memoryless property of the exponential distribution, the time from the start of observation until the first event has same exponential distribution as the times between the k^{th} and $(k+1)^{st}$ events*



Bayesian Inference – Poisson and Exponential

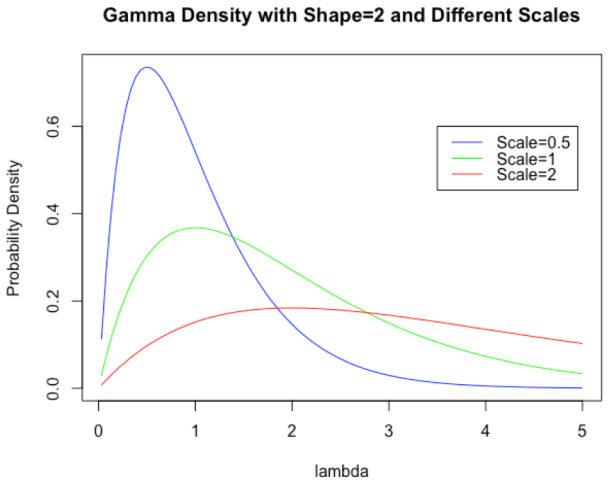
- The data:
  - Number of observations in time period of length 1 has Poisson distribution with parameter Λ
  - Time between observations has exponential distribution with parameter $\Theta=\Lambda^{-1}$
  - Prior distribution of Λ has Gamma distribution with shape α and scale β
- Case 1: Poisson data (watch for n time periods and see how many events occur)
  - n observations $X_1, \ldots, X_n$ of Poisson (Λ) data
  - Posterior distribution for Λ is Gamma with parameters $\alpha^* = \alpha+\Sigma x_i$ and $\beta^* = (\beta^{-1}+n)^{-1}$
- Case 2: Exponential data (watch until m events occur)
  - m observations $Y_1, \ldots, Y_m$ of exponential(Θ) data (where $\Theta = \Lambda^{-1}$)
  - Posterior distribution for Θ is inverse-Gamma with shape $\alpha^* = \alpha+m$ and inverse-scale $\beta^*=(\beta^{-1}+\Sigma y_i)^{-1}$
- In both cases:
  - Λ is rate at which events occur per unit time; $\Theta = \Lambda^{-1}$ is mean time between events
  - Posterior shape $\alpha^* = \alpha$ + number of events observed
  - Posterior scale [inverse-scale] $\beta^* = (1/\beta$ + total length of time process was observed$)^{-1}$
  - Posterior distribution of Λ is Gamma(α^*, β^*)
  - Posterior distribution of Θ is Inverse-gamma(α^*, β^*)

A Note on Terminology

- In statistics, a scale parameter determines how "spread out" a distribution is
  - If $f(x)$ is the gpdf for a "standard" distribution with scale 1
  - Then $\frac{1}{\beta} f(x/\beta)$ is the gpdf for the distribution with scale β
- The parameter β is a scale parameter in the gamma distribution with density $g_{Gamma}(\lambda|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}e^{\lambda/\beta}$
- The parameter $1/\beta$ is a scale parameter in the inverse-gamma distribution with density

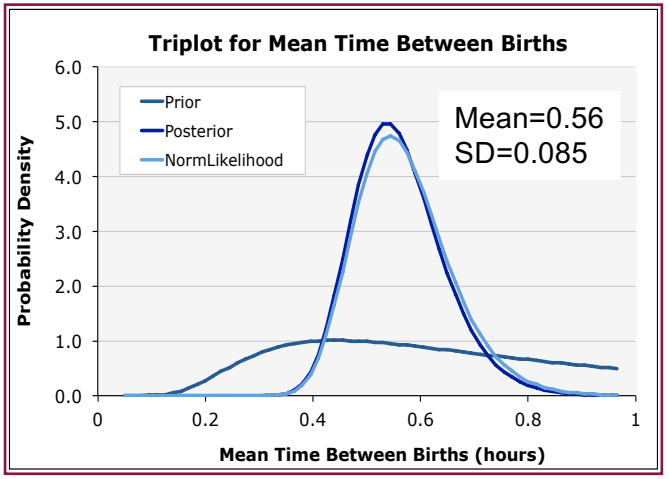
$$g_{Invgamma}(\theta|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{1/(\theta\beta)}$$

- We call β the "inverse-scale" of $g_{Invgamma}(\theta|\alpha,\beta)$
- The `invgamma` package in R calls this parameter "scale" although it is not a true scale parameter



Example: Baby Births

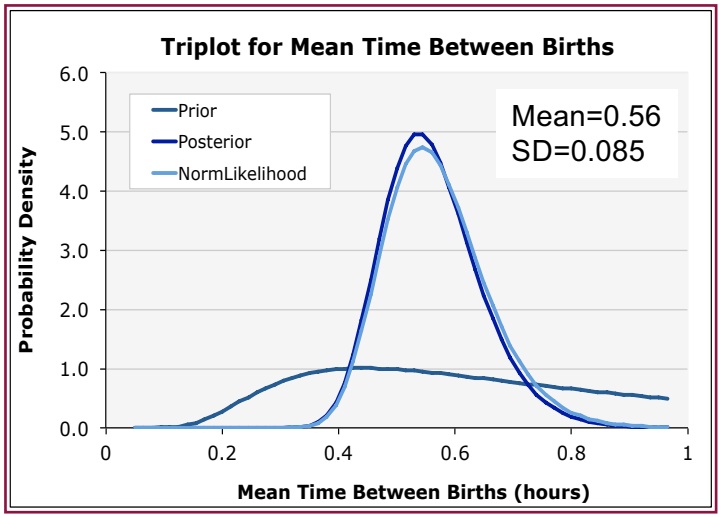
- Data were collected on the time of births at a hospital in Brisbane, Australia
  - http://www.amstat.org/publications/jse/secure/v7n3/datasets.dunn.cfm
  - 44 births occurred between midnight and 23.92 hours after midnight
- Assume:
  - Time between births are iid exponential random variables with mean time between births Θ
  - Expert provides this prior information:
    - Median of Θ is about 0.8 hours (24 hours for 30 births)
    - 90th percentile of Θ is about 2.5 hours (25 hours for 10 births)
  - These judgments fit an inverse-gamma prior distribution with α=2 and β=0.75
    - Median 0.79, 90th percentile 2.51
- Then:
  - Posterior distribution for Θ is inverse-Gamma with α^*=46 and $\beta^* = (0.75^{-1} + 23.92)^{-1} = 0.0396$



Counts and Inter-Birth Times

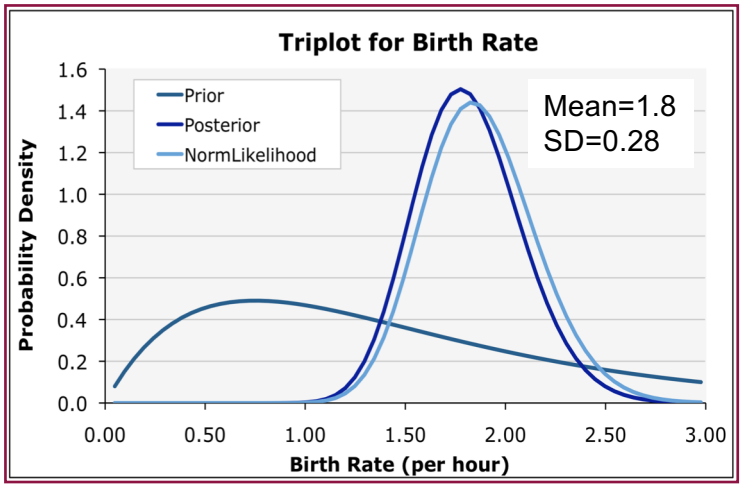
- Case 1: Poisson counts of baby births:
  - Number of observations n = 24 one-hour time periods
  - Sum of observations Σx_i = 44 total births
  - Posterior distribution of the rate is Gamma with parameters $\alpha^* = \alpha+44$ and $\beta^* = (\beta^{-1}+ 24)^{-1}$
- Case 2: Exponential times between births
  - Number of observations m = 44 events
  - Sum of observations Σy_i = 23.92 hours until 44th birth
  - Posterior distribution of mean is inverse Gamma with parameters $\alpha^* = \alpha+44$ and $\beta^* = (\beta^{-1}+ 23.92)^{-1}$
- In both cases:
  - Λ is rate at which births occur per hour; $\Theta = \Lambda^{-1}$ is mean time between births
  - Posterior distribution for Λ is Gamma(α^*, β^*) with mean $\alpha^*\beta^*$
  - Posterior shape $\alpha^* = \alpha$ + number of events observed = $\alpha + 44$
  - Posterior scale [inverse-scale] β^* = ($1/\beta$ + total length of time process was observed)$^{-1}$= $(1/\beta + 24)^{-1}$

