Chapter 7: Estimation

Sections

- 7.1 Statistical Inference

Bayesian Methods:

- 7.2 Prior and Posterior Distributions
- 7.3 Conjugate Prior Distributions
- 7.4 Bayes Estimators

Frequentist Methods:

- 7.5 Maximum Likelihood Estimators
- 7.6 Properties of Maximum Likelihood Estimators
  - Skip: p. 434-441 (EM algorithm and Sampling Plans)
- 7.7 Sufficient Statistics
- Skip: 7.8 Jointly Sufficient Statistics
- Skip: 7.9 Improving an Estimator

Statistical Inference

We have seen statistical models in the form of probability distributions:

$$f(x|\theta)$$

- In this section the general notation for any parameter will be θ
- The parameter space will be denoted by Ω

For example:

- Life time of a christmas light series follows the $\mathrm{Expo}(\theta)$
- The average of 63 poured drinks is approximately normal with mean θ
- The number of people that have a disease out of a group of N people follows the $\mathrm{Binomial}(N, \theta)$ distribution.

In practice the value of the parameter θ is unknown.

Statistical Inference

Statistical Inference: Given the data we have observed what can we say about θ?

- I.e. we observe random variables $X_1, \ldots, X_n$ that we assume follow our statistical model and then we want to draw probabilistic conclusions about the parameter θ.

For example:

- If I tested 5 christmas light series from the same manufacturer and they lasted for

$$21, 103, 76, 88 \text{ and } 96 \text{ days.}$$

  Assuming that the life times are independent and follow $\mathrm{Expo}(\theta)$, what does this data set tell me about the failure rate θ?

Statistical Inference – Another example

Say I take a random sample of 100 people and test them all for a disease.

If 3 of them have the disease, what can I say about $\theta =$ the prevalence of the disease in the population?

- Say I estimate θ as $\hat{\theta} = 3/100 = 3\%$.
- How sure am I about this number?
  I want uncertainty bounds on my estimate.
- Can I be confident that the prevalence of the disease is higher than 2% ?

Statistical Inference

Examples of different types of inference

Prediction

- Predict random variables that have not yet been observed
- E.g. If we test 40 more people for the disease, how many people do we predict have the disease?

Estimation

- Estimate (predict) the unknown parameter θ
- E.g. We estimated the prevalence of the disease as $\hat{\theta} = 3\%$.

Statistical Inference

Examples of different types of inference

Making decisions

- Hypothesis testing, decision theory
- E.g. If the disease affects 2% or more of the population, the state will launch a costly public health campaign.
  Can we be confident that θ is higher than 2% ?

Experimental Design

- What and how much data should we collect?
- E.g. How do I select people in my clinical trial? How many do I need to be comfortable making decision based on my analysis?
- Often limited by time and / or budget constraints

Bayesian vs. Frequentist Inference

Should a parameter be treated as a random variable?

- Do we think about $f(\mathbf{x}|\theta)$ as the conditional pdf/pf of $\mathbf{X}$ given θ or
- do we think about $f(\mathbf{x}|\theta)$ as a pdf/pf indexed by θ that is unknown?

E.g. consider the prevalence of a disease.

Frequentists:

- No, the proportion q of the population that has the disease, is not a random phenomenon but a fixed number that is simply unknown
- Example: 95% confidence interval:
  Wish to find random variables T_1 and T_2 that satisfy the probabilistic statement $P(T_1 \leq q \leq T_2) \geq 0.9$

  Interpretation: $P(T_1 \leq q \leq T_2)$ is the probability that the random interval $[T_1, T_2]$ covers q

Bayesian vs. Frequentist Inference

Should a parameter be treated as a random variable?

E.g. consider the prevalence of a disease.

Bayesians:

- Yes, the proportion Q of the population that has the disease is unknown and the distribution of Q is a subjective probability distribution that expresses the experimenters (prior) beliefs about Q
- Example: 95% credible interval:
  Wish to find constants t_1 and t_2 that satisfy the probabilistic statement $P(t_1 \leq Q \leq t_2 \mid \text{data}) \geq 0.9$

  Interpretation: $P(t_1 \leq Q \leq t_2)$ is the probability that the parameter Q is in the interval $[t_1, t_2]$.

