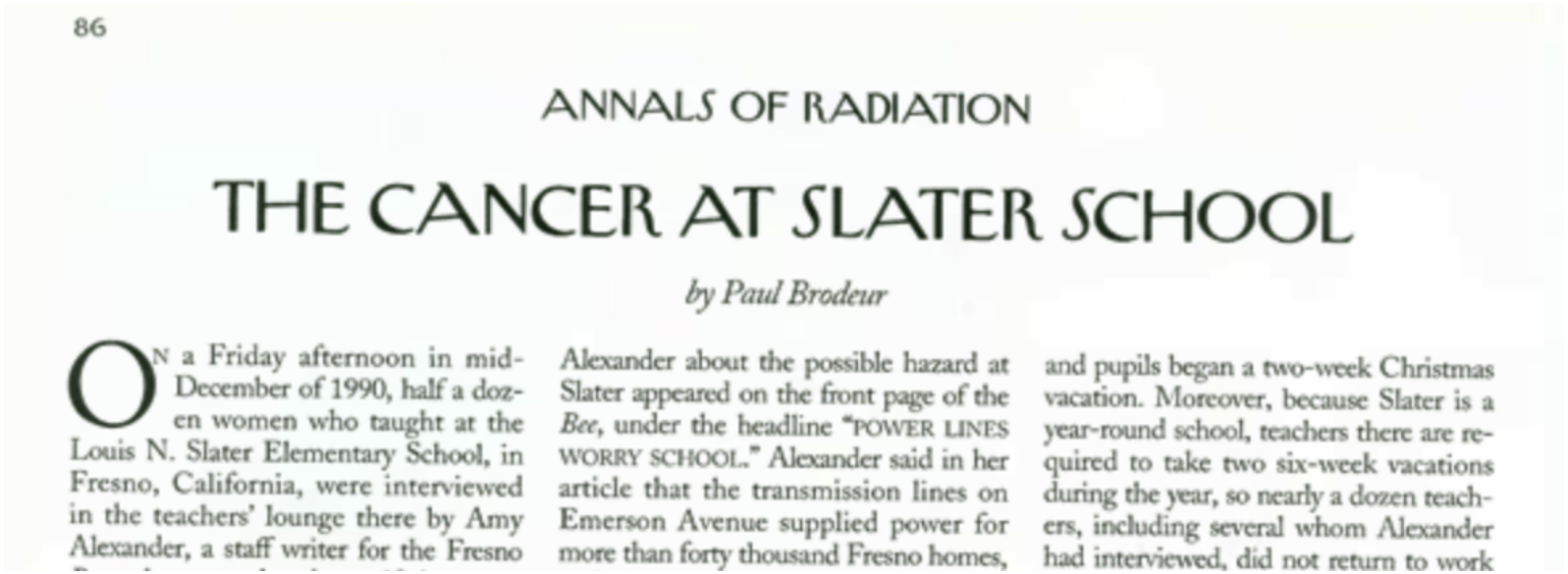
Biostatistics 682: Applied Bayesian Inference
Lecture 2: Single parameter models

Jian Kang

Department of Biostatistics
University of Michigan, Ann Arbor

"What is Bayesian statistics
and why everything else is wrong"
by Michael Lavine

The New Yorker of Dec 7, 1992



86

ANNALS OF RADIATION

THE CANCER AT SLATER SCHOOL

by Paul Brodeur

On a Friday afternoon in mid-December of 1990, half a dozen women who taught at the Louis N. Slater Elementary School, in Fresno, California, were interviewed in the teachers' lounge there by Amy Alexander, a staff writer for the Fresno

Alexander about the possible hazard at Slater appeared on the front page of the *Bee*, under the headline "POWER LINES WORRY SCHOOL." Alexander said in her article that the transmission lines on Emerson Avenue supplied power for more than forty thousand Fresno homes,

and pupils began a two-week Christmas vacation. Moreover, because Slater is a year-round school, teachers there are required to take two six-week vacations during the year, so nearly a dozen teachers, including several whom Alexander had interviewed, did not return to work

- The Slater school is an elementary school in Fresno, California
- Teachers and staff were "*concerned about the presence of two high-voltage transmission lines that ran past the school ...*"
- Their concern centered on the "*high incidence of cancer at Slater ...*"
- To address their concern, Dr. Raymond Neutra of the California Department of Health Services's Special Epidemiological Studies Program conducted a statistical analysis

The Cancer at Slater School

- *"eight cases of invasive cancer, ..., the total years of employment of the hundred and forty-five teachers, teachers' aides, and staff members, ... "*
- *"The number of person-years in terms of National Cancer Institute statistics showing the annual rate of invasive cancer in American women between the ages of forty and forty-four – the age group encompassing the average age of the teachers and staff at Slater – [which] enabled him to calculate that 4.2 cases of cancer could have been expected to occur among the Slater teachers and staff members ..."*
- Assumptions for statistical analysis:
  - The 145 employees develop (or not) cancer independently of each other
  - The chance of cancer, θ, is the same for each employee.
- Let Y be the number of cancers among the 145 employees, what is probability distribution of Y?

Binomial distribution

In general, the binomial distribution provides a natural model for data that arise from a sequence of n exchangeable trials or draws from a large population, where each trial gives rise to one of two possible outcomes, conventionally labeled as "success" and "failure", where Y is the total number of successes in the n trails. Its probability mass function is given by

$$\Pr(Y = y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, \ldots, n,$$

We write

$$Y \sim \text{Bin}(n, \theta).$$

The Cancer at Slater School (Continued)

- Since Y is the number of cancers among the 145 employees, then

$$Y \sim \text{Bin}(145, \theta).$$

- What we observed in this case? The event $\{Y = 8\}$.
- According to Dr. Neutra, the expected number of cancers is 4.2. We formulate a theory:

$$\text{Theory A:} \quad \theta = 0.03.$$

  This implies that the underlying cancer rate at slater is just like the national average.

- Other theories:

$$\text{Theory B:} \quad \theta = 0.04;$$

$$\text{Theory C:} \quad \theta = 0.05;$$

$$\text{Theory D:} \quad \theta = 0.06.$$

The Likelihood

- To compare the theories we see how well each one explains the data. That is, for each value of θ, we calculate

$$\Pr(Y = 8 \mid \theta) = \binom{145}{8} \theta^8 (1-\theta)^{137}.$$

which says how well each value of θ explains the observed data $Y = 8$. The results are

$$\Pr(Y = 8 \mid \theta = 0.03) \approx 0.036,$$
$$\Pr(Y = 8 \mid \theta = 0.04) \approx 0.096,$$
$$\Pr(Y = 8 \mid \theta = 0.05) \approx 0.134,$$
$$\Pr(Y = 8 \mid \theta = 0.06) \approx 0.136.$$

or roughly in the ratio of 1:3:4:4.

- Thus we can make statements

"Theory B explains the data about three times as well as Theory A"

"Theory C and D explains the data equally well"

The Likelihood Principal

- $\Pr(Y = y \mid \theta)$ is a function of two variables y and θ, Once $Y = 8$ has been observed, then $\Pr(Y = 8 \mid \theta)$ describes how well each theory, or value of θ, explains the data.
- It is a function only of θ; no value of Y other than 8 is relevant.
- Is $\Pr(Y = 9 \mid \theta = 0.03)$ relevant? Does it describe how well theory explains the observed data.
- **Likelihood Principal**: once Y has been observed, say $Y = y_o$, then no other value of Y matters and we should treat as $\Pr(Y = y_o \mid \theta)$ as the function only of θ.
- This principle is central to Bayesian thinking.

Prior Specifications

What prior knowledge do we have?

- In fact, there are other sources of information about whether cancer can be induced by proximity to high-voltage transmission lines.
  - Some epidemiological studies showing a positive correlation between cancer rates and proximity and others failing to show such correlations.
  - Some statements from physicists and biologists argued that the energy in magnetic fields associated with high-voltage transmission lines (purported to be the cause of increased cancer rates) is too small to have an appreciable biological effects

- The above information is inconclusive. We can assume
  - No other theory $\Rightarrow \Pr(A) + \Pr(B) + \Pr(C) + \Pr(D) = 1$.
  - Theory A is just as likely to be true as false $\Rightarrow \Pr(A) = 1/2$.
  - No information to suggest any of theory B, C or D is more likely than others.

$$\Rightarrow \Pr(B) = \Pr(C) = \Pr(D) = 1/6.$$

  - These probabilities are called prior distribution.