Inference about Exponential Mean Θ and Poisson Rate Λ



- Prior distribution for rate Λ is Gamma with shape α=2 and scale β = 0.75
- Posterior distribution for rate Λ is gamma with shape α^*=46 and scale $\beta^* = (0.75^{-1} + 23.92)^{-1} = 0.0396$

- Prior distribution for mean Θ is inverse-Gamma with shape α=2 and inverse-scale β = 0.75
- Posterior distribution for mean Θ is inverse-gamma with shape α^*=46 and inverse-scale $\beta^* = (0.75^{-1} + 23.92)^{-1} = 0.0396$



Confidence and Credible Intervals

For the Brisbane birth data set:

- Posterior distribution for birth rate is Gamma with shape 46 and scale 0.0396
- 90% posterior credible interval for the birth rate Λ is [1.403, 2.285] births per hour
  - In R we use `qgamma(p,shape=46,scale=0.0396)` with p = 0.05 and 0.95
  - In Excel we use =GAMMA.INV(0.05,46,0.0396) and =GAMMA.INV(0.95,46,0.0396)
- 90% posterior credible interval for the mean time between births is [1/2.285, 1/1.403] = [0.438, 0.713]
  - Can you explain why?
- A 90% frequentist confidence interval for the birth rate is [1.404, 2.357] births / hour
  - In R we use `poisson.test(44,24,conf.level=0.9)$conf.int`
- For many commonly used statistical models, a frequentist confidence interval with coverage probability p has Bayesian coverage probability approximately equal to p if:
  - Sample size is sufficiently large
  - Prior distribution is not too informative

Summary: Poisson and Exponential Distributions

- Two different views of the same process:
  - Counts of events
  - Times between events
- If counts are iid Poisson with rate Λ then inter-event times are iid exponential with mean $\Theta = \Lambda^{-1}$
- If rate Λ has Gamma distribution then mean Θ has inverse-gamma distribution
- Anything we learn about Λ can be translated to information about Θ (and vice versa)

Binomial Distribution

- The binomial distribution is used to model a process having two possible outcomes ("success" and "failure")
  - *X* is the number of successes in *n* independent trials with probability θ of success on each trial
  - Probability mass function for *X*:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient

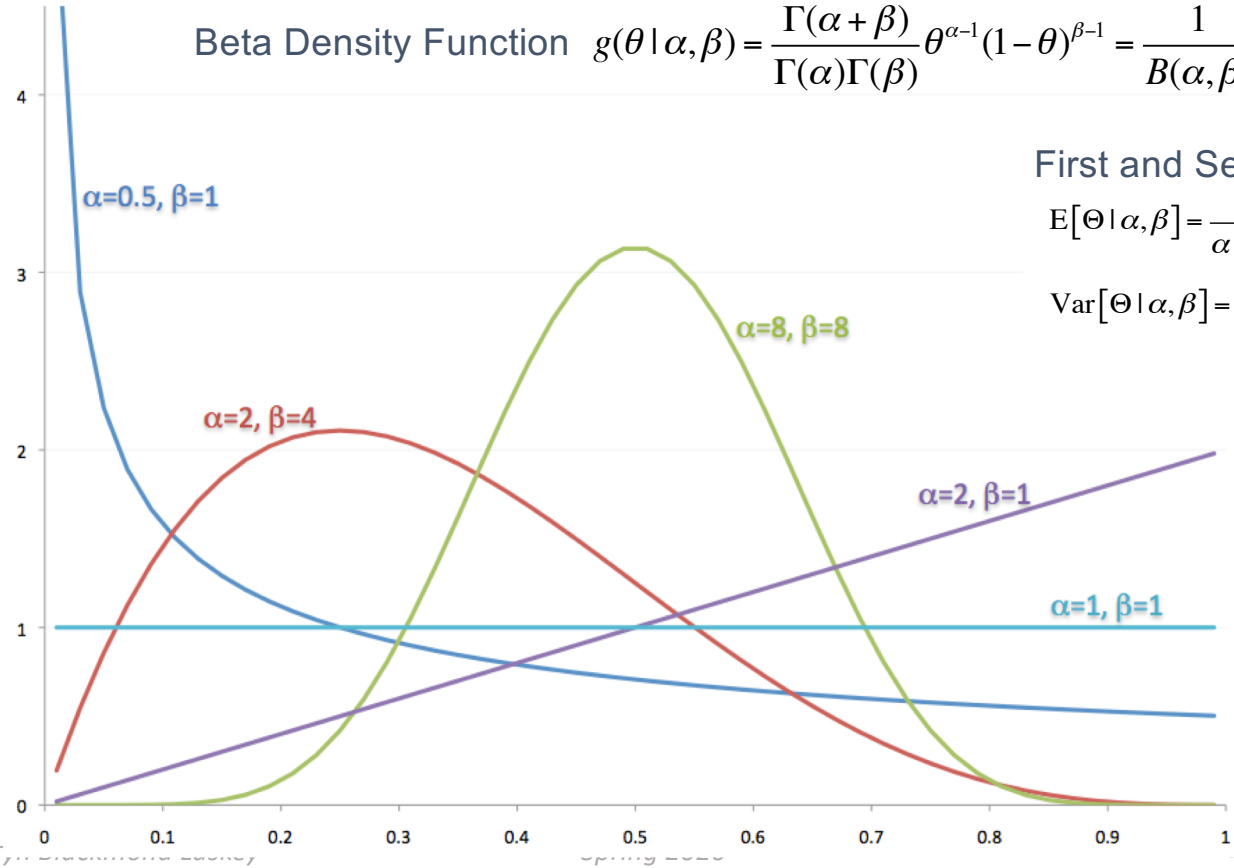
- Some facts about the binomial distribution:
  - Binomial (1, θ) distribution is also called Bernoulli(θ) distribution
  - The sum of n independent Bernoulli(θ) random variables has the binomial(n, θ) distribution
  - If X_1, …, X_r are independent and identically distributed binomial(n, θ) random variables, then:
    - The total number of successes $\sum X_i$ is a sufficient statistic for θ
    - $\sum X_i$ has a binomial ($r\,n$, θ) distribution

Bayesian Inference about a Binomial Probability (recall example from Unit 1)

- An analyst models the number X of ill patients arriving at a clinic using a binomial distribution with size n (total number of patients) and unknown probability Θ
  - Likelihood function $f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$
- The analyst models her prior information about Θ as a beta distribution with shape parameters α and β
  - Prior density function $g(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The analyst has chosen a conjugate prior distribution for the binomial likelihood

Beta Distribution

Beta Density Function $g(\theta \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$



First and Second Moments

$$\mathrm{E}[\Theta \mid \alpha,\beta] = \frac{\alpha}{\alpha+\beta}$$

$$\mathrm{Var}[\Theta \mid \alpha,\beta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The integral $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 q^{\alpha-1}(1-q)^{\beta-1}dq$

is called the beta function

Binomial / Beta Inference

- Data X distributed Binomial(n,Θ) with unknown Θ
  - $f(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$
- Parameter Θ has Beta(α,β) prior distribution
  - $g(\theta|\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Joint gpdf for (X, Θ):
  - $f(x|\theta)g(\theta|\alpha,\beta) = \binom{n}{x}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} = \binom{n}{x}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}$

    $\alpha^* = \alpha + x$

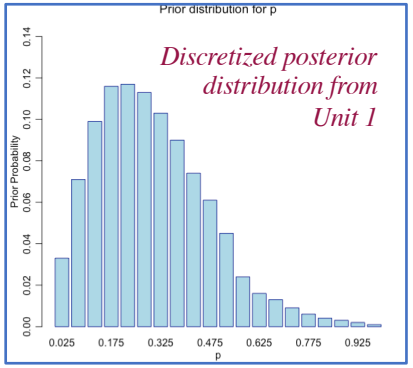
    $\beta^* = \beta + n - x$
- Posterior distribution for Θ is Beta(α^*, β^*)
  - $g(\theta|x,\alpha,\beta) = \frac{\Gamma(\alpha^*)\Gamma(\beta^*)}{\Gamma(\alpha^*+\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}$
- The parameters α and β of the Beta distribution are called the *virtual counts* (or pseudo counts)

