# Bayesian Inference for Simple Problems

Simon Jackman

Stanford University
http://jackman.stanford.edu/BASS

February 11, 2012

# Conjugacy

- Bayes Rule says $p(\theta|y) \propto p(\theta)p(y|\theta)$
- Mantra: "Posterior is proportional to prior times likelihood"
- This math easy to do when we use prior densities that are **conjugate** with respect to the likelihood $p(y|\theta)$.

# Conjugacy

**Definition 1.2 (p15 BASS)**: Suppose a prior density $p(\theta)$ belongs to a class of parametric densities, $\mathcal{F}$. Then the prior density is said to be conjugate with respect to a likelihood $p(\mathbf{y}|\theta)$ if the posterior density $p(\theta|\mathbf{y})$ is also in $\mathcal{F}$.

- Up until the Markov-chain Monte Carlo revolution in the 1990s, Bayesian inference was almost all done with
  - conjugate priors
  - simple problems; e.g., rates and proportions (Bernoulli trials, coin-flipping), counts (Poisson), means, variances and regression (normal data).
- Bayes estimates such as the mean of the posterior density $E(\theta|y)$ have a simple mathematical form; can be computed "by hand"; are **"precision-weighted averages"** of estimates based on the prior and on the data.

# Example: coin flipping (p50)

- $y_i \in \{0, 1\}$, exchangeable
- unknown success probability $\theta \in [0, 1]$,
- data: $y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), i = 1, \ldots, n$
- $r \sim \text{Binomial}(\theta; n), r = \sum y_i$.
- binomial likelihood $p(r|\theta)$:

$$\binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

- What prior density $p(\theta)$ is conjugate wrt the likelihood $p(r|\theta)$?
- That is, what form does $p(\theta)$ have to be such that $p(\theta|r, n) \propto p(r|\theta, n)p(\theta)$ is of the same form?

# Example: coin flipping (p50)

- $y_i \in \{0, 1\}$, exchangeable
- unknown success probability $\theta \in [0, 1]$,
- data: $y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta), i = 1, \ldots, n$
- $r \sim \text{Binomial}(\theta; n), r = \sum y_i$.
- binomial likelihood $p(r|\theta)$:

$$\binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

- What prior density $p(\theta)$ is conjugate wrt the likelihood $p(r|\theta)$?
- That is, what form does $p(\theta)$ have to be such that $p(\theta|r, n) \propto p(r|\theta, n)p(\theta)$ is of the same form?
- Answer: the **Beta** density is conjugate wrt a binomial likelihood.

# Beta density

- A prior density for $\theta \in [0, 1]$ must have the properties:
  1. $p(\theta) \geq 0, \theta \in [0, 1]$.
  2. $\int_0^1 p(\theta) d\theta = 1$.
- A conjugate prior must also have the property that $p(\theta|r, n) \propto p(r|\theta, n)p(\theta)$ is of the same form as $p(\theta)$.

## Definition

**Beta density:**

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha\text{-}1} (1 - \theta)^{\beta\text{-}1},$$