Prior to Posterior

- Applying the Bayes's Theorem, we can compute the posterior distribution

$$\begin{aligned}
\Pr(A \mid Y = 8) &= \frac{\Pr(A \text{ and } Y = 8)}{\Pr(Y = 8)} \\
&= \frac{\Pr(A \text{ and } Y = 8)}{\Pr(A \text{ and } Y = 8) + \Pr(B \text{ and } Y = 8) + \Pr(C \text{ and } Y = 8) + \Pr(D \text{ and } Y = 8)} \\
&= \frac{\Pr(A \text{ and } Y = 8)}{\Pr(A)\Pr(Y = 8 \mid A) + \Pr(B)\Pr(Y = 8 \mid B) + \Pr(C)\Pr(Y = 8 \mid C) + \Pr(D)\Pr(Y = 8 \mid D)} \\
&\approx \frac{(1/2)(.036)}{(1/2)(.036) + (1/6)(.096) + (1/6)(.134) + (1/6)(.136)} \\
&\approx 0.23
\end{aligned}$$

$$\Pr(B \mid Y = 8) = 0.21 \quad \Pr(C \mid Y = 8) = \Pr(D \mid Y = 8) = 0.28$$

- What statement we can make?
  - Which theory is more likely? The four theories are about equally likely
  - What are the odds that the underlying cancer rate at Slater is higher than 0.03. The odds are about 3 to 1

A frequentist approach

- Consider the hypothesis testing problem

$$H_0 : \theta = 0.03, \quad \text{versus} \quad H_1 : \theta > 0.03.$$

- What is the p-value?
  - The probability under H_0 of observing an outcome at least as extreme as the outcome actually observed.
  - In the Slater problem,

$$\text{p-value} = \Pr(Y = 8 \mid \theta = 0.03) + \ldots + \Pr(Y = 145 \mid \theta = 0.03) \approx 0.07$$

- Why the p-value is not appropriate here?
  - Hypotheses should be compared by how well they explain the data,
  - the p-value does not account for how well the alternative hypotheses explain the data, and
  - the summands of $\Pr(Y = 9 \mid \theta = 0.03), \ldots, \Pr(Y = 145 \mid \theta = 0.03)$ are irrelevant because they do not describe how well any hypothesis explains any observed data
- The p-value does not obey the Likelihood Principle. Why?

Discussion

- Whether the two approaches agree?
- The classical p-value is approximately 0.07, or very close to the widely accepted critical value of 0.05 (why 0.05?), below which null hypothesis are rejected. It is close to rejecting H_0.
- The Bayesian analysis say that the evidence against $\theta = 0.03$ is not very strong, is only about 3 to 1.
- Bayesian analysis used four discrete values of θ to construct the prior model. A better approach is to treat θ as continuous with values between 0 and 1.
- We might expect $\Pr\{\theta \in (0.03, 0.06)\} = 0.99$ or let $\theta \sim \text{Unif}(0, 1)$ (all values of θ are equally likely).

Beta Distribution

- If $\theta \sim \text{Beta}(\alpha, \beta)$, then

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

$$E(\theta) = \frac{\alpha}{\alpha+\beta}, \qquad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

  where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function.

- Continuous prior distribution for θ in the binomial model $Y \sim \text{Bin}(\theta, n)$.

- We can show that the posterior distribution of θ given $Y = y$ is still beta distribution

$$\theta \mid Y = y \sim \text{Beta}(y+\alpha, n-y+\beta).$$

-

$$E(\theta \mid y) = (y+\alpha)/(n+\alpha+\beta)$$

$$\text{Var}(\theta \mid y) = \frac{(y+\alpha)(n-y+\beta)}{(n+\alpha+\beta)^2(n+1+\alpha+\beta)}$$

Predictive Distribution

- The beta function:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

- $$\binom{n}{m} = \frac{\Gamma(n+1)}{\Gamma(m+1)\Gamma(n-m+1)}$$

- Prior predictive distribution

$$\Pr(\tilde{Y} = \tilde{y}) = \frac{\Gamma(n+1)}{\Gamma(\tilde{y}+1)\Gamma(n-\tilde{y}+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\tilde{y}+\alpha)\Gamma(n-\tilde{y}+\beta)}{\Gamma(\alpha+\beta+n)}$$

- Posterior predictive distribution: $\tilde{y} \sim \text{Bin}(\theta, \tilde{n})$ and $y \sim \text{Bin}(\theta, n)$.

$$\Pr(\tilde{Y} = \tilde{y} \mid Y = y) = \frac{\Gamma(\tilde{n}+1)}{\Gamma(\tilde{y}+1)\Gamma(\tilde{n}-\tilde{y}+1)} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)} \frac{\Gamma(\alpha+y+\tilde{y})\Gamma(\beta+n-y+\tilde{n}-\tilde{y})}{\Gamma(\alpha+\beta+n+\tilde{n})}$$

Posterior a compromise between prior and data

- The posterior mean is a compromise between prior information and data
- The posterior mean is a weighted average of the sample mean and the prior mean

$$E(\theta \mid y) = \frac{\alpha + y}{\alpha + \beta + n} = \frac{n}{\alpha + \beta + n}\left(\frac{y}{n}\right) + \frac{\alpha + \beta}{\alpha + \beta + n}\left(\frac{\alpha}{\alpha + \beta}\right).$$

- The prior hyperparameter interpretation
  - α : Prior number of successes
  - $\alpha + \beta$: Prior sample size
- When sample size goes to infinity, what happened?

$$\lim_{n \to \infty} \left\{ E(\theta \mid y) - \frac{y}{n} \right\} = 0$$

Improper prior distribution

- A proper distribution whose density integrates to 1.

$$\sum_{y=0}^{n} \pi(Y = y \mid \theta) = 1, \quad \int \pi(\theta \mid \alpha, \beta) d\theta = 1.$$

- An improper distribution whose "density" does not integrate to one (or even a finite number for that matter) over the support of its argument.

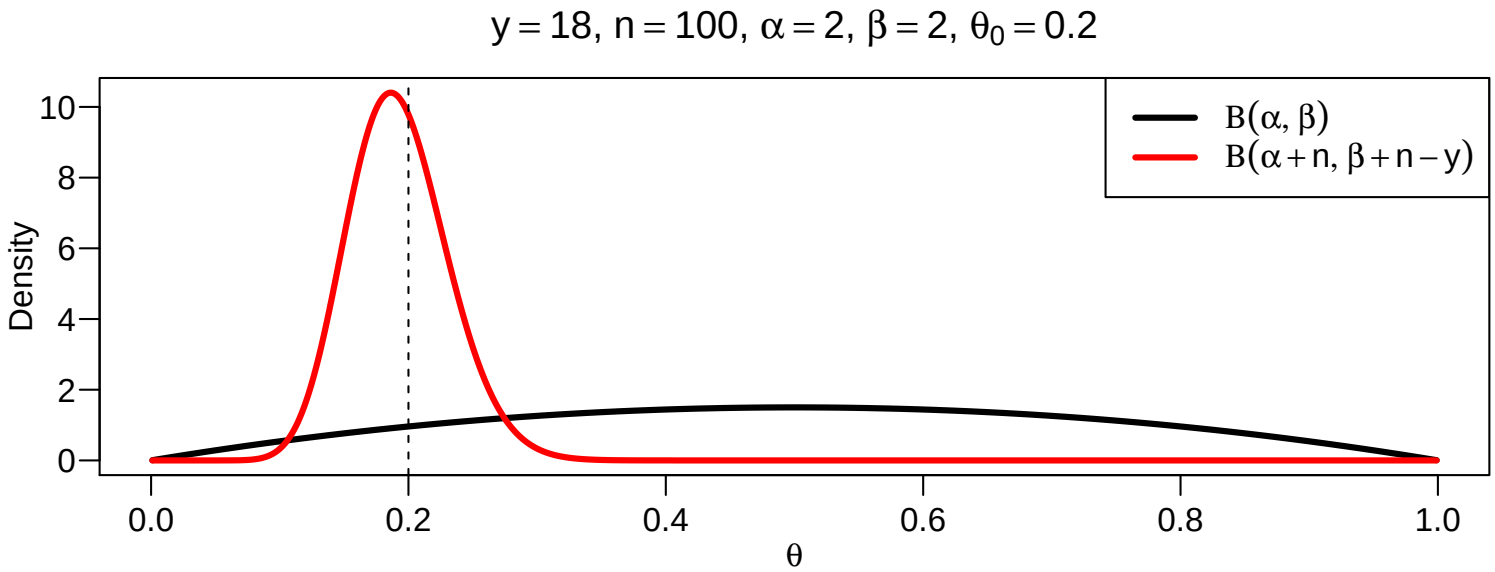
$$h(\theta) = \lim_{\alpha \to 0, \beta \to 0} \pi(\theta \mid \alpha, \beta), \qquad \int h(\theta) d\theta = \infty.$$

- When the posterior distribution is proper?

$$\int \pi(y \mid \theta) \pi(\theta) d\theta < \infty.$$

- Caution: You have to be very careful when using improper priors that the answer makes sense. Probability theory is not guarantee that because improper priors are not probability distirbutions

Prior versus Posterior



- Prior expectation is the average of the posterior expectation by *the law of iteration expectation:*

$$E(\theta) = E\{E(\theta \mid y)\}.$$

- Posterior variance is (on average) smaller than the prior variance by *the law of total variation:*

$$\mathrm{Var}(\theta) = E\{\mathrm{Var}(\theta \mid y)\} + \mathrm{Var}\{E(\theta \mid y)\}.$$