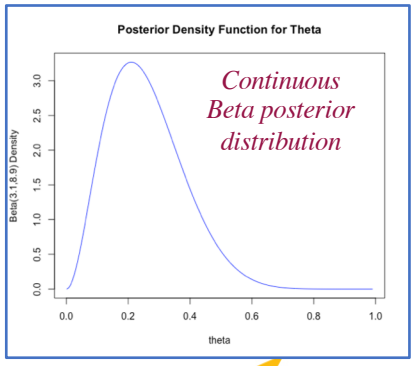
Disease Example (from Unit 1)

- The problem:
  - "Middle-aged male patients who complain of symptom S are drawn at random from a population with a proportion Θ who have disease D"
    - Observations are independent & identically distributed (iid) Bernoulli(Θ) draws or a single Binomial(n,θ) draw
  - Prior distribution for Θ is Beta(2.1,4.9)
    - Mean is 0.30, 90% credible interval is [0.070, 0.598]
    - (To find find interval use qbeta(p,2.1, 4.9) for p=0.05, 0.95)
  - Observations: 3 of 10 patients have disease
  - Objective is to use the data to draw inferences about Θ
- Conditional on X the posterior distribution of Θ is Beta(5.1,11.9)
  - Posterior density function:

$$g(\theta|x) = \frac{\Gamma(5.1)\Gamma(11.9)}{\Gamma(17)}\theta^{5.1}(1-\theta)^{11.9}$$

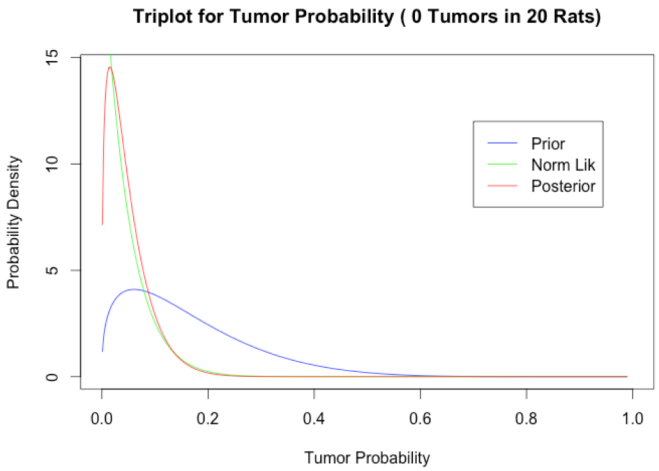
  - Posterior mean is 0.30, 90% credible interval for Θ : [0.137, 0.491]
    - To find interval use qbeta(p,5.1,14.9) for p=0.05, 0.95





Example: Rat Tumor Probability

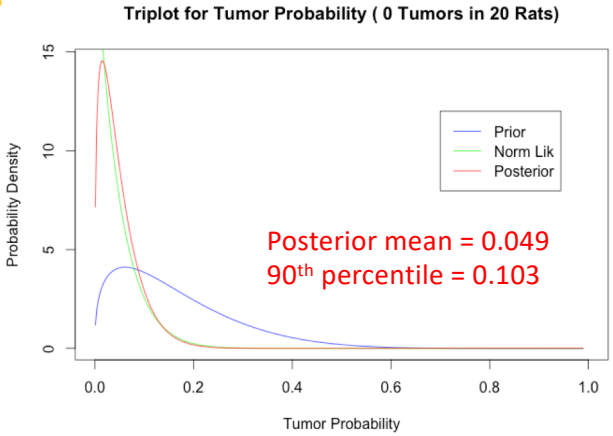
- Observations: tumor incidence in rats given a drug being tested for safety
- Analyst works with the expert to specify a Beta(1.4, 7.2) prior distribution for Θ
  - Mode of the prior distribution is 0.06
  - Mean of the prior distribution is 0.163
  - Median of the prior distribution is 0.136
  - 90th percentile of the prior distribution is 0.331
- No tumors were observed in 20 rats
- Posterior distribution for Θ is Beta(1.4, 27.2)
  - Mode of the posterior distribution is 0.015
  - Mean of the posterior distribution is 0.049
  - Median of the posterior distribution is 0.039
  - 90th percentile of the posterior distribution is 0.103
- *Note that posterior mean of tumor probability is non-zero even when sample frequency is zero*

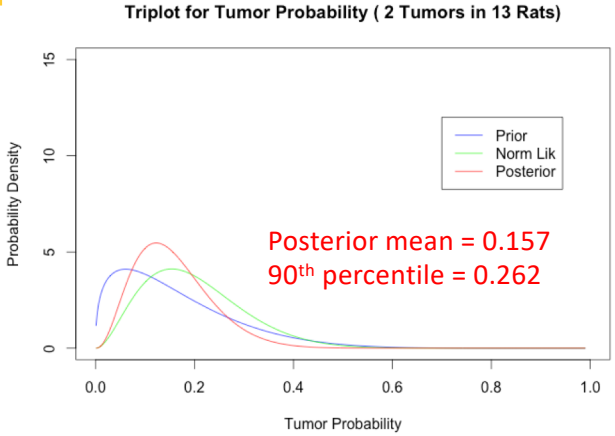


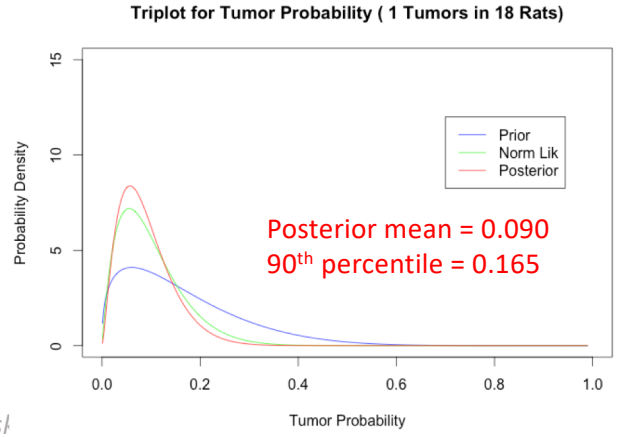
Data taken from Tarone, R. E. The Use of Historical Control Information in Testing for a Trend in Proportions. *Biometrics* **38**, 215-220. 1982. This article reports data from 71 studies of tumor incidence in rats.

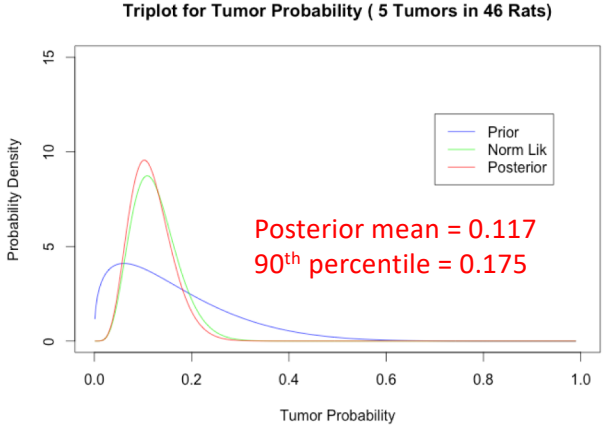
Posterior Distributions from Different Studies

(same prior distribution)









Data taken from Tarone (1982)

The Beta / Binomial Conjugate Pair

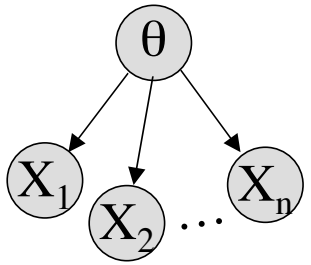
- The binomial and beta families of distributions are a *conjugate pair*:

  IF Observations $X_1, \ldots, X_n$ are a random sample from a binomial(m, Θ) and prior distribution for Θ is beta(α, β)

  THEN Posterior distribution for Θ is beta(α^*, β^*), another member of the conjugate family, where $\alpha^* = \alpha + \sum_i x_i$ and $\beta^* = \beta + nm - \sum_i x_i$

Marginal Likelihood

- The marginal likelihood $f(\underline{x})$ is the probability distribution for the data integrated over all values of the parameter:
  - $$f(\underline{x}) = \int_{\theta} \prod_{i=1}^{n} f(x_i \mid \theta) g(\theta) d\mu(\theta)$$
    - This is *not* a product of individual density functions
    - Observations are *conditionally independent* given parameter but *not independent* – each new observation changes our predictions for subsequent observations