Bayesian Inference

Prior distribution

Prior distribution: The distribution we assign to parameters before observing the random variables. Notation for the *prior pdf/pf*: We will use $p(\theta)$, the book uses $\xi(\theta)$

Likelihood

When the joint pdf/pf $f(\mathbf{x}|\theta)$ is regarded as a function of θ for given observations $x_1, \ldots, x_n$ it is called the *likelihood function*.

Posterior distribution

Posterior distribution: The conditional distribution of the parameters θ given the observed random variables $X_1, \ldots, X_n$. Notation for the *posterior pdf/pf*: We will use

$$p(\theta|x_1, \ldots, x_n) = p(\theta|\mathbf{x})$$

Bayesian Inference

Theorem 7.2.1: Calculating the posterior

Let $X_1, \ldots, X_n$ be a random sample with pdf/pf $f(x|\theta)$ and let $p(\theta)$ be the prior pdf/pf of θ. The the posterior pdf/pf is

$$p(\theta|\mathbf{x}) = \frac{f(x_1|\theta) \times \cdots \times f(x_n|\theta)p(\theta)}{g(\mathbf{x})}$$

where

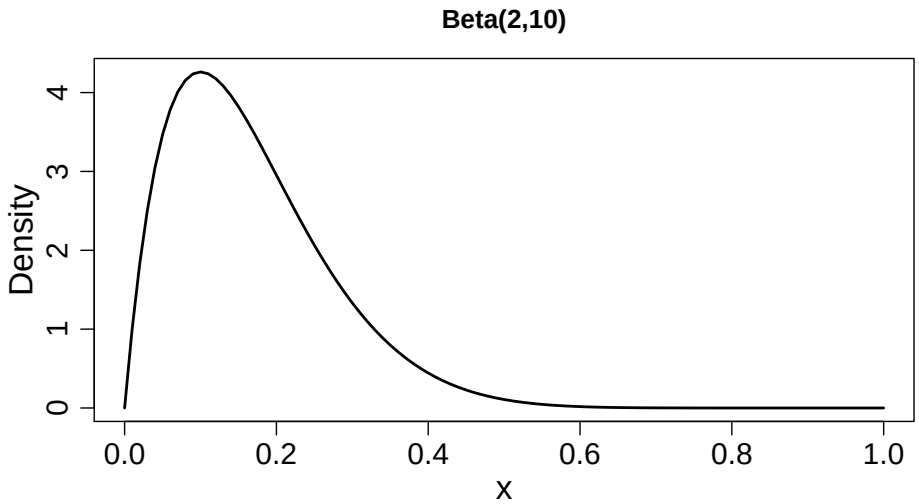
$$g(\mathbf{x}) = \int_\Omega f(\mathbf{x}|\theta)p(\theta)d\theta$$

is the marginal distribution of $X_1, \ldots, X_n$

Example: Binomial Likelihood and a Beta prior

I take a random sample of 100 people and test them all for a disease. Assume that