Example: Placenta Previa

- *A study conducted in Germany of 980 births from women with placenta previa. Out of the 980 births, $y = 437$ were baby girls. The established proportion of female births in the genral population is 0.485. The scientific question of interest is whether the proportion of female births in this subpopulation is less than that in the general population*
- Let θ denote the proportion of female births. We can assign the uniform prior for θ.

$$\theta \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1).$$

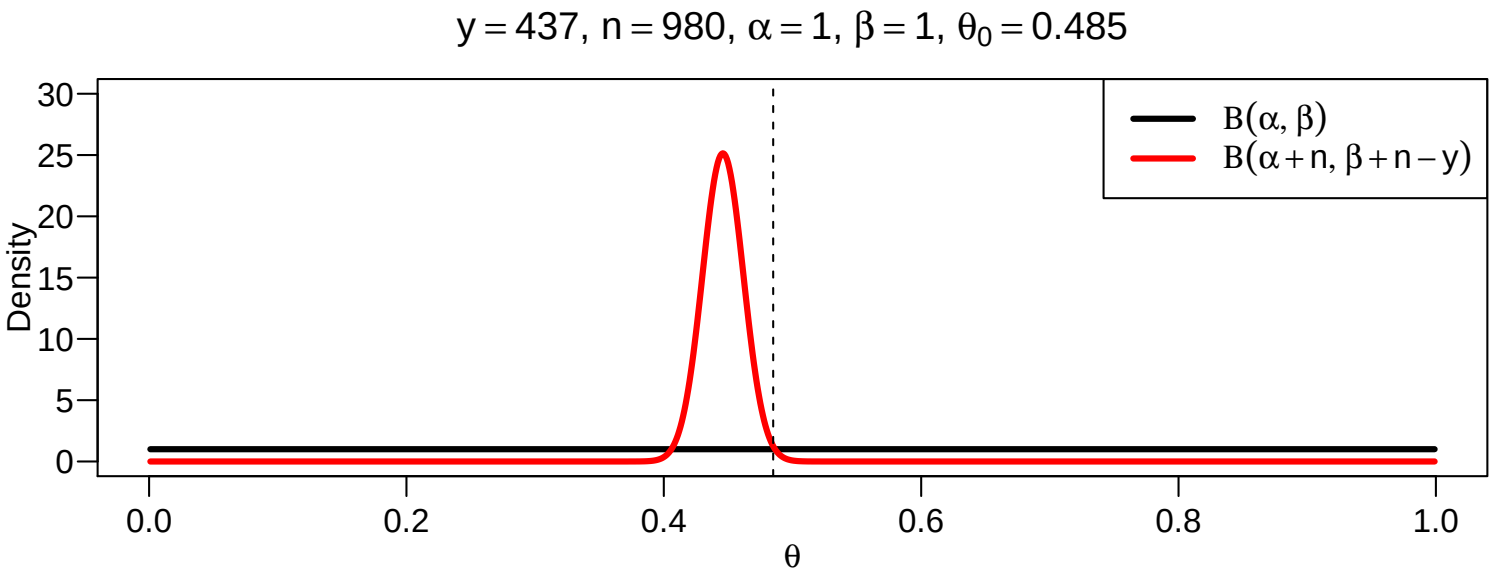
- The posterior distribution of θ is given by

$$\theta \mid y \sim \text{Beta}(438, 544).$$

This implies

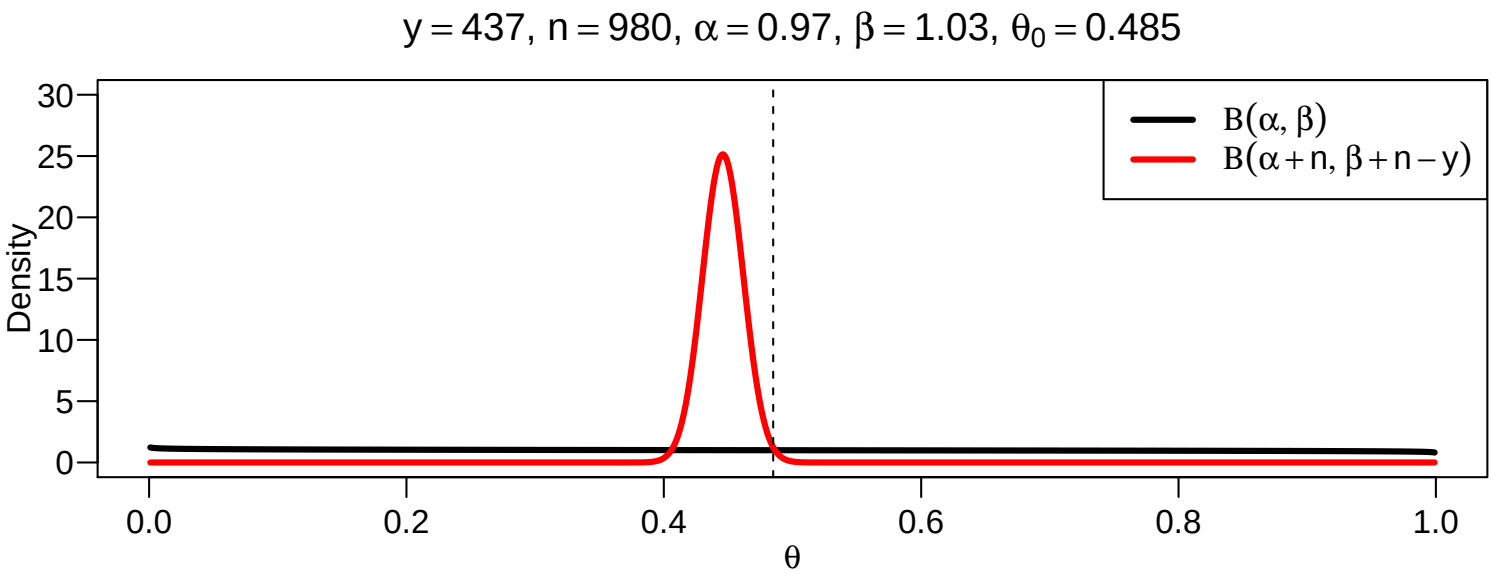
  - $E(\theta \mid y) = 0.446$ and $\text{Sd}(\theta \mid y) = 0.016$.
  - $\text{Med}(\theta \mid y) = 0.446$.
  - 95% credible interval of $(\theta \mid y) = (0.415, 0.477)$.
  - $\Pr(\theta < 0.485) = 0.993$.

Example: Placenta Previa (continued)



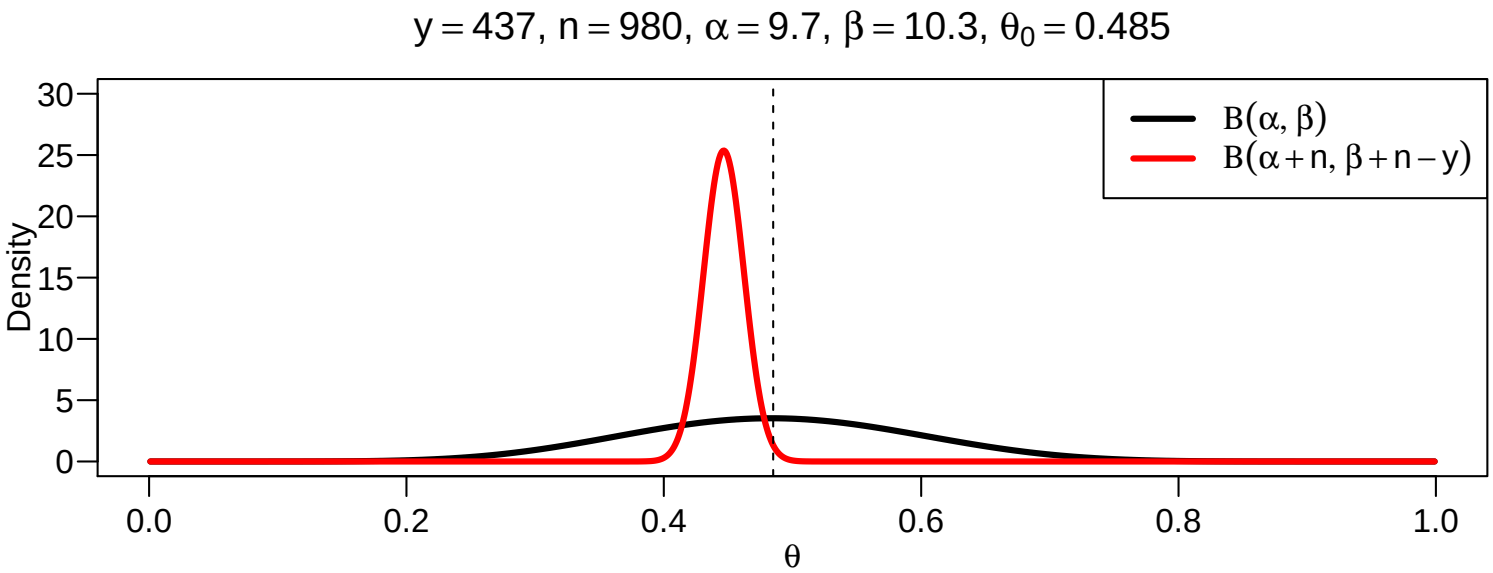
$E(\theta)$	$\alpha + \beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