- Marginal likelihood combines two sources of uncertainty:
  - The parameter is uncertain
  - Given the parameter, the observations are uncertain

Uses of Marginal Likelihood

- The denominator in Bayes Rule: $g(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)g(\theta)}{f(\underline{x})}$
- To predict future observations given our current knowledge (including *both* uncertainty about θ and uncertainty about data given θ)
- To measure how surprising an observed data set is given our prior assumptions about θ and the form of the likelihood function
- To compare hypotheses about the parameter (e.g., different functional forms, different dimensionality)
- To calculate weights in Bayesian model averaging

Joint Marginal Likelihood for Poisson / Gamma Observations

- Prior Gamma density function

$$g(\lambda|\alpha,\beta) = \underbrace{\frac{1}{\beta^{\alpha}\Gamma(\alpha)}}_{\text{Normalizing constant}} \lambda^{\alpha-1} e^{-\lambda/\beta}$$

- Poisson likelihood function

$$f(\underline{x}\,|\,\lambda) \underbrace{\frac{1}{\prod_i x_i!}}_{\text{Data-dependent constant}} \lambda^{\sum_i x_i} \exp\{-n\lambda\}$$

- Posterior Gamma density function

$$g(\lambda|\alpha^*,\beta^*) = \underbrace{\frac{1}{(\beta^*)^{\alpha^*}\Gamma(\alpha^*)}}_{\text{Normalizing constant}} \lambda^{\alpha^*-1} e^{-\lambda/\beta^*}$$

$$\alpha^* = \alpha + \sum_i x_i$$

$$\beta^* = \frac{1}{\beta^{-1} + n}$$

The marginal likelihood is equal to the ratio of prior and posterior normalizing constants times a function that depends on data but not the hyperparameters

- Marginal likelihood:

$$f(\underline{x}\,|\,\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \frac{1}{\prod_i x_i!} \int_{\lambda} \lambda^{\alpha+\sum_i x_i - 1} e^{-\lambda(\beta^{-1}+n)}\, d\lambda = \frac{(\beta^*)^{\alpha^*}\Gamma(\alpha^*)}{\beta^{a}\Gamma(\alpha)} \frac{1}{\prod_i x_i!}$$

Poisson / Gamma Predictive Distribution

- Predictive distribution for next Poisson observation X with prior distribution $\Lambda \sim \text{Gamma}(\alpha,\beta)$

$$f(x\,|\,\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\frac{1}{x!}\int_{\lambda}\lambda^{\alpha+x-1}\exp\left\{-\lambda\left(\beta^{-1}+1\right)\right\}d\lambda$$

$$= \frac{(\beta^{*})^{\alpha^{*}}\Gamma(\alpha^{*})}{\beta^{a}\Gamma(\alpha)}\frac{1}{x!} = \frac{\left(\frac{\beta}{1+\beta}\right)^{\alpha+x}\Gamma(\alpha+x)}{\beta^{a}\Gamma(\alpha)}\frac{1}{x!}$$

*$f(x|\alpha,\beta)$ is the pmf for a **negative binomial distribution** with size=α and prob=$1/(1+\beta)$*

- Predictive distribution for sum $Y = \Sigma X_i$ of next *n* Poisson observations (sufficient statistic for next sample):

$$f(y\,|\,\alpha,\beta) = \frac{\left(\frac{n\beta}{1+n\beta}\right)^{\alpha+y}\Gamma(\alpha+y)}{(n\beta)^{a}\Gamma(\alpha)}\frac{1}{y!}$$

$Y \mid \Lambda$ has Poisson($n\Lambda$) distribution

Λ has Gamma(α,β) prior distribution

*$f(y|\alpha,\beta)$ is the pmf for a **negative binomial distribution** with size=α and prob=$1/(1+n\beta)$*

Bayesian Belief Dynamics
Using the Marginal Likelihood

- A Bayesian forecaster uses the marginal likelihood to predict each new block of observations using information from previous blocks
  - Begin with prior distribution $\Lambda \sim \text{Gamma}(\alpha, \beta)$
  - Use prior marginal likelihood to predict sum $Y_{1:n} = \sum_{1:n} X_i$ of next n observations
    - $Y_{1:n} \sim \text{NegBinom}(\alpha, 1/(1+n\beta))$
  - Observe $Y_{1:n}$ and update according to Bayes rule to obtain posterior distribution $\Lambda \sim \text{Gamma}(\alpha^*, \beta^*)$
  - Use posterior marginal likelihood to predict sum $Y_{n+1:2n} = \Sigma_{n+1:} X_i$ of next n observations
    - $Y_{n+1:2n} \sim \text{NegBinom}(\alpha^*, 1/(1+n\beta^*))$
  - Continue as additional blocks of data are observed
- As number of observations becomes large the predictive distribution becomes approximately equal to $f(X_{n+1} | \lambda_{MAP})$ where λ_{MAP} is the posterior mode (*maximum a posteriori* estimate)
  - Prediction using MAP estimate understates uncertainty and ignores dependence among future observations due to parameter uncertainty
  - Using the marginal likelihood to predict gives better assessment of uncertainty when the parameter is uncertain

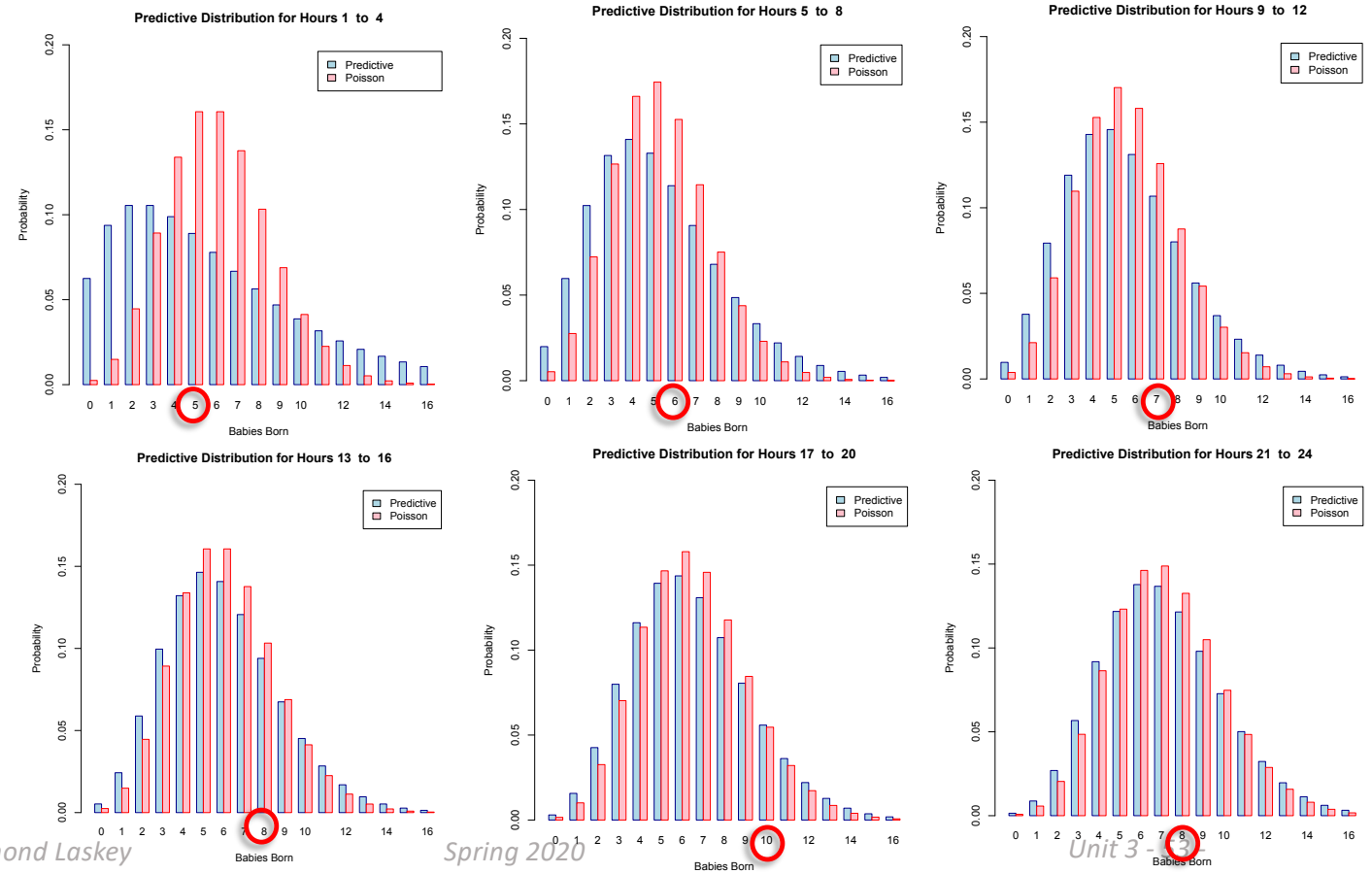
Example: Australian Babies

- Objective: predict Australian baby births in 4-hour blocks
- Earlier, we fit a Gamma(2, 0.75) prior distribution to expert's judgments
  - Expect about 1.5 births per hour