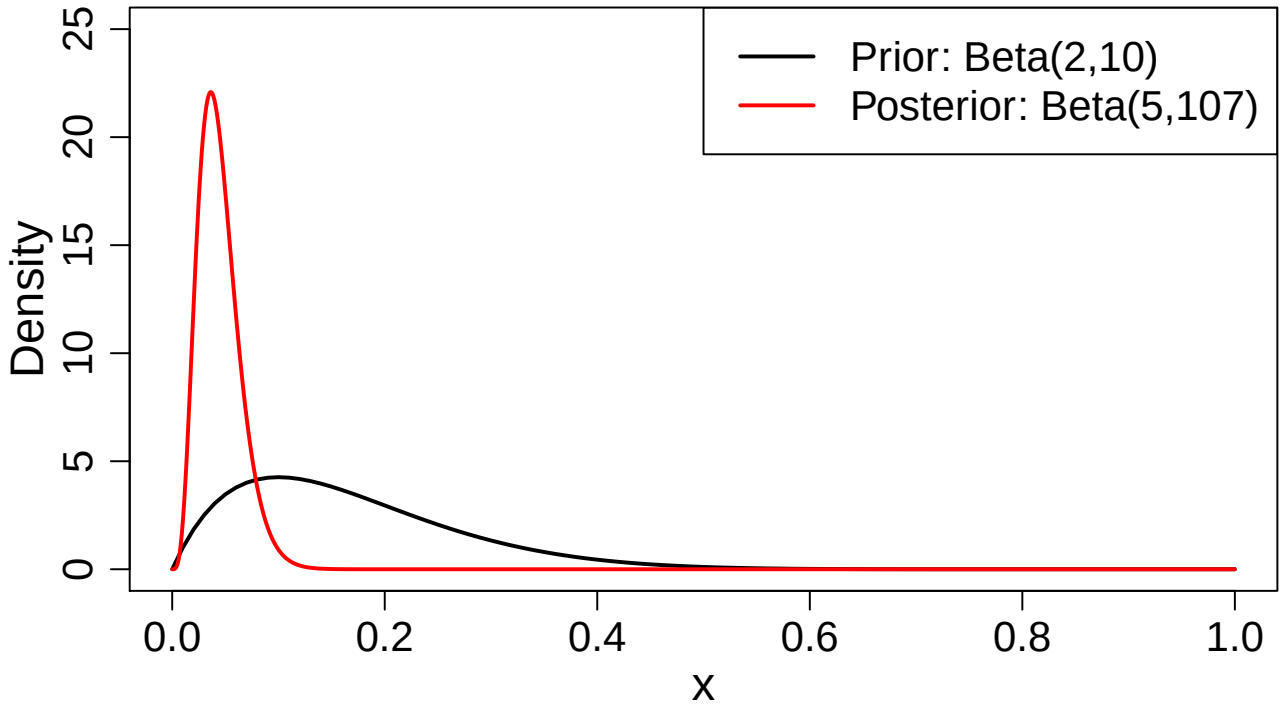
- Likelihood: $X|\theta \sim \text{Binomial}(100, \theta)$, where X denotes the number of people with the disease
- Prior: $\theta \sim \text{Beta}(2, 10)$



- I observe $X = 3$ and I want to find the posterior distribution of θ

Do the general case first: Find the posterior distribution of θ when $X|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$ where n, α and β are known.

Example: Binomial Likelihood and a Beta prior



Notice how the posterior is more concentrated than the prior.
After seeing the data we know more about θ

Bayesian Inference

Recall the formula for the posterior distribution:

$$p(\theta|\mathbf{x}) = \frac{f(x_1|\theta) \times \cdots \times f(x_n|\theta)p(\theta)}{g_n(\mathbf{x})}$$

where $g(\mathbf{x}) = \int_\Omega f(\mathbf{x}|\theta)p(\theta)d\theta$ is the marginal distribution

- $g(\mathbf{x})$ does not depend on θ
- We can therefore write

$$p(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)p(\theta)$$

- In many cases we can recognize the form of the distribution of θ from $f(\mathbf{x}|\theta)p(\theta)$, eliminating the need to calculate the marginal distribution

Example: The Binomial - Beta case

Sequential Updates

If our observations are a random sample, we can do Bayesian Analysis sequentially:

- Each time we use the posterior from the previous step as a prior:

$$
\begin{aligned}
p(\theta|x_1) &\propto f(x_1|\theta)p(\theta) \\
p(\theta|x_1, x_2) &\propto f(x_2|\theta)p(\theta|x_1) \\
p(\theta|x_1, x_2, x_3) &\propto f(x_3|\theta)p(\theta|x_1, x_2) \\
&\vdots \\
p(\theta|x_1, \ldots x_n) &\propto f(x_n|\theta)p(\theta|x_1, \ldots, x_{n-1})
\end{aligned}
$$

For example:

- Say I test 40 more people for the disease and 2 tested positive.
- What is the new posterior?

Prior distributions

- The prior distribution should reflect what we know apriori about θ
- For example: $\mathrm{Beta}(2, 10)$ puts almost all of the density below 0.5 and has a mean $2/(2 + 10) = 0.167$, saying that a prevalence of more then 50% is very unlikely
- Using $\mathrm{Beta}(1, 1)$, i.e. the $\mathrm{Uniform}(0, 1)$ indicates that *a priori* all values between 0 and 1 are equally likely.