Example: Placenta Previa (continued)



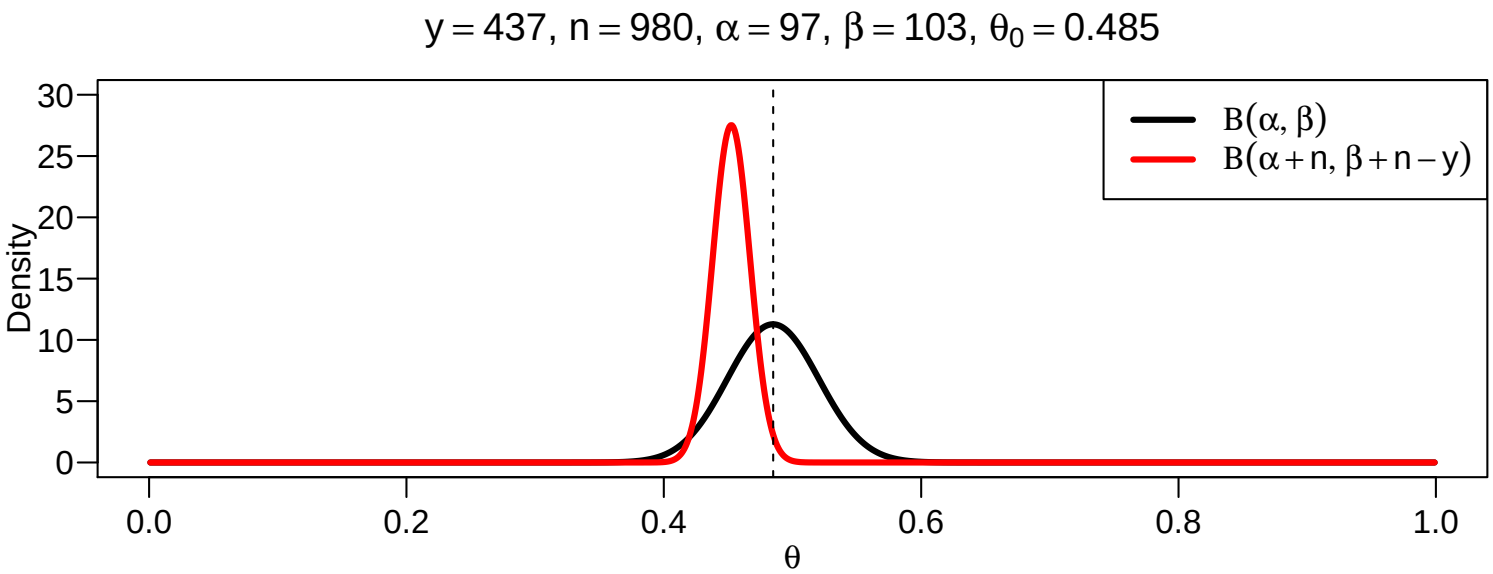
$E(\theta)$	$\alpha+\beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

Example: Placenta Previa (continued)



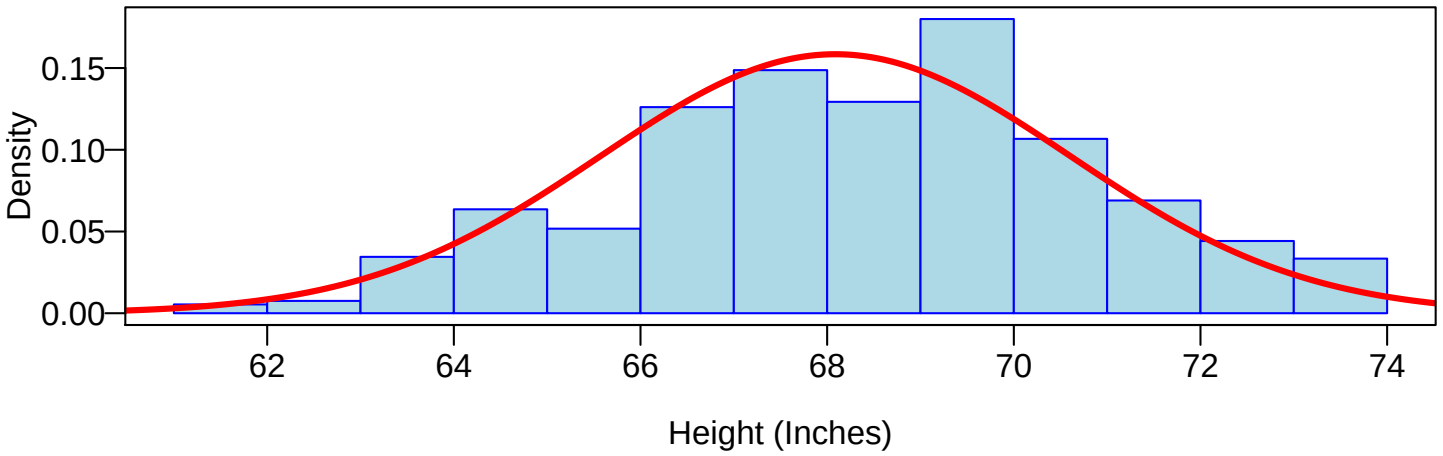
$E(\theta)$	$\alpha+\beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

Example: Placenta Previa (continued)



$E(\theta)$	$\alpha+\beta$	$E(\theta \mid y)$	95% Credible Interval
0.5	2	0.446	(0.415, 0.477)
0.485	2	0.446	(0.415, 0.477)
0.485	20	0.447	(0.416, 0.478)
0.485	200	0.453	(0.424, 0.481)

Children's Height



- *In 1886, Francis Galton collected data in a table, showing a cross-tabulation of 928 adult children born to 205 fathers and mothers, by their height and their mid-parent's height. Let us first focus on the children's height data. The sample mean is 68.09 inches and the sample standard deviation is 2.62 inches*
- What is the sampling distribution of data?
- Why we can assume that?

Normal model

- The normal distribution is fundamental to most statistical modeling.
- The central limit theorem helps to justify using the normal likelihood in many statistical problems, as an approximation to a less analytically convenient actual likelihood.
- Even when the normal distribution does not itself provide a good model fit, it can be useful as a component of a more complicated model involving t or finite mixture distributions. We will see it in the future lectures.

Estimating a normal mean with known variance

- $y_i \mid \mu \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2), i = 1, \ldots, n$ with σ^2 known. Let $y = (y_1, \ldots, y_n)$.
- What is the likelihood?

$$\pi(y \mid \mu) = (2\pi\sigma^2)^{n/2} \exp\left\{ -(2\sigma^2)^{-1} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

- What prior models should we consider?

Conjugate prior distributions

If $\mathcal{F}$ is a class of sampling distributions $\pi(y \mid \theta)$, and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ is

$$\pi(\theta \mid y) \in \mathcal{P}, \text{ for all } \pi(\cdot \mid \theta) \in \mathcal{F} \text{ and } \pi(\cdot) \in \mathcal{P}.$$

- Is beta distribution conjugate for binomial distribution?
- What distributions are conjugate for normal distribution?

The natural conjugate prior families: $\mathcal{P}$ is the set of all densities having the same functional form as the likelihood.

Natural conjugate prior for normal model

- We can write

$$\pi(y \mid u) \propto \exp\{a(y)\mu^2 + b(y)\mu + c(y)\}$$

$$\begin{aligned} a(y) &= -n(2\sigma^2)^{-1} \\ b(y) &= \sigma^{-2}\sum_{i=1}^{n} y_i = n\sigma^{-2}\bar{y} \\ c(y) &= -(2\sigma^2)^{-1}\sum_{i=1}^{n} y_i^2 \end{aligned}$$

- To specify the natural conjugate prior, we set

$$\pi(\mu) \propto \exp\left\{\tilde{a}\mu^2 + \tilde{b}\mu + \tilde{c}\right\} = \exp\left\{-\frac{1}{2\tau^2}(\mu - \zeta)^2\right\} \Rightarrow \mu \sim N(\zeta, \tau^2).$$

- This further implies that

$$\pi(\mu \mid y) \sim \mathrm{N}(\tilde{\zeta}, \tilde{\tau}^2)$$

$$\begin{aligned} \tilde{\zeta} = E(\mu \mid y) &= (n\sigma^{-2} + \tau^{-2})^{-1}(n\sigma^{-2}\bar{y} + \tau^{-2}\zeta) \\ \tilde{\tau}^2 = \mathrm{Var}(\mu \mid y) &= (n\sigma^{-2} + \tau^{-2})^{-1} \end{aligned}$$