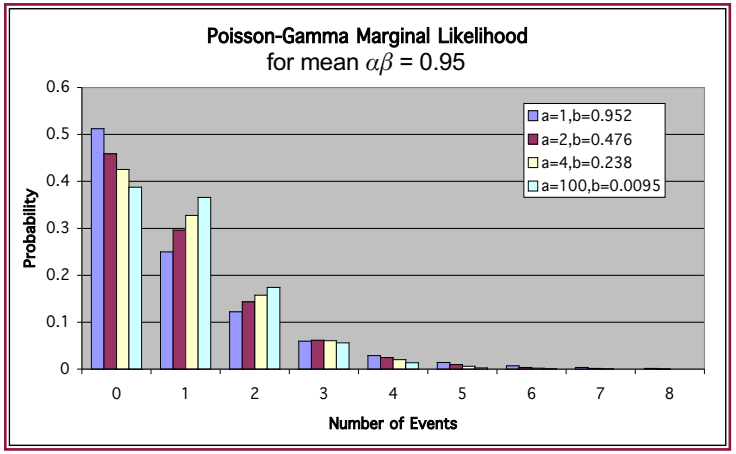
- Hours 1 to 4
  - $\Lambda \sim$ Gamma(2, 0.75)
  - Predict $Y_{1:4} \sim$ NegBinom(2,0.25)
  - Observe 5 births
- Hours 5 to 8
  - $\Lambda \sim$ Gamma(7, 1/5.33)
  - Predict $Y_{5:8} \sim$ NegBinom(7,0.571)
  - Observe 6 births
- Hours 9 to 12
  - $\Lambda \sim$ Gamma(13, 1/9.33);
  - predict $Y_{9:12} \sim$ NegBinom(13,0.7)
  - observe 7 births
- Hours 13 to 16
  - $\Lambda \sim$ Gamma(20,1/13.33);
  - predict $Y_{13:16} \sim$ NegBinom(20,0.769)
  - observe 8 births
- Hours 17 to 20
  - $\Lambda \sim$ Gamma(28, 1/17.33);
  - Predict $Y_{17:20} \sim$ NegBinom(28,0.813)
  - Observe 10 births
- Hours 20 to 24:
  - $\Lambda \sim$ Gamma(38, 1/21.33);
  - Predict $Y_{17:20} \sim$ NegBinom(38,0.842)
  - Observe 8 births

Predicting Baby Births in 4-Hour Blocks

Compare NegBinom(α, 1/(1+$n\beta$)) and Poisson($\alpha\beta$) Distributions

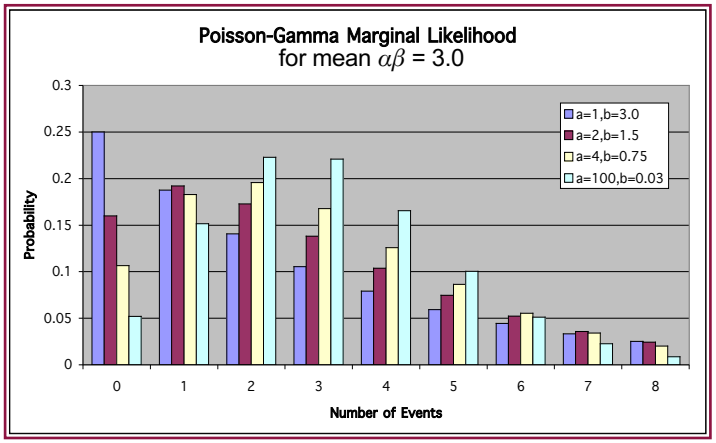


Marginal Likelihood for Poisson-Gamma Pair: Plots of Probability Mass Function



- *These plots show negative binomial marginal likelihoods for prior distributions for and different values of α and β with prior mean αβ held fixed*
- *Holding αβ fixed and increasing α increases the amount of information about the parameter Λ while holding central tendency fixed*
- *As α increases, the plot becomes more similar to a Poisson distribution with the same mean*

The marginal likelihood incorporates uncertainty from observations and lack of knowledge about parameter



α governs shape; β governs scale

Joint Marginal Likelihood for Exponential / Inverse-Gamma Observations

- Prior Inverse-Gamma density function

$$g(\theta \mid \alpha, \beta) = \underbrace{\frac{1}{\beta^{\alpha}\Gamma(\alpha)}}_{\text{Normalizing constant}} \theta^{-(\alpha+1)} e^{-1/(\theta\beta)}$$

- Exponential likelihood function

$$f(\underline{x} \mid \theta) = \theta^{-n} e^{-\sum_i x_i / \theta}$$

(*Data-dependent constant is equal to 1*)

- Posterior Inverse-Gamma density function

$$g(\theta \mid \alpha^*, \beta^*) = \underbrace{\frac{1}{(\beta^*)^{\alpha^*}\Gamma(\alpha^*)}}_{\text{Normalizing constant}} \theta^{-(\alpha^*+1)} e^{-1/(\theta\beta^*)} \qquad \alpha^* = \alpha + n \qquad \beta^* = \frac{1}{\beta^{-1} + \sum x_i}$$

The marginal likelihood is equal to the ratio of prior and posterior normalizing constants times a function that depends on data but not the hyperparameters

- Marginal likelihood:

$$f(\underline{x} \mid \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int \theta^{-(\alpha+n+1)} e^{-(\sum_i x_i + \beta^{-1})/\theta} d\theta = \frac{(\beta^*)^{\alpha^*}\Gamma(\alpha^*)}{\beta^{\alpha}\Gamma(\alpha)} = \frac{\left(\frac{1}{\beta^{-1}+n}\right)^{-\alpha}\Gamma(\alpha + \sum x_i)}{\beta^{\alpha}\Gamma(\alpha)}$$

Exponential / Inverse-Gamma Predictive Distribution

- Predictive distribution for next exponential observation X if our current prior information is $\Theta \sim$ Inverse-Gamma(α,β)

$$f(x \mid \alpha,\beta) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{\left(\frac{\beta}{1+\beta x}\right)^{\alpha+1}}{\beta^{\alpha}} = \alpha \frac{\beta}{(1+\beta x)^{\alpha+1}}$$

Useful identity: $\Gamma(\alpha)=(\alpha-1)\,\Gamma(\alpha-1)$

$f(x|\alpha,\beta)$ is the pdf for a gamma-gamma distribution with shape1=α, shape2=1, scale= β (or rate 1/ β)

- Predictive distribution for sum $Y = \Sigma X_i$ of next *n* exponential observations
  - Y is sufficient statistic for next sample and has Gamma(n, Θ) distribution:

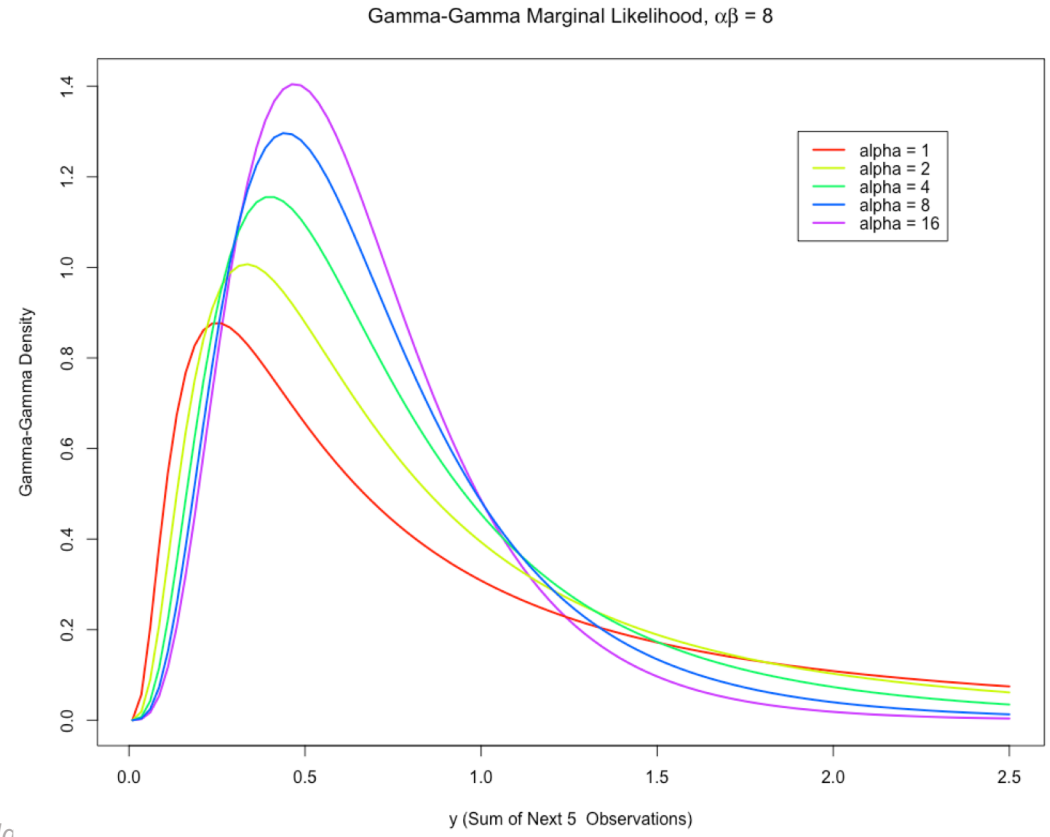
$$f(y \mid \alpha,\beta) = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \frac{\left(\frac{\beta}{1+\beta y}\right)^{\alpha+n}}{\beta^{\alpha}} \frac{y^{n-1}}{(n-1)!} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \frac{\beta^{n}}{(1+\beta y)^{\alpha+n}} \frac{y^{n-1}}{(n-1)!}$$

$Y \mid \Theta$ has Gamma(n, Θ) distribution

Θ has Inverse-Gamma(α,β) prior distribution

$f(y|\alpha,\beta)$ is the pdf for a gamma-gamma distribution with shape1=α, shape2=n, scale= β (or rate 1/ β)

Comparing Gamma-Gamma Marginal Likelihoods
with different shape parameters



As amount of information becomes larger the plot becomes closer to a Gamma distribution with shape 5 and scale 1/8

The marginal likelihood incorporates uncertainty from observations <u>and</u> lack of knowledge about parameter

In R, the BAEssd package has density function and random generation for gamma-gamma distribution

Joint Marginal Likelihood for Binomial / Beta Observations

- Prior Beta density function

$$g(\theta \mid \alpha,\beta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{Normalizing constant}} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Likelihood function:
  r iid observations from Binomial(n,Θ) distribution

$$f(\underline{x} \mid \theta) = \underbrace{\prod_{i=1}^{r}\binom{n}{x_i}}_{\text{Data-dependent constant}} \theta^{\sum_i x_i}(1-\theta)^{nr-\sum_i x_i}$$

- Posterior Beta density function

$$g(\theta \mid \alpha^*,\beta^*) = \underbrace{\frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}}_{\text{Normalizing constant}} \theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}$$

$$\alpha^* = \alpha + \sum x_i$$

$$\beta^* = \beta + nr - \sum x_i$$

The marginal likelihood is equal to the ratio of prior and posterior normalizing constants times a function that depends on data but not the hyperparameters

- Marginal likelihood:

$$f(\underline{x} \mid \alpha,\beta) = \int_\theta \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=1}^{r}\binom{n}{x_i} \theta^{\sum_i x_i+\alpha-1}(1-\theta)^{nr-\sum_i x_i+\beta-1} d\theta$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=1}^{r}\binom{n}{x_i} \frac{\Gamma(\alpha+\sum_i x_i)\Gamma(\beta+nr-\sum_i x_i)}{\Gamma(\alpha+\beta+nr)}$$

Binomial / Beta Predictive Distribution

- Predictive distribution for next Binomial observation X if our current prior information is $\Theta \sim \text{Beta}(\alpha,\beta)$

$$f(x \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{n}{x}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}$$

f(x|α,β) is the pmf for a beta-binomial distribution with shape1=α, shape2 = β, size=n (or probability α/(α + β), overdispersion α + β, size = n)

  - If *n* is equal to 1 the predictive pmf simplifies to:

$$f(x \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(\beta+1-x)}{\Gamma(\alpha+\beta+1)} = \begin{cases} \frac{\alpha}{\alpha+\beta} & \text{if } x=1 \\ \frac{\beta}{\alpha+\beta} & \text{if } x=0 \end{cases}$$

Probability of success on next trial is α/(α+β)

- Predictive distribution for sum $Y = \sum X_i$ of next *r* Binomial(*n*, Θ) observations (sufficient statistic for next sample):

$$f(y \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{nr}{y}\frac{\Gamma(\alpha+y)\Gamma(\beta+nr-y)}{\Gamma(\alpha+\beta+nr)}$$

$Y \mid \Theta$ has Binomial(nr,Θ) distribution

Θ has Beta(α, β) prior distribution

f(y|α,β) is the pmf for a beta-binomial distribution with shape1=α, shape2 = β, size=nr (or probability α/(α + β), overdispersion α + β, size = nr)

The Beta-Binomial Family of Distributions

- Sample space: non-negative integers
- Parameters:
  - Standard parameterization: size parameter n, shape parameters α and β
  - Alternative parameterization: size parameter n, probability parameter $p = \frac{\alpha}{\alpha+\beta}$, overdispersion parameter $m = \alpha + \beta$
- Probability mass function:

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}$$

- Moments:

$$E[X|\alpha,\beta,n] = \frac{n\alpha}{\alpha+\beta} = np \quad Var[X|\alpha,\beta,n] = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = np(1-p)\frac{(m+n)}{m+1}$$

- Functions in R
  - `[d,p,q,r]betabinom` in `rmutil` package (uses *p*, *m*, *n* parameterization)
  - `[d,p,q,r]bb` in `tailrank` package (uses α, β, *n* parameterization; depends on Biobase package which must be installed from bioconductor.org)

Prediction Using the Beta-Binomial Distribution

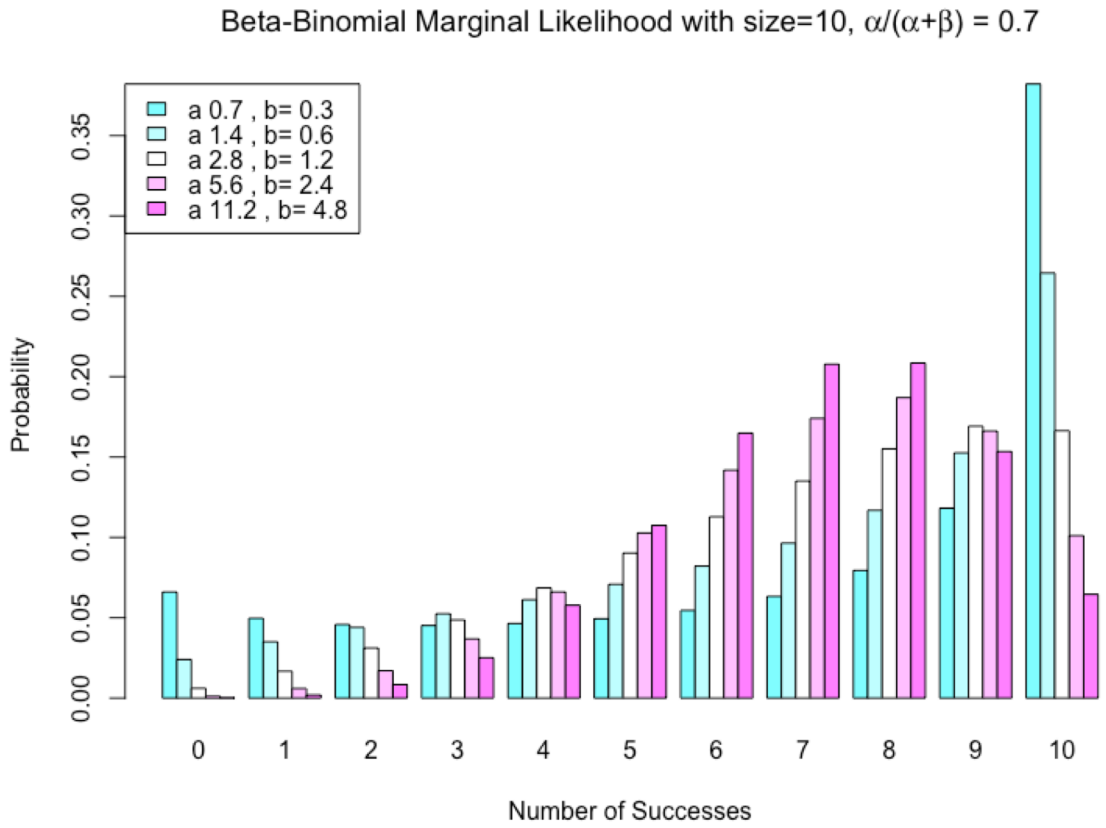
- Assume:
  - Our current information about Θ is expressed by a Beta(α, β) distribution
  - We want to predict the number of successes in a random sample of size *r* from a Binomial(*n*, Θ) distribution
- We use the beta-binomial predictive probability mass function:

$$f(y \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{nr}{y}\frac{\Gamma(\alpha+y)\Gamma(\beta+nr-y)}{\Gamma(\alpha+\beta+nr)}$$

```
dbetabinom(y, n*r, alpha/(alpha+beta), alpha + beta) # uses rmutil package
```

- We use the *same* predictive distribution to predict the number of successes in *nr* Bernoulli(Θ) trials, or a single Binomial(*nr*, Θ) trial
- After we have observed *k* total successes out of *nr* trials, we update our knowledge to a Beta(α+*k*, β+*nr*-*k*) distribution
- We use this updated knowledge to predict our *next* sample

Comparing Beta-Binomial Marginal Likelihoods
with common mean, different overdispersions



As amount of information becomes larger the plot becomes close to a Binomial distribution with the same mean

The marginal likelihood incorporates uncertainty from observations <u>and</u> lack of knowledge about parameter

Marginal Likelihood: Summary

- The marginal likelihood is the marginal distribution for the data integrated over the parameter
  - Randomly sampled observations are independent conditional on parameter but are not independent marginally
- We can use the marginal likelihood to predict the sufficient statistic for the next batch of observations
  - Incorporates uncertainty about both parameter and observation conditional on parameter

Likelihood	Prior	Marginal Likelihood
Poisson	Gamma (for rate)	Negative binomial
Exponential	Inverse-Gamma (for mean)	Gamma-Gamma
Binomial	Beta (for probability)	Beta-binomial

Expressing Vague Prior Information with Reference Priors

- If prior information is vague compared to the information in the sample then it is not worth much effort to do a careful assessment of your prior distribution.
- For many problems there are convenient standard distributions we use to approximate the posterior distribution when our prior knowledge is vague compared with the data we expect to get
- We call these "reference priors"
- Examples:
  - Jeffreys Rule and invariant priors
  - Uniform prior (posterior distribution is normalized likelihood)

Improper Reference Priors

- Sometimes reference prior distributions are "improper" (density function integrates to infinity)
  - Example: a uniform distribution on the positive real numbers
- It may be that even if $g(\Theta)$ is improper, $f(x|\Theta)g(\Theta)$ has a finite integral
  - If $f(x|\Theta)g(\Theta)$ has a finite integral, we get a well-defined answer when we plug $f(x|\Theta)$ and $g(\Theta)$ into Bayes Rule:

$$g(\theta \mid x) = \frac{f(x \mid \theta)g(\theta)}{\int f(x \mid \theta)g(\theta)d\mu(\theta)}$$

  - Improper reference priors can be useful if prior knowledge is vague and sample is informative

Ignorance and Invariance to Parameterization

- We would like to have standard prior distributions that express the idea that we have very little knowledge of the parameter
- Sometimes a uniform prior is used to express ignorance
  - Uniform distribution is improper when the parameter space is infinite
  - Posterior distribution is often proper
- Example: Uniform prior distribution for Poisson rate parameter
  - If we are ignorant about Λ, the rate parameter of a Poisson distribution, we might specify a uniform distribution for Λ
  - A uniform distribution $g(\lambda) \propto 1$ is a limit of a gamma$(1, \beta)$ distribution as $\beta \rightarrow \infty$
  - The posterior distribution given n observations with sufficient statistic X_i is Gamma$(\sum X_i + 1, 1/n)$, a proper distribution
- Issue: lack of invariance to how we parameterize the likelihood
  - If Λ has a uniform prior distribution, then the density function for $\Theta = \Lambda^{-1}$ is $g(\theta) \propto \theta^{-2}$, which is not uniform!
  - If we are ignorant about the rate, shouldn't we also be ignorant about the scale?
  - If a uniform distribution means ignorance, which distribution should be uniform?

Invariance to Transformation: Jeffreys Prior

- Jeffreys prior uses a quantity called the Fisher information

$$I(\theta;\underline{X}) = -E_{\underline{x}}\left[\partial^2 \log f(\underline{X}\,|\,\theta)/\partial\theta^2\right]$$

  - Note: Some texts use the notation $I(\theta \mid x)$
  - This notation is confusing because it suggests conditioning on x
  - Our notation $I(\theta;\ X)$ emphasizes that the expectation is taken over the random variable X, while θ is treated as a fixed parameter
  - For an iid sample of size n, $I(\theta;\ X)\ =\ nI(\theta;\ X)$, where $I(\theta;\ X)$ is the Fisher information for a single observation
  - Cramer-Rao lower bound: If $\hat{\theta}$ is any unbiased estimator of θ then $\mathrm{Var}(\hat{\theta}) \geq I(\theta;\underline{X})$
- Jeffreys' Prior $g(\theta) \propto \sqrt{I(\theta;X)}$ *Square root of information in one observation*
- Jeffreys' prior is invariant to changes of variable:
  - If we re-parameterize as $\psi\ =\ h(\theta)$ then $I(\psi;\ X)\ =\ I(\theta;\ X)(d\theta/d\psi)^2$
  - If we use the reference prior $g(\theta) \propto \sqrt{I(\theta;X)}$ for Θ and apply standard change of variables rule then we obtain the reference prior $g(\psi) \propto \sqrt{I(\psi;X)}$ for Ψ

Examples of Jeffreys Prior

- If X is Binomial with unknown success probability Θ then
  - Jeffreys' prior is $g(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, or a Beta (1/2, 1/2) distribution
- If X has a Poisson distribution with unknown mean Λ
  - Jeffreys' prior is $g(\lambda) \propto \lambda^{-1/2}$
  - This is the limit of the Gamma(1/2, β) distribution as β tends to infinity
  - This distribution is improper (integrates to infinity)
- If X has an exponential distribution with unknown mean Θ
  - Jeffreys' prior is $g(\theta) \propto \theta^{-1}$
  - This is the limit of the Inverse-gamma(α, β) distribution as α tends to zero and β tends to infinity
  - This distribution is improper (integrates to infinity)

Example: Jeffreys Prior for Poisson Parameter

$$f(x \mid \lambda) = \left(\prod_i x_i!\right)^{-1} \lambda^{n\bar{x}} e^{\lambda n}$$

$$\log\left(f(x \mid \lambda)\right) = -\log\left(\prod_i x_i!\right)^{-1} + n\bar{x}\log\lambda + \lambda n$$

$$\frac{\partial}{\partial\lambda}\log\left(f(x \mid \lambda)\right) = \frac{n\bar{x}}{\lambda} + n$$

$$\frac{\partial^2}{\partial\lambda^2}\log\left(f(x \mid \lambda)\right) = -\frac{n\bar{x}}{\lambda^2}$$

$$I(\lambda;\underline{x}) = -E\left[\frac{\partial^2}{\partial\lambda^2}\log\left(f(x \mid \lambda)\right)\right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

$$\sqrt{I(\lambda;\underline{x})} = \left(\frac{n}{\lambda}\right)^{1/2}$$

$$g_{Jeffreys}(\lambda) \propto \lambda^{-1/2}$$

$$g(\theta) \propto \sqrt{I(\theta;X)}$$

$$I(\theta;\underline{X}) = -E_{\underline{x}}\left[\partial^2 \log f(\underline{X} \mid \theta)/\partial\theta^2\right]$$

This is the limit of a Gamma density with $\alpha = 1/2$ and $\beta \to \infty$

Interpreting a Conjugate Prior Distribution

- Prior information can be thought of as "memory" of a prior sufficient statistic
  - A parameter called the "virtual count" or "pseudo count" represents the amount of prior information
  - Other parameter(s) represent other "remembered" features