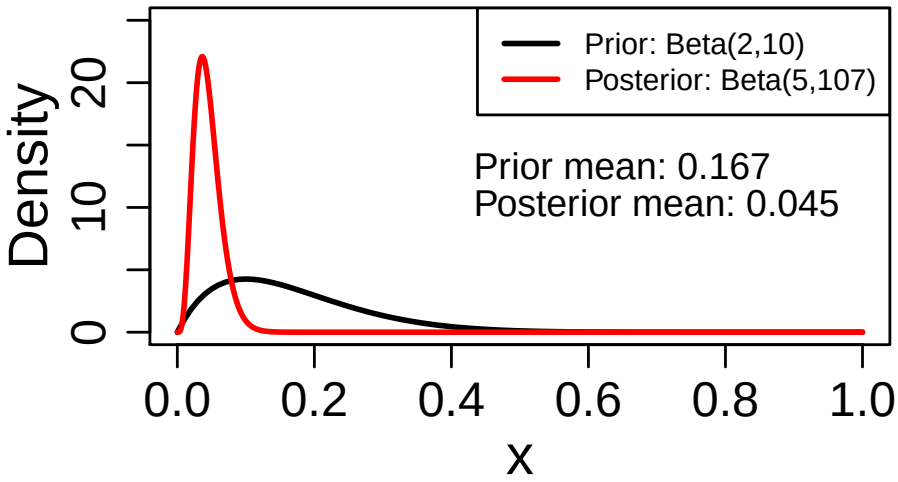
Choosing a prior

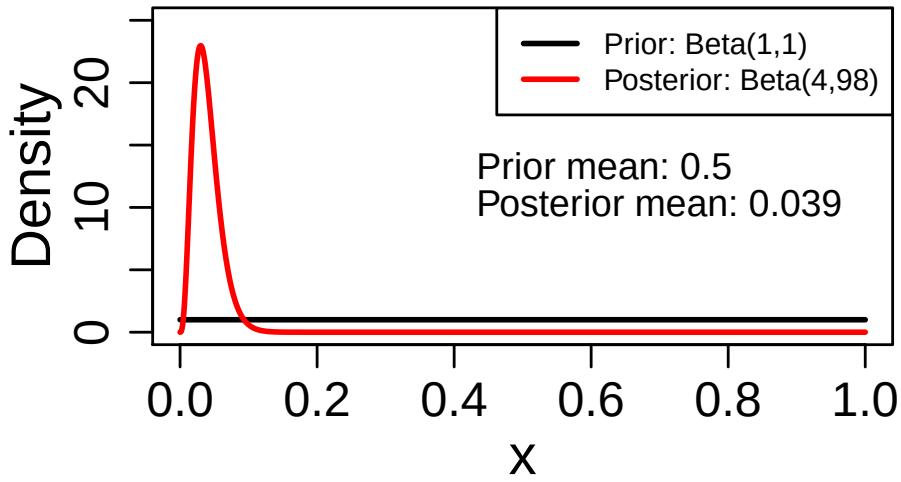
- Deciding what prior distribution to use can be very difficult
- We need a distribution (e.g. Beta) and its *hyperparameters* (e.g. α, β)
- When hyperparameters are difficult to interpret we can sometimes set a mean and a variance and solve for parameters

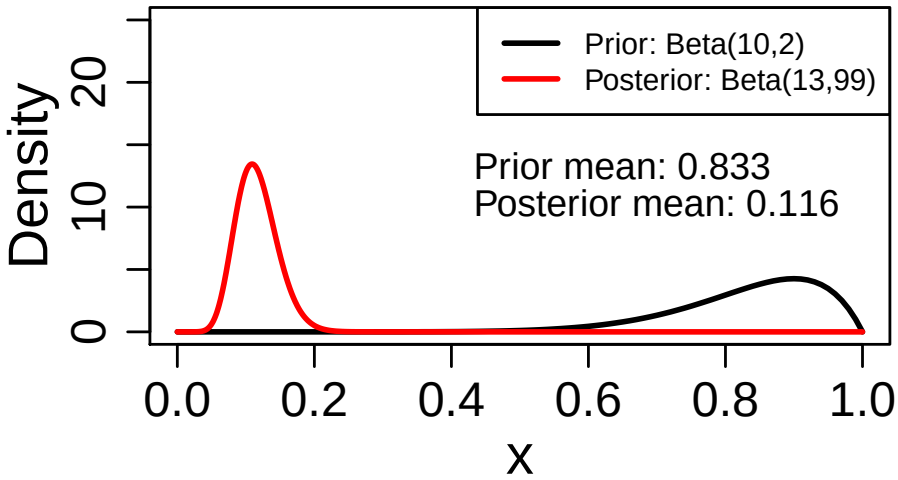
  E.g: What Beta prior has mean 0.1 and variance 0.1^2 ?
- If more than one option seems sensible, we perform *sensitivity analysis*:

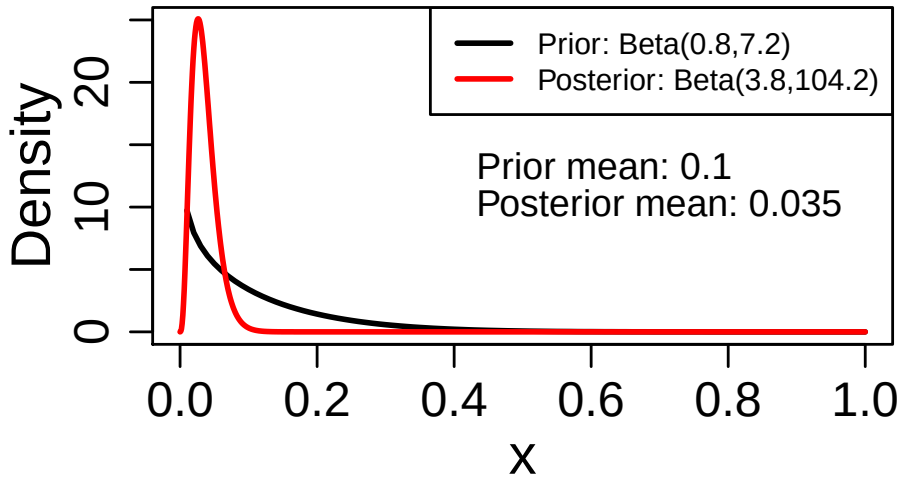
  We compare the posteriors we get when using the different priors.