Limiting case

- Posterior mean is a weighted average of the prior mean and the sample mean

$$\tilde{\zeta} = \frac{n\sigma^{-2}}{n\sigma^{-2} + \tau^{-2}}\bar{y} + \frac{\tau^{-2}}{n\sigma^{-2} + \tau^{-2}}\zeta.$$

- Posterior mean is prior mean adjusted toward the sample mean

$$\tilde{\zeta} \quad = \zeta + \frac{n\sigma^{-2}}{n\sigma^{-2} + \tau^{-2}}(\bar{y} - \zeta)$$

- Posterior mean is sample mean that are shrunk toward the prior mean

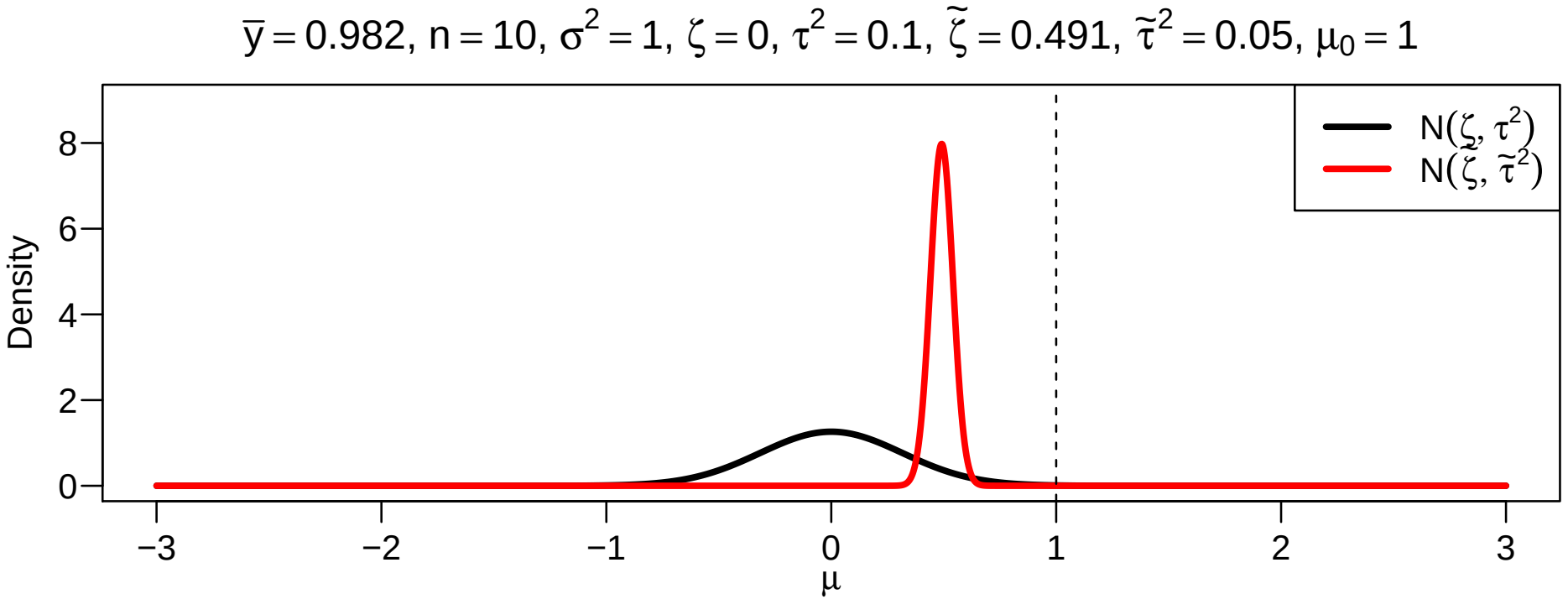
$$\tilde{\zeta} \quad = \bar{y} - \frac{\tau^{-2}}{n\sigma^{-2} + \tau^{-2}}(\bar{y} - \zeta)$$

- A non-informative prior when $\tau^2 \to \infty$. We write $\pi(\mu) \propto 1$.

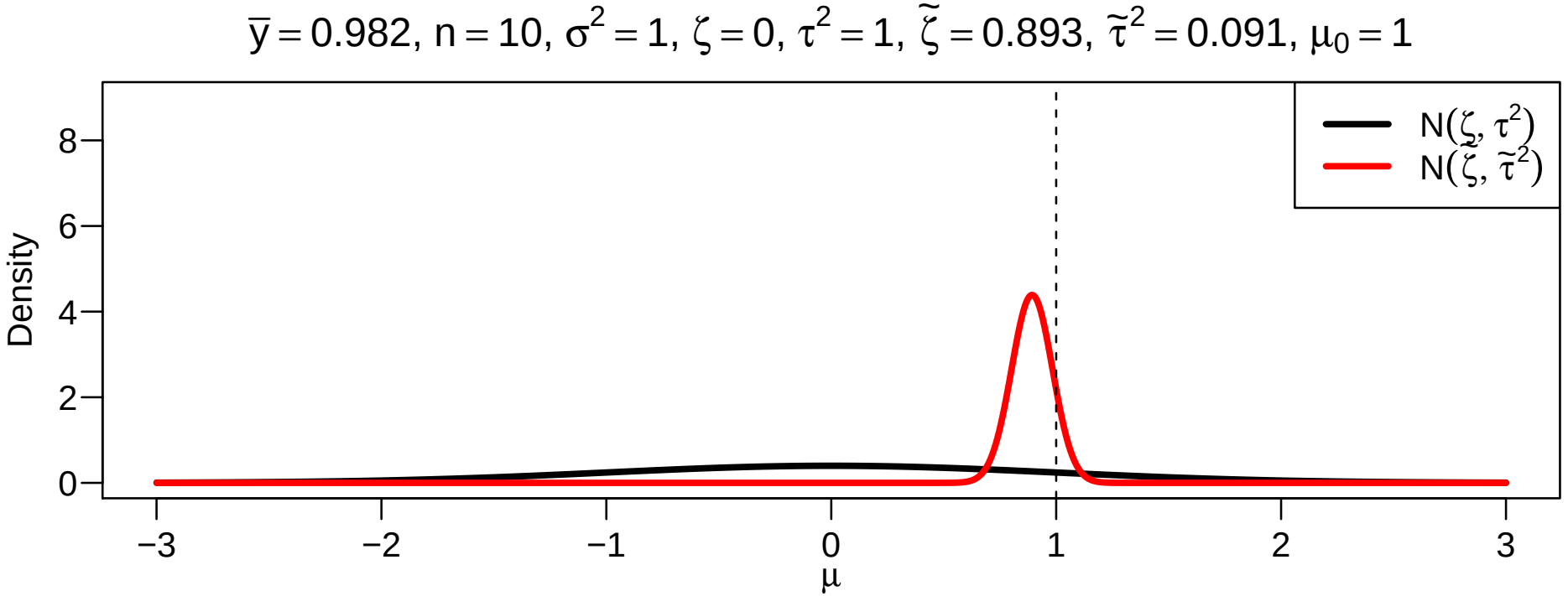
$$\lim_{\tau^2 \to \infty} \tilde{\zeta} = \bar{y}, \qquad \lim_{\tau^2 \to \infty} \tilde{\tau}^2 = \frac{\sigma^2}{n}$$

It is the MLE!