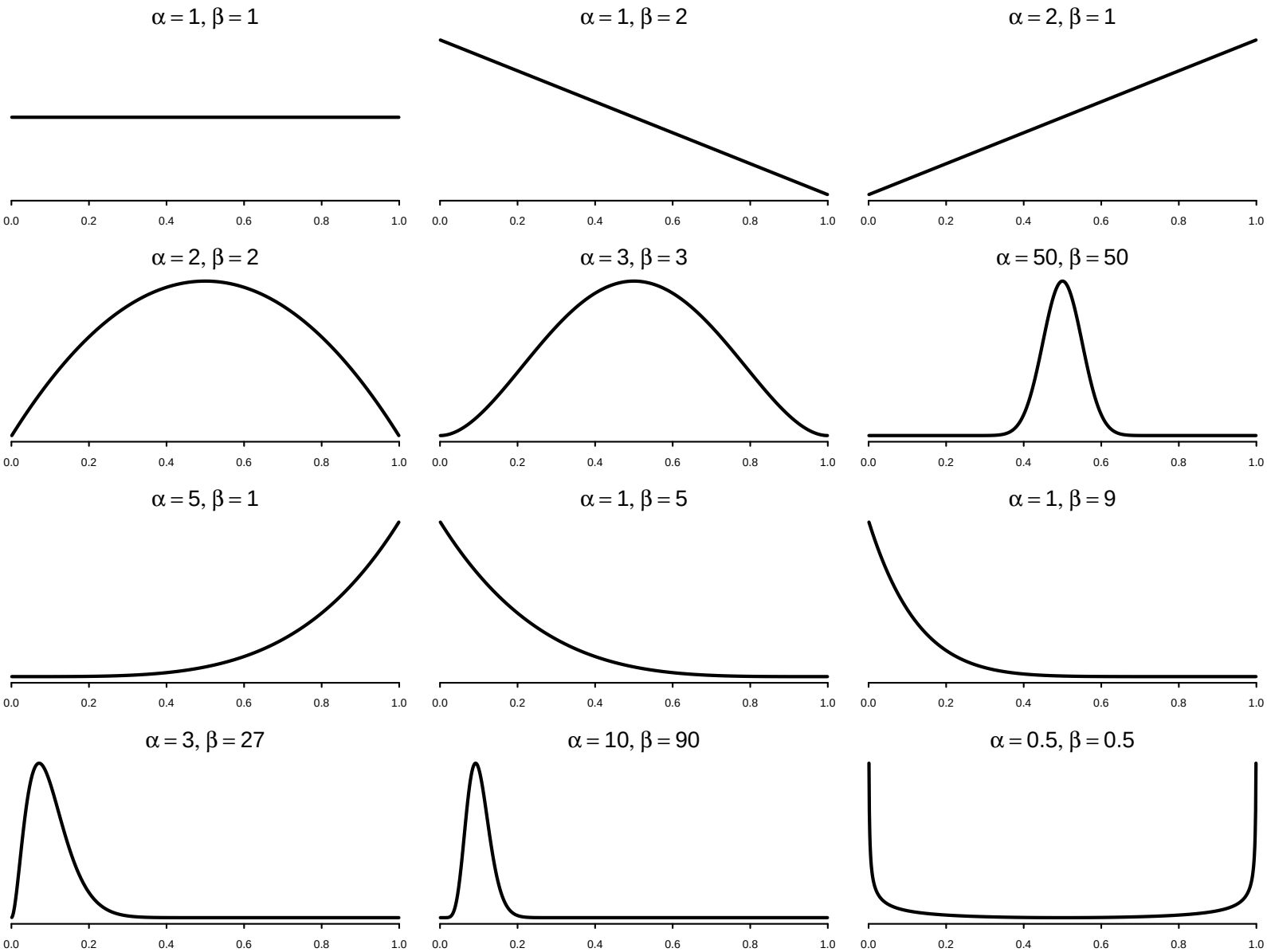
where $\theta \in [0, 1]$, $\alpha, \beta > 0$ and $\Gamma(z) = \int_0^\infty t^{z\text{-}1} \exp(-t) dt$ is the Gamma function (Definition B.19).

# Beta density

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha\text{-}1} (1 - \theta)^{\beta\text{-}1}$$

- Note that the leading terms involving the Gamma functions do not involve $\theta$: $p(\theta; \alpha, \beta) \propto \theta^{\alpha\text{-}1} (1 - \theta)^{\beta\text{-}1}$.
- A uniform density on $[0, 1]$ is a special case of the Beta density, arising when $\alpha = \beta = 1$.
- Symmetric densities with a mode/mean/median at .5 are generated when $\alpha = \beta$ for $\alpha, \beta > 1$.
- the mean, $E(\theta) = \frac{\alpha}{\alpha + \beta}$
- the mode: $\frac{\alpha - 1}{\alpha + \beta - 2}$
- the variance: $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

# Beta density



# Conjugacy of the Beta wrt the Binomial Likelihood

## Theorem

***Conjugacy of Beta Prior, Binomial Data.*** *Given a binomial likelihood over r successes in n Bernoulli trials, each independent conditional on an unknown success parameter* $\theta \in [0, 1]$*, i.e.,*

$$\mathcal{L}(\theta; r, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

*then the prior density* $p(\theta) = Beta(\alpha, \beta)$ *is conjugate with respect to the binomial likelihood, generating the posterior density* $p(\theta|r, n) = Beta(\alpha + r, \beta + n - r)$.

# Conjugacy of the Beta wrt the Binomial Likelihood

## Theorem

***Conjugacy of Beta Prior, Binomial Data.*** *Given a binomial likelihood over r successes in n Bernoulli trials, each independent conditional on an unknown success parameter* $\theta \in [0, 1]$*, i.e.,*

$$\mathcal{L}(\theta; r, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

*then the prior density* $p(\theta) = Beta(\alpha, \beta)$ *is conjugate with respect to the binomial likelihood, generating the posterior density* $p(\theta|r, n) = Beta(\alpha + r, \beta + n - r)$.

Shorthand:

$\theta \sim \text{Beta}(\alpha, \beta), r \sim \text{Binomial}(\theta, n) \Rightarrow \theta|r, n \sim \text{Beta}(\alpha + r, \beta + n - r)$.