Sensitivity analysis – Binomial-Beta example







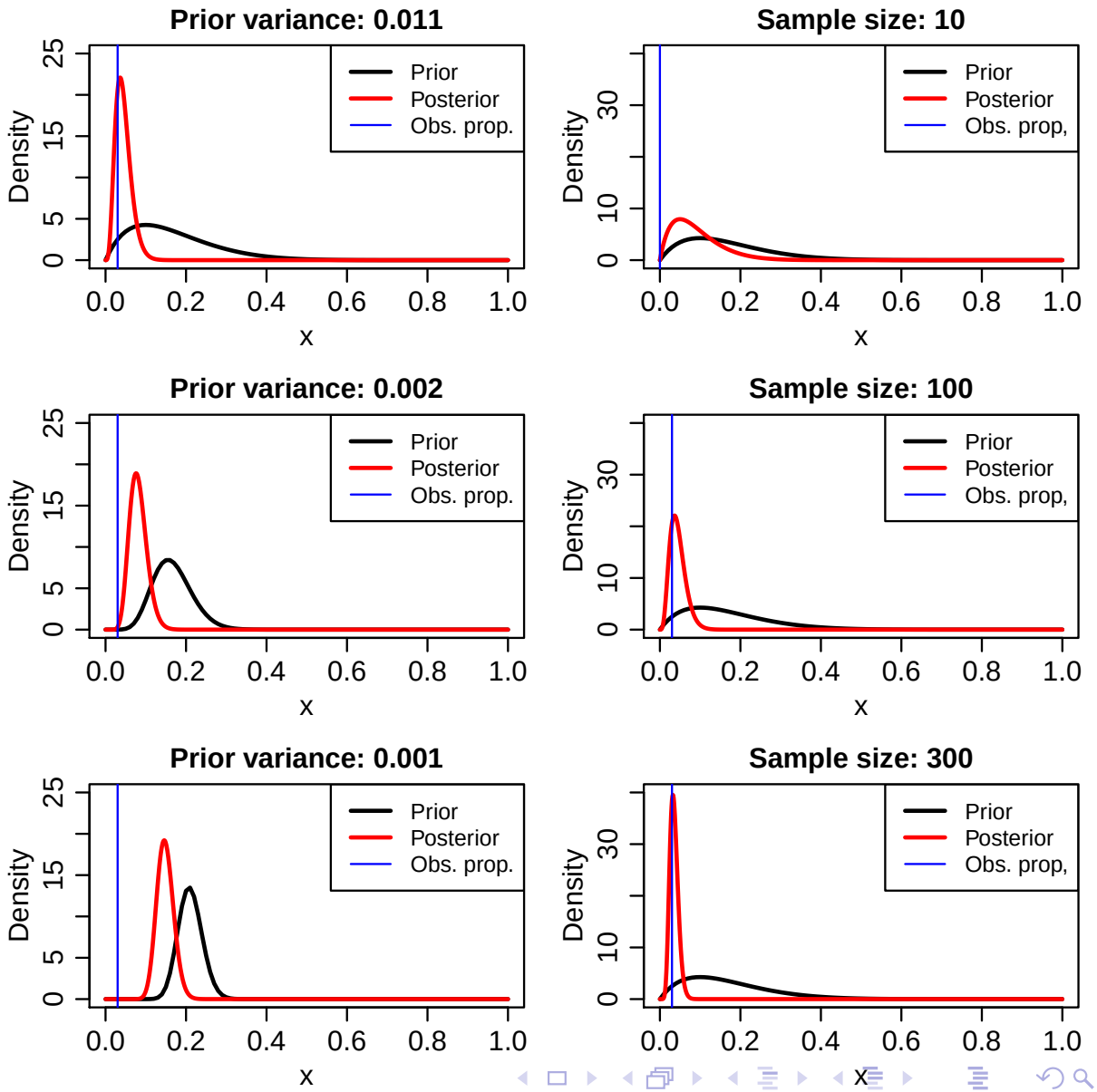


Notice: The posterior mean is always between the prior mean and the observed proportion 0.03

Effect of sample size and prior variance

The posterior is influenced both by sample size and the prior variance

- Larger sample size $\Rightarrow$ less the prior influences the posterior
- Larger prior variance $\Rightarrow$ the less the prior influences the posterior



Example - Normal distribution

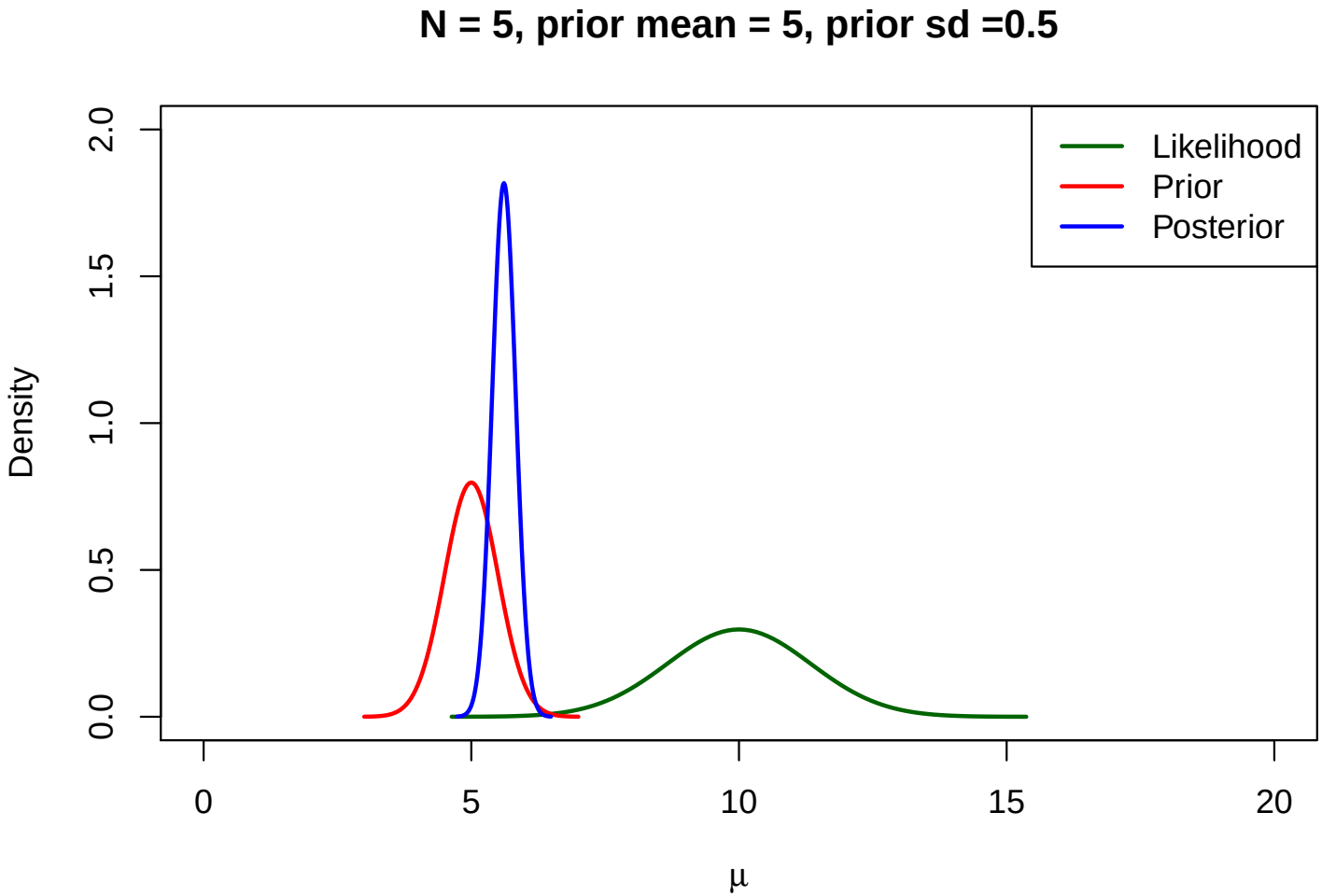
- Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \sigma^2)$ where σ^2 is known
- Let the prior distribution of θ be $N(\mu_0, \nu_0^2)$ where μ_0 and ν^2 are known.
- Show that the posterior distribution $p(\theta|\mathbf{x})$ is $N(\mu_1, \nu_1^2)$ where