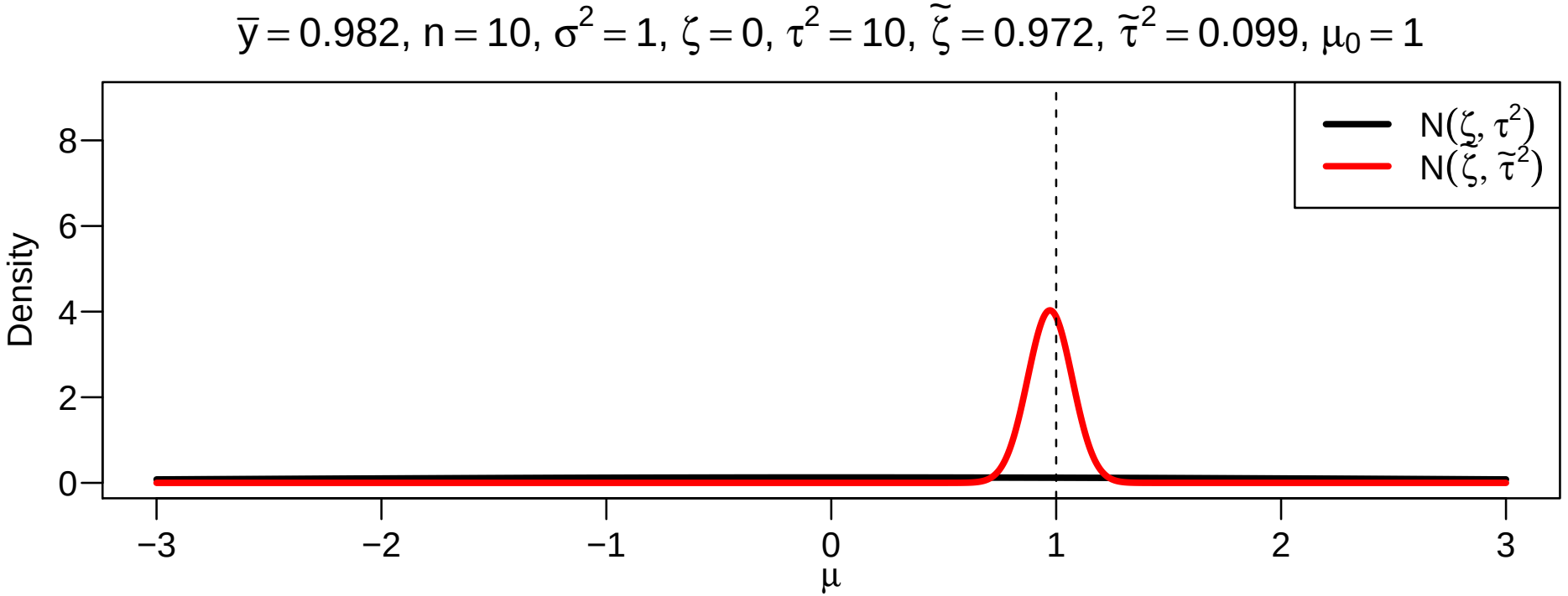
Limiting case: $\tau^2 \to \infty$



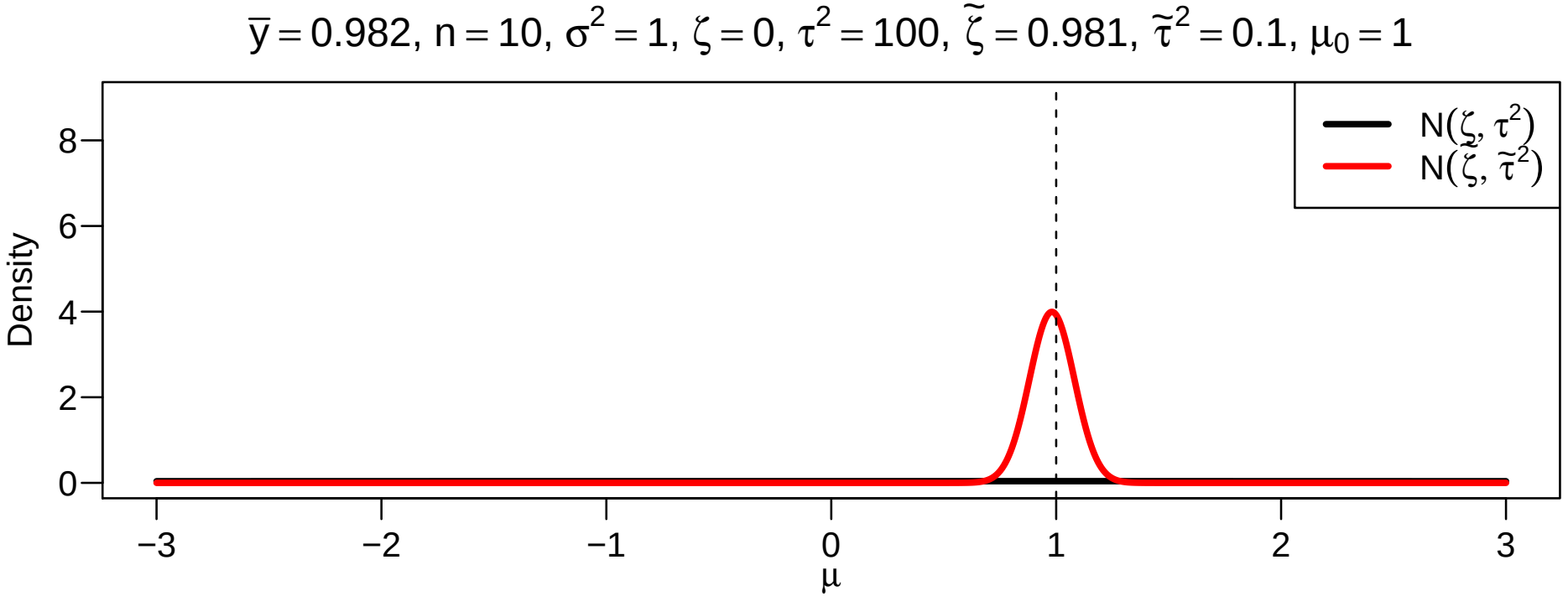
Limiting case: $\tau^2 \to \infty$



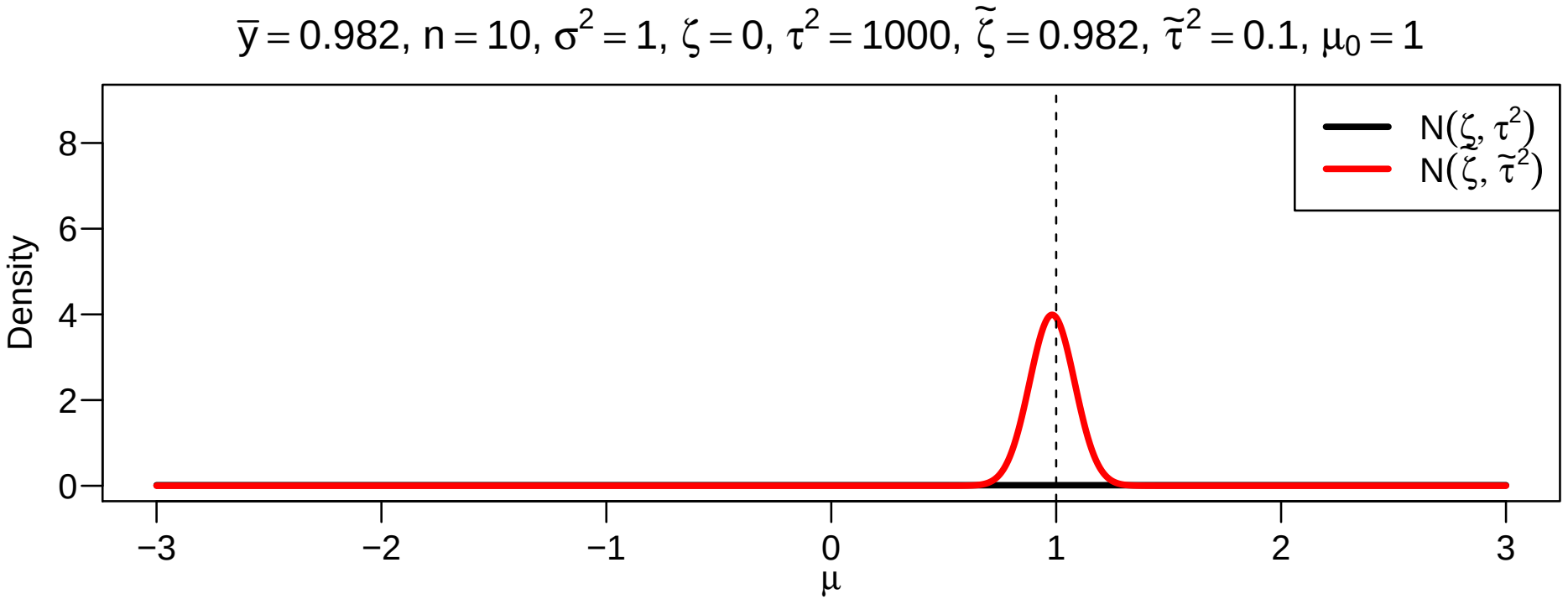
Limiting case: $\tau^2 \to \infty$



Limiting case: $\tau^2 \to \infty$



Limiting case: $\tau^2 \to \infty$



Predictions

- Prior predictive distribution:

$$\pi(\tilde{y}) = \int_{-\infty}^{\infty} \pi(\tilde{y} \mid \mu)\pi(\mu)d\mu = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left\{-\frac{(\tilde{y}-\zeta)^2}{2(\sigma^2+\tau^2)}\right\}$$

- Posterior predictive distribution:

$$\pi(\tilde{y} \mid y) = \int_{-\infty}^{\infty} \pi(\tilde{y} \mid \mu)\pi(\mu \mid y)dy = \frac{1}{\sqrt{2\pi(\sigma^2+\tilde{\tau}^2)}} \exp\left\{-\frac{(\tilde{y}-\tilde{\zeta})^2}{2(\sigma^2+\tilde{\tau}^2)}\right.$$

  where

- Do we have a simpler way to figure out those distributions?
  - The density of $(\tilde{y}, \mu)^{\mathrm{T}}$ must be bivariate normal and so the marginal density of $\tilde{y} \mid y$ must be normal.
  - The normal distribution is completely specified by its mean and variance.