# Conjugacy of the Beta wrt the Binomial Likelihood

**Proof of Proposition: Conjugacy of Beta Prior, Binomial Data.**
By Bayes Rule,

$$p(\theta|r,n) = \frac{\mathcal{L}(\theta;r,n)p(\theta)}{\int_0^1 \mathcal{L}(\theta;r,n)p(\theta)d\theta} \propto \mathcal{L}(\theta;r,n)p(\theta)$$

Ignoring terms that do not depend on $\theta$,

$$\begin{aligned} p(\theta|r,n) &\propto \underbrace{\theta^r(1-\theta)^{n-r}}_{\text{likelihood}} \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} \\ &= \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1} \end{aligned}$$

which is the *kernel* of a Beta density. That is,
$p(\theta|r,n) = c\theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$ where $c$ is the normalizing constant

$$\frac{\Gamma(n+\alpha+\beta)}{\Gamma(r+\alpha)\Gamma(n-r+\beta)},$$

In other words, $\theta|r,n \sim \text{Beta}(\alpha+r, \beta+n-r)$.

# Interpretation of Conjugacy in Data-Equivalent Terms

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ r &\sim \text{Binomial}(\theta, n) \\ \theta | r, n &\sim \text{Beta}(\alpha + r, \beta + n - r) \end{aligned}$$

- *as if* our prior distribution represents the information in a sample of $\alpha + \beta - 2$ independent Bernoulli trials, in which we observed $\alpha - 1$ "successes" or $y_i = 1$.
- $p(\theta) \equiv \text{Unif}(0, 1) \equiv \text{Beta}(1, 1)$ has the "data equivalent" interpretation of $\alpha + \beta - 2 = 0$ prior observations.

# Example: Florida Polling

- Florida poll, March 2000, voting intentions for the November 2000 presidential election.
- $n = 621$. Bush 45% ($n = 279$), Gore 37% (230), Buchanan 3% (19) and undecided 15% (93).
- For simplicity, we ignore the undecided and Buchanan vote share, leaving Bush with 55% of the two-party vote intentions, and Gore with 45%, and $n = 509$ respondents expressing a preference for the two major party candidates.
- We assume that the survey responses are independent, and (perhaps unrealistically) that the sample is a random sample of Floridian voters.
- Thus, the binomial likelihood is (ignoring constants that do not depend on $\theta$),

$$p(r|\theta, n) \propto \theta^{279}(1 - \theta)^{509-279}.$$

The maximum likelihood estimate of $\theta$ is
$\hat{\theta}_{\text{MLE}} = r/n = 279/509 = .548$

# Example: Florida Polling

- Prior information from previous elections.
- Forecasting model produces a forecast of Bush vote share of 49.1%, with a standard error of 2.2 percentage points.
- Convert this to a Beta density: we seek values for $\alpha$ and $\beta$ such that

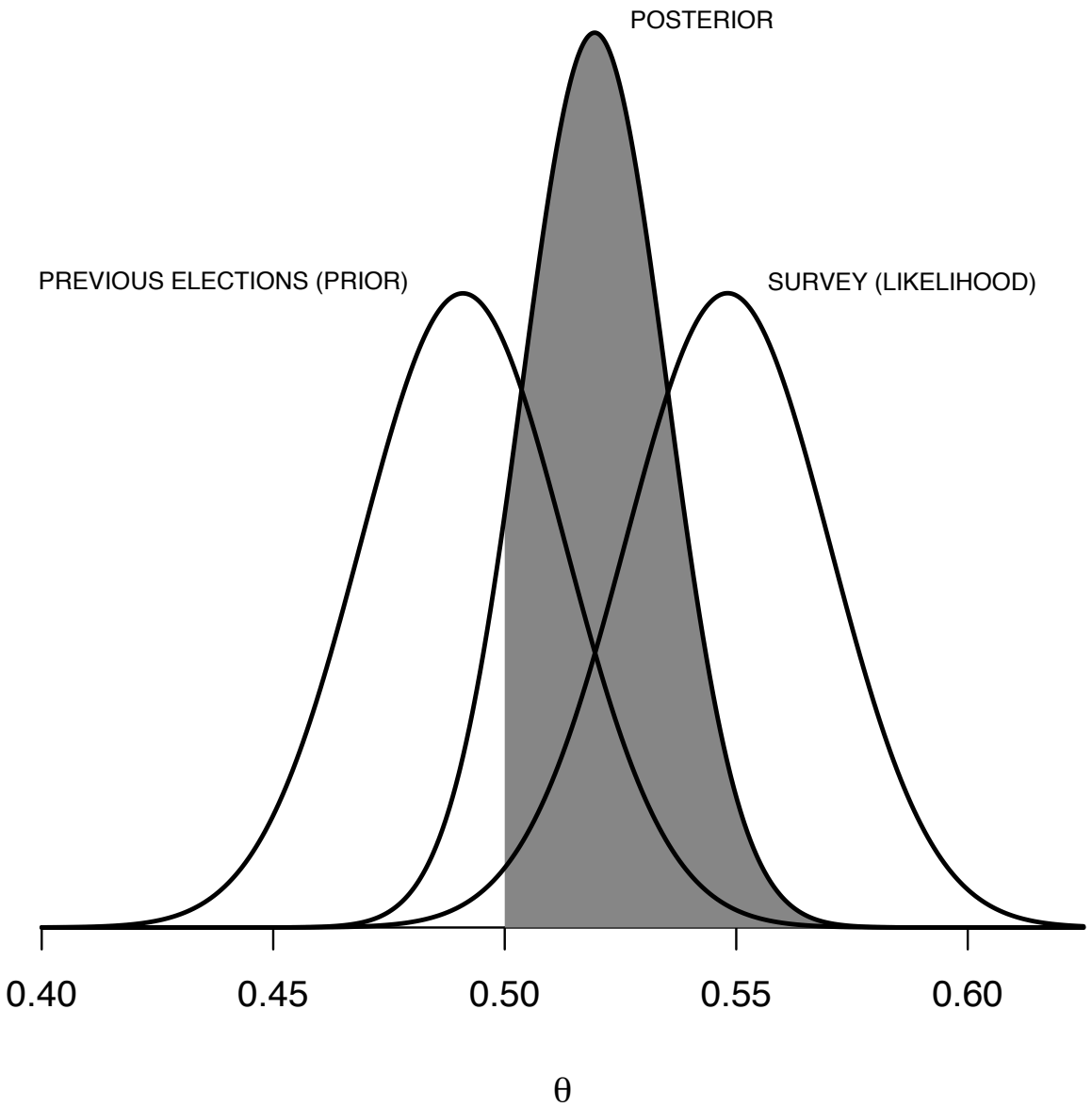
$$\begin{aligned} E(\theta; \alpha, \beta) &= \alpha/(\alpha+\beta) = .491 \\ V(\theta; \alpha, \beta) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = .022^2 \end{aligned}$$