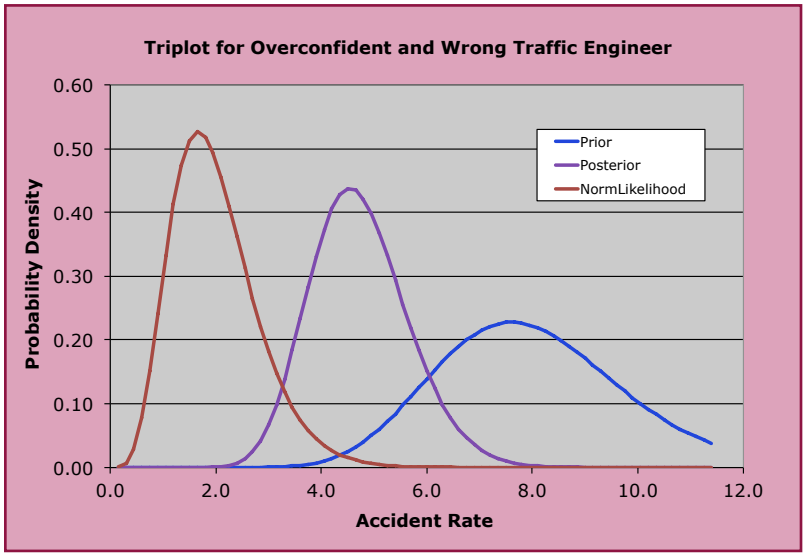
Recall from Unit 2: A sufficient statistic is a data summary (function of sample of observations) such that the observations are independent of the parameter given the sufficient statistic

Prior as Remembered Past Sample: Example

- For the Exponential/Inverse-Gamma and Poisson/Gamma pairs:
  - α represents the remembered number of events
  - $\alpha\beta$ represents the remembered mean (number of events per unit time)
  - β^{-1} represents the remembered period of observation
- The shape hyperparameter α is sometimes called the "virtual count" (or pseudo count)
  - The virtual count can be thought of as measuring "amount of information"
  - As virtual count grows large the distribution becomes more Gaussian in shape and highly concentrated about the mode
  - If we increase the virtual count α, the relative contribution of the prior distribution increases relative to the sample

Example: Overconfident But Wrong Prior

- Engineer assesses Gamma(20,0.4) distribution for daily accident rate
  - Expected value 8 accidents per day
  - Standard deviation 1.8
- Observe 6 accidents in 3 days of collecting data
- Posterior distribution is Gamma(26,0.18)
  - Expected value 4.7 accidents per day
  - Standard deviation 0.9
- Posterior distribution is centered on values considered very unlikely a priori and not favored by data



"A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule." – Stephen Senn

- *This joke is actually incorrect – only strong prior and strong but inconsistent likelihood can lead to strong posterior inconsistent with both*
- *We do need to beware of strong and overconfident prior*

Recap: Statistical Inference

- General inference problem:
  - Parameter $\underline{\Theta}$ is unknown and unobservable
  - Use observable data $\underline{X}=\underline{x}$ to draw inferences about $\underline{\Theta}$
- Bayesian approach:
  - Quantify prior knowledge about ($\underline{\Theta}$, $\underline{X}$) by conditional distribution $f(\underline{x}|\underline{\theta})$ for $\underline{X}$ and prior distribution $g(\underline{\theta})$ for $\underline{\Theta}$
  - Posterior distribution $g(\underline{\theta}|\underline{x})$ quantifies new knowledge about $\underline{\Theta}$ after observing $\underline{X}=\underline{x}$
- Canonical statistical inference problem:
  - $\underline{X}=X_1, \ldots, X_n$ is an iid sample from a probability distribution with gpdf $f(\underline{x}|\underline{\theta})$
  - Prior distribution $g(\underline{\theta})$ is given for unknown parameter $\underline{\Theta}$
  - The goal is to find the posterior distribution $g(\underline{\theta}|\underline{x})$ and properties of $g(\underline{\theta}|\underline{x})$
- Finding the posterior distribution
  - Conjugate distributions simplify analysis when available and appropriate
  - Reference distributions are simple and convenient for large sample sizes
  - Be sure to check adequacy of prior distribution

Recap: Conjugate Pairs of Distributions

- A gpdf family $g(\theta|\alpha)$ is *conjugate* to the gpdf family $f(x|\theta)$ if it is *closed under sampling* from $f(x|\theta)$, that is:

  IF Data $X_1, \ldots, X_n$ are a random sample from $f(x|\theta)$ AND prior distribution for unknown parameter Θ is $g(\theta|\alpha)$

  THEN Posterior distribution for parameter Θ is $g(\theta|\alpha^*)$, another member of the conjugate family

- There is a simple updating rule to find α^* from α and the observations
- Explicit form for the marginal likelihood supports predictions that include uncertainty about both parameter and observations given parameter

Remarks on Conjugate Pairs

- Analysis with conjugate pairs is convenient:
  - Easy to specify with virtual sufficient statistic
  - Exact posterior distribution and marginal likelihood for sufficient statistic
- However:
  - Conjugate family is limited in the kinds of prior knowledge it can express
    - Prior information is "like" a previous sample with known virtual sufficient statistic
    - Cannot express uncertainty about virtual count ("amount" of prior information)
  - Conjugate family may not be robust to mis-specification of prior parameters (especially virtual count)
- Conjugate distributions can be extended to more flexible prior distributions that retain:
  - Ease of specification
  - Ease of computation

Summary: Conjugate Pairs

Data $\underline{X}=(X_1,\ldots,X_n)$	**Prior**	**Sufficient Statistic**	**Posterior**	**Marginal likelihood for sufficient statistic**
$\underline{X}\|\Lambda \sim$ Poisson(Λ)	$\Lambda \sim$ Gamma(α,β)	$Y=\sum_i X_i$	$\Lambda\|\underline{X} \sim$ Gamma(α^*,β^*) $\alpha^*=\alpha+\sum_i X_i$, $\beta^*=(\beta^{-1}+n)^{-1}$	$f(y\|\alpha,\beta)=\frac{\left(\frac{n\beta}{1+n\beta}\right)^{\alpha+y}\Gamma(\alpha+y)}{(n\beta)^{\alpha}\Gamma(\alpha)}\frac{1}{y!}$ $Y \sim$ NegBinom(α, $1/(1+n\beta)$)
$\underline{X}\|\Theta \sim$ Exponential(Θ)	$\Theta \sim$ Inverse-Gamma(α,β)	$Y=\sum_i X_i$	$\Theta\|\underline{X} \sim$ Inverse-Gamma(α^*,β^*) $\alpha^*=\alpha+n$, $\beta^*=(\beta^{-1}+\sum_i X_i)^{-1}$	$f(y\|\alpha,\beta)=\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\frac{\beta^n}{(1+\beta y)^{\alpha+n}}\frac{y^{n-1}}{(n-1)!}$ $Y \sim$ gamma-gamma(α, $1/\beta$, n)
$\underline{X}\|\Theta \sim$ Binomial(r,Θ)	$\Theta \sim$ Beta(α,β)	$Y=\sum_i X_i$	$\Theta\|\underline{X} \sim$ Beta(α^*,β^*) $\alpha^*=\alpha+\sum_i X_i$, $\beta^*=\beta+(nr-Y)$	$f(y\|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{nr}{y}\frac{\Gamma(\alpha+y)\Gamma(\beta+nr-y)}{\Gamma(\alpha+\beta+nr)}$ $Y \sim$ beta-binomial(α, β, nr)

This table is a useful reference

Summary and Synthesis

- A conjugate prior/posterior pair is closed under sampling
  - If observations are a random sample from a likelihood family that has a conjugate prior family, and prior is from the conjugate prior family, then so is the posterior distribution
  - There is a simple updating rule to find the posterior hyperparamter
- With conjugate families there is an explicit marginal likelihood (predictive distribution) for the sufficient statistic of a future sample
  - Marginal likelihood includes uncertainty about both the parameter and the observations given the parameter
- We examined several conjugate pairs for single-parameter models:
  - Poisson likelihood / Gamma prior
  - Exponential likelihood / Inverse-Gamma prior
  - Binomial likelihood / Beta prior
- Prior information can be thought of as "memory" of a prior sufficient statistic
- We can use reference prior distributions when we have very little prior knowledge of the parameter relative to the amount of information in the observations