Estimating a normal variance with known mean

- Let
$$y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2),$$
where μ is known and σ^2 is unknown.
- The likelihood function
$$\pi(y \mid \sigma^2) = (2\pi\sigma^2)^{n/2} \exp\left\{-(2\sigma^2)^{-1} \sum_{i=1}^{n} (y_i - \mu)^2\right\}$$
- For a natural conjugate prior, we seek a prior that has the same functional form"
$$\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right),$$
which is the kernel of an inverse gamma distribution: $G^{-1}(\alpha, \beta)$. Thus the natural conjugate prior in this case is
$$\sigma^2 \sim G^{-1}(\alpha, \beta).$$
what is the normalizing constant in $\pi(\sigma^2)$? $\frac{\beta^\alpha}{\Gamma(\alpha)}$.
- Posterior distribution
$$\pi(\sigma^2 \mid y) \propto \quad (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\nu/2 + \beta}{\sigma^2}\right),$$
where $\nu = n^{-1} \sum_i (y_i - \mu)^2$.

An improper prior

- An improper prior results in the limit as $\alpha \downarrow 0$ and $\beta \downarrow 0$. In this case,

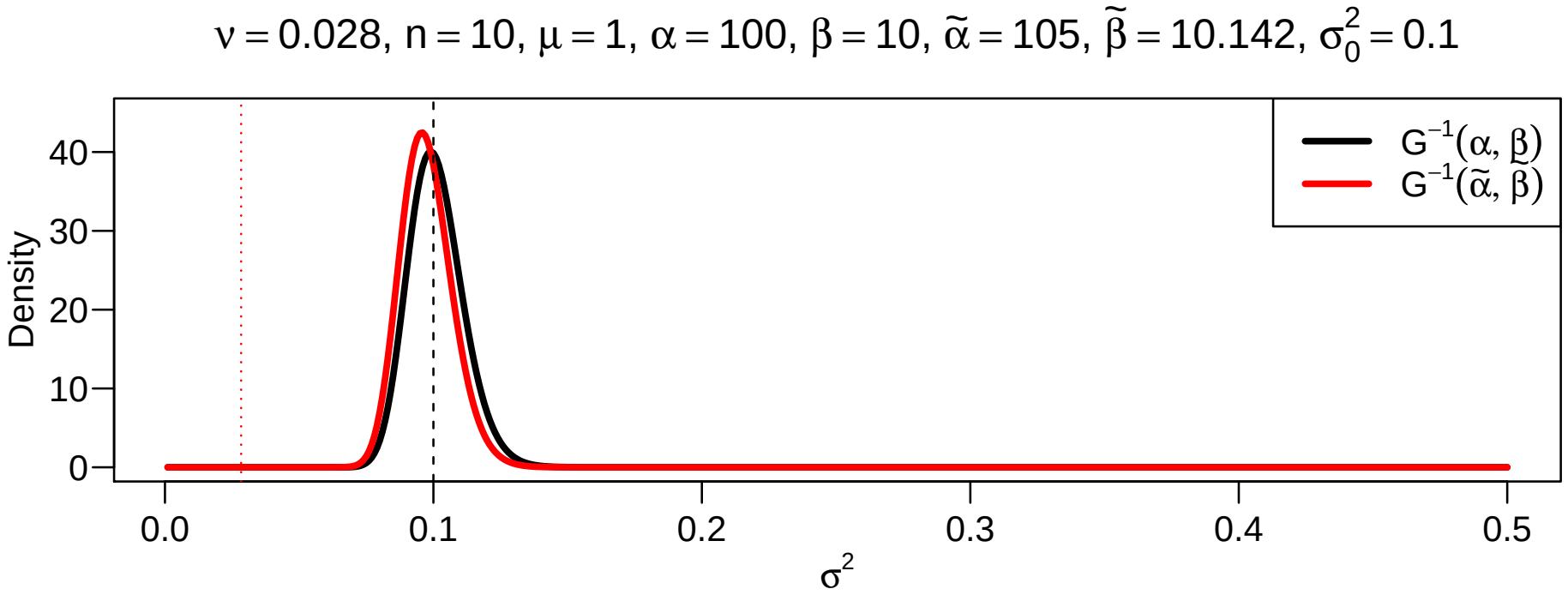
$$\lim_{\alpha \to 0, \beta \to 0} \pi(\sigma^2; \alpha, \beta) = \sigma^{-2}$$

- What is the posterior distribution? Is it proper or not?

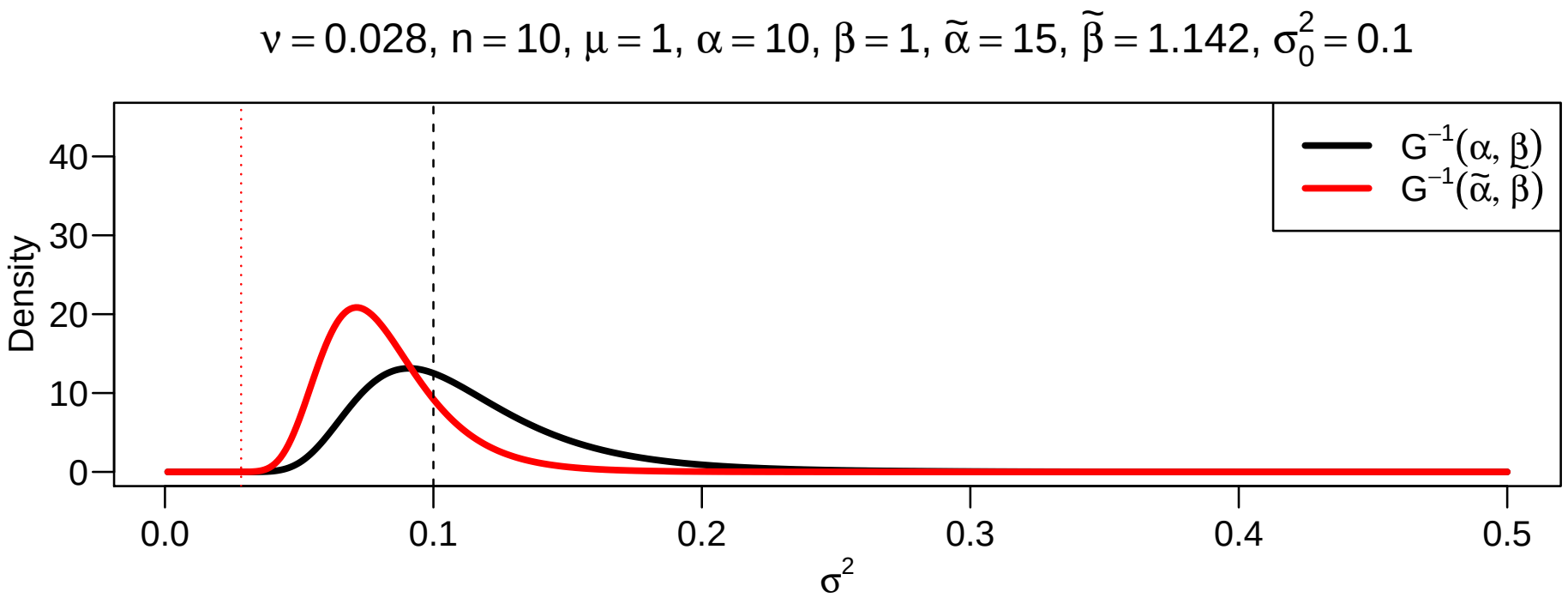
$$\sigma^2 \mid y \sim \mathrm{G}^{-1}(n/2, \nu/2)$$

It is proper.

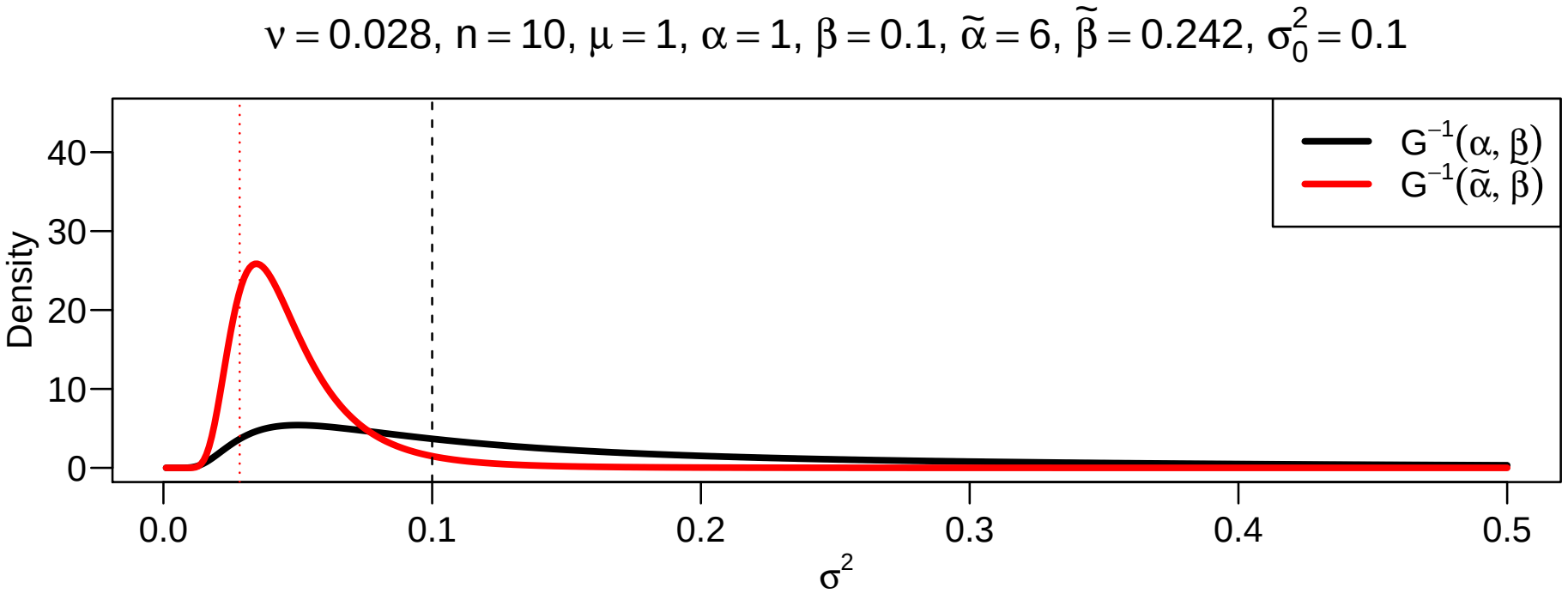
Limiting case: $\alpha \downarrow 0$ and $\beta \downarrow 0$