  which yields a system of equations in two unknowns.
- Solving yields

$$\begin{aligned} \alpha &= 515.36 \times .491 = 253.04, \\ \beta &= 515.36 \times (1 - .491) = 262.32. \end{aligned}$$

- information in the previous elections is equivalent to having ran another poll with $n \approx 515$ in which $r \approx 253$ respondents said they would vote for the Republican presidential candidate.

# Example: Florida Polling



# Bayes Estimate as a Convex Combination of Prior and Data

Consider a Bayes estimate such as the posterior mean:

$$E(\theta|r,n) = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\alpha + r}{\alpha + \beta + n} = \frac{n_0\theta_0 + n\hat{\theta}}{n_0 + n}$$

where $\hat{\theta} = r/n$ is the maximum likelihood estimate of $\theta$, $n_0 = \alpha + \beta$ and $\theta_0 = \alpha/(\alpha + \beta) = E(\theta)$ is the mean of the prior density for $\theta$.
Alternatively,

$$E(\theta|r,n) = \gamma\theta_0 + (1 - \gamma)\hat{\theta}$$

where $\gamma = n_0/(n_0 + n)$, and since $n_0, n > 0$, $\gamma \in [0, 1]$. Alternatively,

$$E(\theta|r,n) = \hat{\theta} + \gamma(\theta_0 - \hat{\theta}).$$

That is, a Bayes estimate of $\theta$ --- is a *weighted average* of the prior mean $\theta_0$ and the maximum likelihood estimate $\hat{\theta}$.

# Conjugate analysis of normal data

We first consider the simple case of normal data, with unknown mean $\mu$ and known variance $\sigma^2$:

$$y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

Likelihood:

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right]$$

Conjugacy: we seek a prior for $\mu$, $p(\mu)$, s.t. the posterior

$$p(\mu|y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\mu, \sigma^2)p(\mu)$$

is in the same class as $p(\mu)$.

# Conjugate analysis of normal data

For normal data with unknown mean $\mu$, a normal prior for $\mu$ is conjugate:

## Theorem

*Let* $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$, *with* $\sigma^2$ *known, and* $\mathbf{y} = (y_1, \ldots, y_n)'$. *If* $\mu \sim N(\mu_0, \sigma_0^2)$ *is the prior density for* $\mu$, *then* $\mu$ *has posterior density*

$$\mu|\mathbf{y} \sim N\left(\frac{\mu_0\sigma_0^{-2} + \bar{y}\frac{n}{\sigma^2}}{\sigma_0^{-2} + \frac{n}{\sigma^2}}, \quad \left(\sigma_0^{-2} + \frac{n}{\sigma^2}\right)^{-1}\right).$$

Note the precision-weighted average form of the mean of the posterior density:

- precision = 1/variance
- the prior has precision $\sigma_0^{-2}$
- the MLE of $\mu$, $\bar{y}$ has precision $n/\sigma^2$.
- the posterior precision $\sigma_0^{-2} + n/\sigma^2$ is the sum of the prior precision and the data precision.