$$\mu_1 = \frac{\sigma^2\mu_0 + n\nu_0^2\overline{\mathbf{x}}_n}{\sigma^2 + n\nu_0^2} \quad \text{and} \quad \nu_1^2 = \frac{\sigma^2\nu_0^2}{\sigma^2 + n\nu_0^2}$$

The posterior mean is a linear combination of the prior mean μ_0 and the observed sample mean.

- What happens when $\nu_0^2 \to \infty$?
- What happens when $\nu_0^2 \to 0$?
- What happens when $n \to \infty$?

Example - Normal distribution



GUI example...

Conjugate Priors

Def: Conjugate Priors

Let $X_1, X_2, \ldots$ be a random sample from $f(x|\theta)$. A family Ψ of distributions is called a *conjugate family of prior distributions* if for any prior distribution $p(\theta)$ in Ψ the posterior distribution $p(\theta|\mathbf{x})$ is also in Ψ

Likelihood	Conjugate Prior for θ
Bernoulli(θ)	The Beta distributions
Poisson(θ)	The Gamma distributions
$N(\theta, \sigma^2)$, σ^2 known	The Normal distributions
Exponential(θ)	The Gamma distributions

Have already see the Bernoulli-Beta and Normal-Normal cases

Conjugate prior families

- The Gamma distributions are a conjugate family for the $\text{Poisson}(\theta)$ likelihood:

  If $X_1, \ldots, X_n$ i.i.d. $\text{Poisson}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$
  then the posterior is

$$\text{Gamma}\left(\alpha + \sum_{i=1}^{n} x_i,\ \beta + n\right)$$

- The Gamma distributions are a conjugate family for the $\text{Expo}(\theta)$ likelihood:

  If $X_1, \ldots, X_n$ i.i.d. $\text{Expo}(\theta)$ and $\theta \sim \text{Gamma}(\alpha, \beta)$
  then the posterior is

$$\text{Gamma}\left(\alpha + n,\ \beta + \sum_{i=1}^{n} x_i\right)$$

Improper priors

- *Improper Prior*: A "pdf" $p(\theta)$ where $\int p(\theta)d\theta = \infty$
- Used to try to put more emphasis on data and down play the prior
- Used when there is little or no prior information about θ.
- Not clear that an improper prior is necessarily "non-informative".
- Danger: We always need to check that the posterior pdf is proper! (Integrates to 1)

Example:

- Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \sigma^2)$ and $p(\theta) = 1$, for $\theta \in \mathbb{R}$.
- Note: Here the prior variance is ∞
- Then the posterior is $N(\overline{x}_n, \sigma^2/n)$