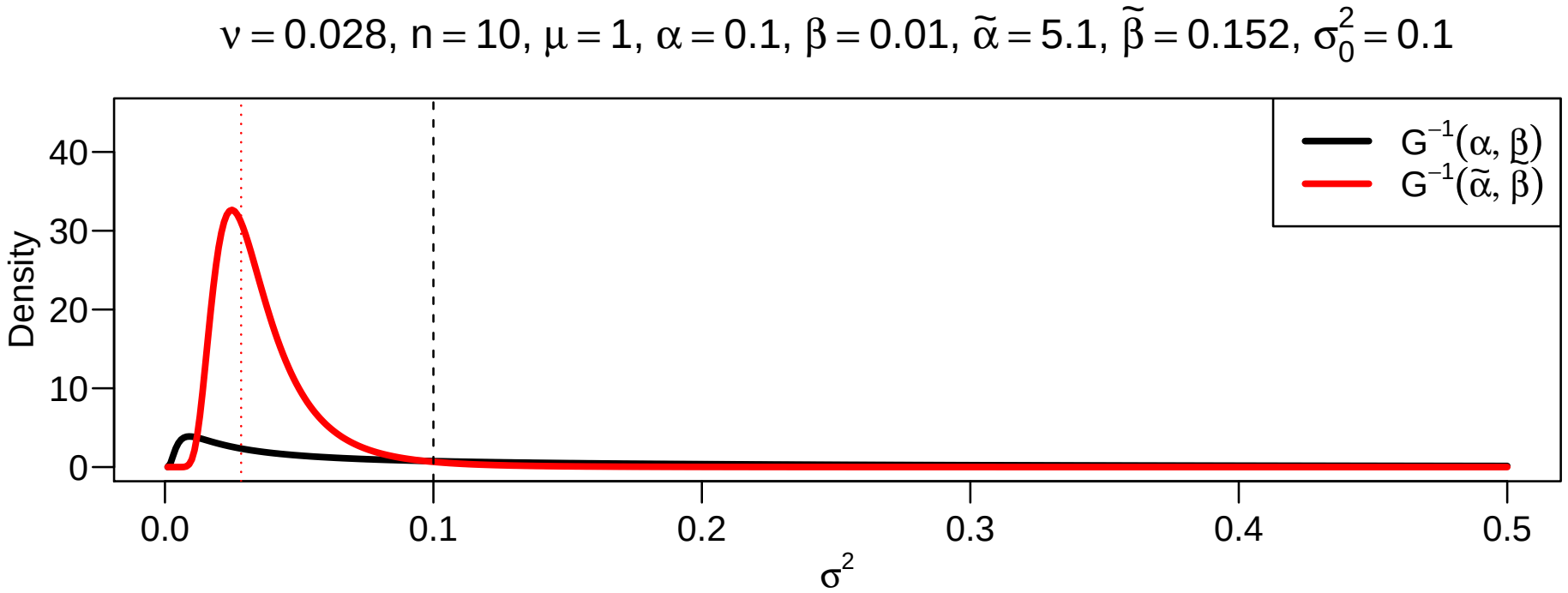
Limiting case: $\alpha \downarrow 0$ and $\beta \downarrow 0$



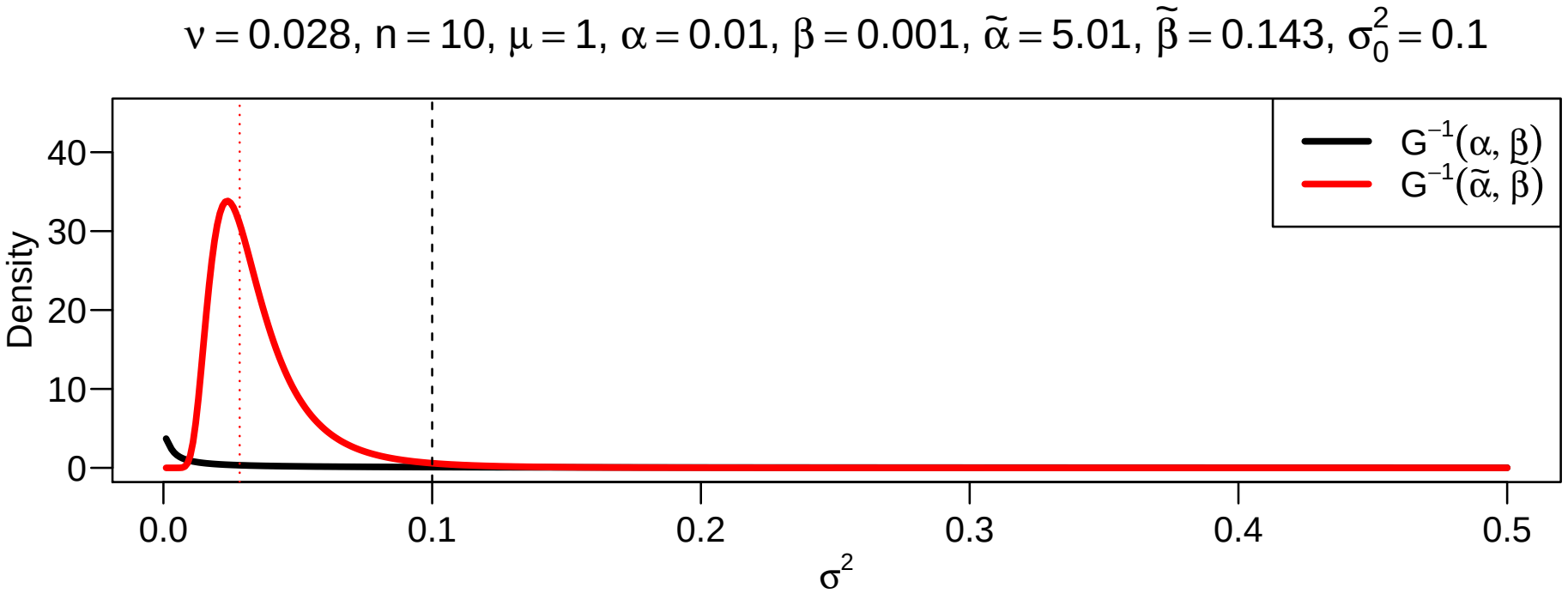
Limiting case: $\alpha \downarrow 0$ and $\beta \downarrow 0$



Limiting case: $\alpha \downarrow 0$ and $\beta \downarrow 0$



Limiting case: $\alpha \downarrow 0$ and $\beta \downarrow 0$



Precision

- Bayesian often work directly with the precision, σ^{-2}, as opposed to the variance.
- The likelihood can be written as

$$\pi(y \mid \sigma^{-2}) \propto (\sigma^{-2})^{n/2} \exp\left\{-0.5\sigma^{-2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}$$

- Using the fact that if $x \sim G^{-1}(\alpha, \beta)$, then $x^{-1} \sim G(\alpha, \beta)$, the natural conjugate prior is

$$\sigma^{-2} \sim G(\alpha, \beta).$$

Therefore,

$$\pi(\sigma^{-2} \mid y) \propto (\sigma^{-2})^{\alpha+n/2-1} \exp\left\{-(\beta + \nu/2)\sigma^{-2}\right\}$$

and

$$\sigma^{-2} \mid y \sim G(\alpha + n/2, \beta + \nu/2)$$

- An improper prior results in the limit as $\alpha \downarrow 0$ and $\beta \downarrow 0$.

$$\pi(\sigma^{-2}) = \sigma^2$$