# Conjugate analysis, normal data, mean and variance unknown

- Same model as in previous section: $y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.
- Likelihood:

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right]$$

- Parameters: a vector $\boldsymbol{\theta} = (\mu, \sigma^2)'$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$; i.e., $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$.
- Prior: it is easier to obtain a conjugate prior if we factor the joint density over $\boldsymbol{\theta}$ into the product of a conditional density for $\mu$ given $\sigma^2$ and a marginal density for $\sigma^2$; i.e., $p(\boldsymbol{\theta}) = p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$.

# Conjugate analysis, normal data, mean and variance unknown

Prior: $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$ where

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \sigma^2/n_0) \\ \sigma^2 &\sim \text{Inverse-Gamma}(\nu_0/2, \nu_0\,\sigma_0^2/2) \end{aligned}$$

where

- $\mu_0 = E(\mu|\sigma^2) = E(\mu)$ is the mean of the prior density for $\mu$
- $\sigma^2/n_0$ is the variance of the prior density for $\mu$, conditional on $\sigma^2$, with $n_0$ interpretable as a "prior sample size"
- $\nu_0 > 0$ is a prior "degrees of freedom" parameter
- $\nu_0\,\sigma_0^2$ is equivalent to the sum of squares one obtains from a (previously observed) data set of size $\nu_0$

# inverse-Gamma density

## Definition

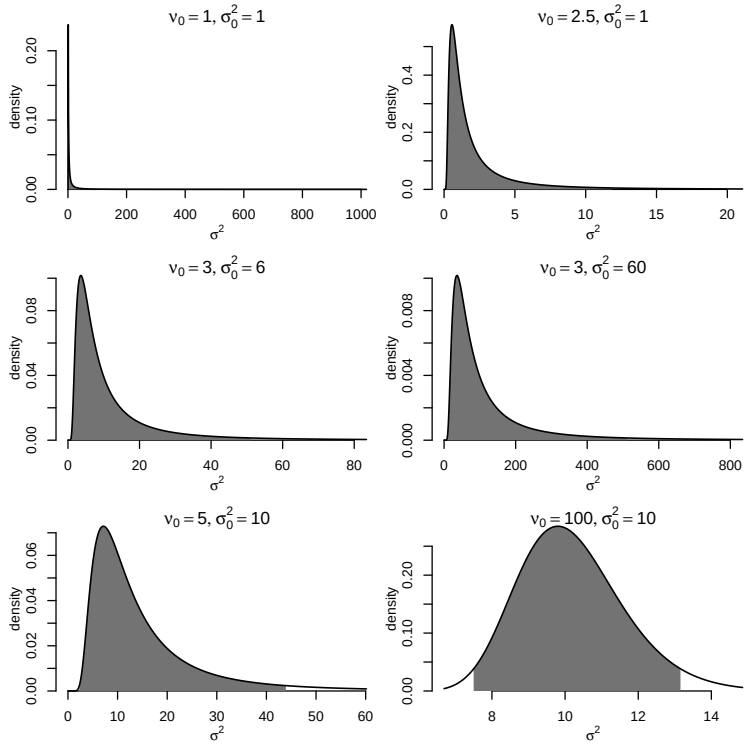
If $x > 0$ follows an inverse-Gamma density with shape parameter $a > 0$ and scale parameter $b > 0$, conventionally written as $x \sim \text{inverse-Gamma}(a, b)$, then

$$p(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(\frac{-b}{x}\right),$$

- $E(x) = \frac{b}{a-1}$ if $a > 1$.
- $V(x) = \frac{b^2}{(a-1)^2(a-2)}$ if $a > 2$.
- $p(x)$ has a mode at $b/(a+1)$.
- If $x \sim \text{inverse-Gamma}(a, b)$ then $1/x \sim \text{Gamma}(a, b)$.
- Typical use is $\sigma^2 \sim \text{inverse-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$. An improper density $p(\sigma^2) \propto 1/\sigma^2$ results with $\nu_0 = 0$.

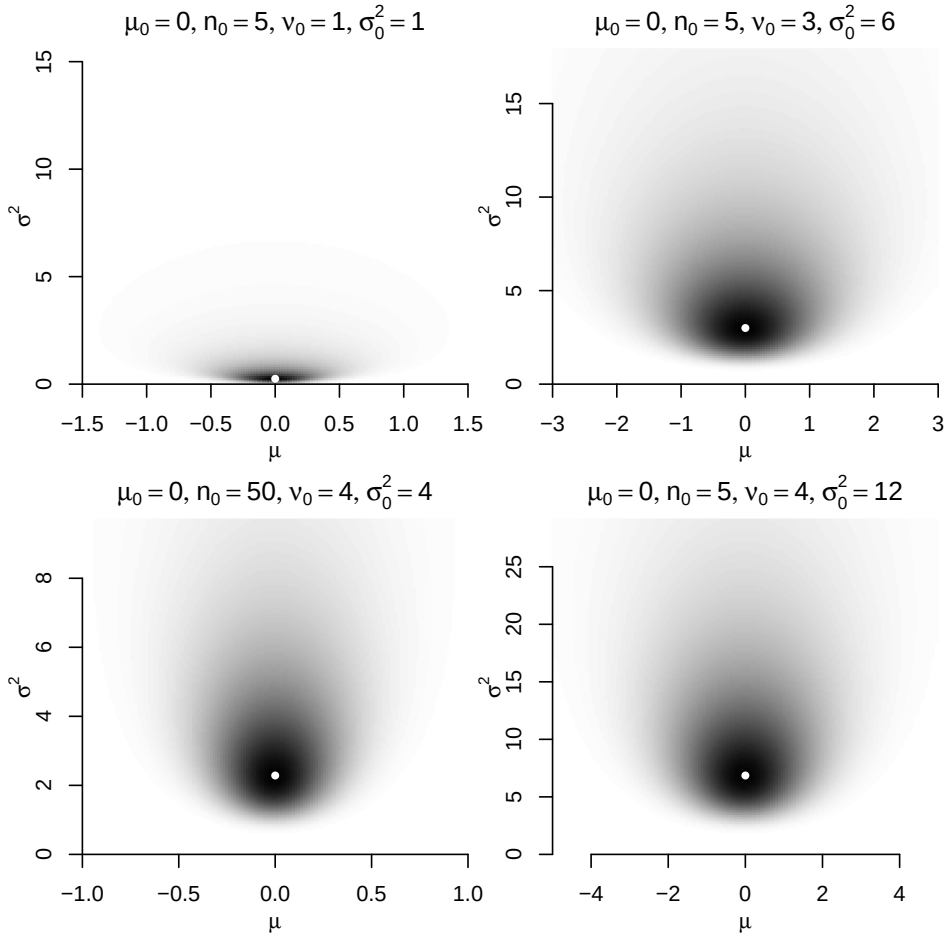
# $\sigma^2 \sim$ inverse-Gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$

- The mean, $E(\sigma^2)$, is $\nu_0\sigma_0^2/(\nu_0 - 2)$, provided $\nu_0 > 2$, otherwise the mean is undefined, and the mode occurs at $\nu_0\sigma_0^2/(\nu_0 + 2)$.
- The mean and the mode tend to coincide as $\nu_0 \to \infty$; i.e., the inverse-Gamma density tends to a (symmetric) normal density as $\nu_0 \to \infty$, but otherwise is skewed right.



# normal/inverse-Gamma density

A density for $\boldsymbol{\theta} = (\mu, \sigma^2)' \in \boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$, indexed by four parameters, $\mu_0, n_0, \nu_0, \sigma_0^2$.



# Conjugacy of the normal/inverse-Gamma prior

## Theorem (Proposition 2.5, BASS)

*Let* $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$. *If*
$\boldsymbol{\theta} = (\mu, \sigma^2)' \sim$ *normal/inverse-Gamma*$(\mu_0, n_0, \nu_0, \sigma_0^2)$, *then*
$\boldsymbol{\theta}|\mathbf{y} \sim$ *normal/inverse-Gamma*$(\mu_1, n_1, \nu_1, \sigma_1^2)$, *where*

$$\begin{aligned} \mu_1 &= \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \\ n_1 &= n_0 + n, \quad \nu_1 = \nu_0 + n \\ \nu_1\sigma_1^2 &= \nu_0\,\sigma_0^2 + S + \frac{n_0\, n}{n_0 + n}(\mu_0 - \bar{y})^2 \end{aligned}$$

*and where* $S = \sum_{i=1}^{n}(y_i - \bar{y})^2$. *That is,*

$$\begin{aligned} \mu|\,\sigma^2, \mathbf{y} &\sim N(\mu_1, \sigma^2/n_1) \\ \sigma^2|\,\mathbf{y} &\sim \textit{inverse-Gamma}\left(\frac{\nu_1}{2}, \frac{\nu_1\,\sigma_1^2}{2}\right) \end{aligned}$$

# Marginal Posterior Density, Normal Mean, Conjugacy

- The conditional posterior density for $\mu$, $p(\mu|\mathbf{y}, \sigma^2)$ is a normal density in which $\sigma^2$ appears in the expression for the variance of the conditional posterior density.
- But if we're interested in $\mu$, we want its marginal posterior density $p(\mu|\mathbf{y})$
- We "integrate out" or "average over" uncertainty with respect to $\sigma^2$; i.e.,

$$p(\mu|\mathbf{y}) = \int_0^\infty p(\mu|\sigma^2, \mathbf{y}) p(\sigma^2|\mathbf{y}) d\sigma^2$$

  where the limits of integration follow from the fact that variances are strictly positive.
- Resulting marginal density for $\mu$ is a student-$t$ density.

# Marginal Posterior Density, Normal Mean, Conjugacy

## Theorem (Proposition 2.6, BASS)

*Assume conditions of the previous theorem. Then the marginal posterior density of $\mu$ is a student-t density (Definition B.37), with location parameter $\mu_1$, scale parameter $\sqrt{\sigma_1^2/n_1}$ and $\nu_1$ degrees of freedom, where* $n_1 = n_0 + n$,

$$\mu_1 = \frac{n_0\mu_0 + n\bar{y}}{n_1},$$

$\sigma_1^2 = S_1/\nu_1$, $S_1 = \nu_0\,\sigma_0^2 + (n-1)s^2 + \frac{n_0\,n}{n_1}(\bar{y}-\mu_0)^2$, $\nu_1 = \nu_0 + n$, $s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2$ *and* $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$.

## Proof.

*Proposition C.7, BASS.* □

# Posterior Predictive Density, normal data, conjugacy

- Consider making a prediction for a future observation, $y^*$.
- This quantity also has a posterior density, called a *posterior predictive density*:

$$\begin{aligned} p(y^*|\mathbf{y}) &= \int_{\mathbb{R}^+} \int_{\mathbb{R}} p(y^*|\mu, \sigma^2, \mathbf{y}) p(\mu, \sigma^2|\mathbf{y}) d\mu d\sigma^2 \\ &= \int_{\mathbb{R}^+} \int_{\mathbb{R}} p(y^*|\mu, \sigma^2, \mathbf{y}) p(\mu|\sigma^2, \mathbf{y}) p(\sigma^2|\mathbf{y}) d\mu d\sigma^2 \end{aligned}$$

- The prediction should not simply condition on particular values of the parameters $\mu$ and $\sigma^2$; we're uncertain about the prediction because we're uncertain about the parameters $\mu$ and $\sigma^2$.
- The double integral might look a little formidable, but the posterior predictive density has a familiar form...

# Posterior Predictive Density, normal data, conjugacy

## Theorem (Proposition 2.7, BASS)

*Let* $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $(\mu, \sigma^2) \sim$ *normal/inverse-Gamma*$(\mu_0, n_0, \nu_0, \sigma_0^2)$. *Then the posterior predictive density for a future observation* $y^*$, $p(y^*|\mathbf{y})$, *is a student-t density, with location parameter*

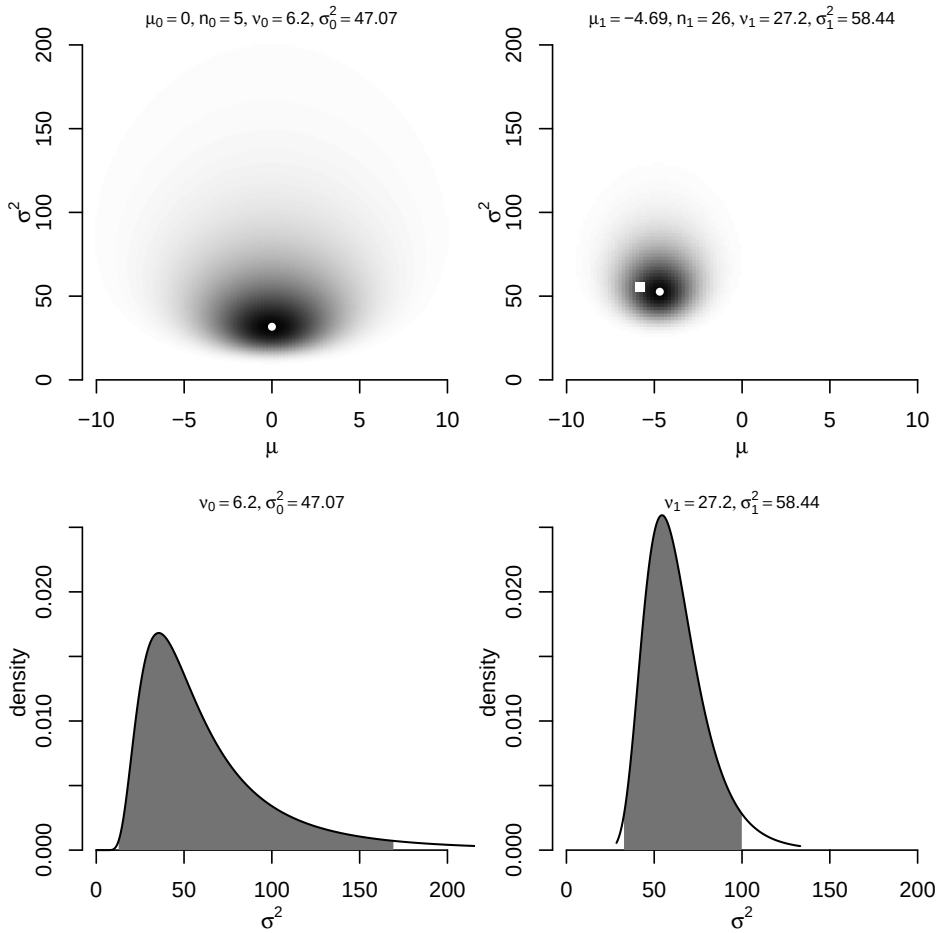
$$E(y^*|\mathbf{y}) = E(\mu|\mathbf{y}) = \mu_1 = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n},$$

*scale parameter* $\sigma_1\sqrt{(n_1+1)/n_1}$ *and* $\nu_1 = n + \nu_0$ *degrees of freedom, where* $n_1 = n_0 + n$, $\sigma_1^2 = S_1/\nu_1$ *and* $S_1 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{n_0 n}{n_0+n}(\mu_0 - \bar{y})^2$.

## Proof.

*Proposition C.8, BASS.* □

# Example 2.13, Suspected voter fraud in Pennsylvania




Prior and Posterior normal/inverse-Gamma Densities. Prior densities on the left; posterior densities on the right. Normal/inverse-Gamma densities for $(\mu, \sigma^2)$ in the upper panels, with darker colors indicating regions of higher density, the circle indicating the mode, and for the posterior density, the square indicating the location of the maximum likelihood estimates. Marginal inverse-Gamma densities for $\sigma^2$ appear in the lower panels, with the shaded area corresponding to a 95% highest density region.

## Theorem (Conjugate Prior Normal Regression Model)

$$
\begin{aligned}
y_i|\mathbf{x}_i &\overset{iid}{\sim} N(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2) \\
\boldsymbol{\beta}|\sigma^2 &\sim N(\mathbf{b}_0, \sigma^2\mathbf{B}_0) \\
\sigma^2 &\sim \textit{inverse-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)
\end{aligned}
$$

*then*

$$
\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X} &\sim N(\mathbf{b}_1, \sigma^2\mathbf{B}_1), \\
\sigma^2|\mathbf{y}, \mathbf{X} &\sim \textit{inverse-Gamma}(\nu_1/2,\ \nu_1\,\sigma_1^2/2) \\
\mathbf{b}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) \\
\mathbf{B}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1} \\
\nu_1 &= \nu_0 + n \qquad \textit{and} \\
\nu_1\sigma_1^2 &= \nu_0\sigma_0^2 + S + r.
\end{aligned}
$$

# Conjugacy Summary

| Prior | Data/Likelihood | Posterior |
|---|---|---|
| $\theta \sim \text{Beta}$ | $r \sim \text{Binomial}(\theta; n)$ | $\theta \vert r, n \sim \text{Beta}$ |
| $\mu \vert \sigma^2 \sim N$ | $y \sim N(\mu, \sigma^2)$ | $\mu \vert y, \sigma^2 \sim N$ |
| | | $\mu \vert y \sim t$ |
| $\sigma^2 \sim \text{inverse-Gamma}$ | $y \sim N(\mu, \sigma^2)$ | $\sigma^2 \vert y \sim \text{inverse-Gamma}$ |
| $\theta \sim \text{Gamma}$ | $y \sim \text{Poisson}(\theta)$ | $\theta \vert y \sim \text{Gamma}$ |
| $\mathbf{\Sigma} \sim \text{inverse-Wishart}$ | $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ | $\mathbf{\Sigma} \vert \mathbf{y} \sim \text{inverse-Wishart}$ |
| $\alpha \sim \text{Dirichlet}$ | $\mathbf{r} \sim \text{Multinomial}(\alpha; n)$ | $\alpha \vert \mathbf{r}, n \sim \text{Dirichlet}$ |

# Limitations of Conjugacy

- seemingly small set of problems amenable to conjugate Bayesian analysis
- e.g., no logit/probit regression!
- prior to MCMC revolution, Bayesian ideas interesting, perhaps even "right", but extremely difficult (impossible) to implement outside small set of problems
- MCMC changed this, circa 1990. Avalanche of Bayesian applications in statistics, econometrics. Now standard.
- Bayes Theorem to 1990: a long time!
- still make use of conjugacy (or conditional conjugacy) in implementing a MCMC algorithm known as the Gibbs sampler.