

The background of the slide is a complex network graph. It consists of numerous small, light blue circular nodes connected by thin, grey lines. The nodes are distributed across the entire slide, with a higher density in the upper left and lower right areas. The lines connecting the nodes form a web-like structure, with some nodes having more connections than others.

# ISI-BUDS PROBABILITY - PART V

Veronica J. Berrocal

July 15, 2022

## PART V PLAN:


- Likelihood and MLE
- Bayesian inference
  - Monte Carlo methods for posterior distributions
  - Approximating posterior distributions via numerical methods

# STATISTICAL INFERENCE

- We have talked **random variables** and talked about different distributions.
- In general, **probability distribution/density functions (p.d.f.)** depend on **parameters**, which define the shape of the distribution.
- How to determine the value of the parameters? That is the goal of **statistical inference**!

**GOAL:** to infer upon the parameters given the observed data.

# LIKELIHOOD FUNCTION

- A fundamental concept in statistical inference is that of likelihood function.
- Suppose we have observed data  $y_1, y_2, \dots, y_n$  observations of random variables  $Y_1, Y_2, \dots, Y_n$  obtained as a result of some experiment/study with a given design,....
- Our understanding of the random variables and of the data  formulation of a stochastic model for the data.
- Commonly we assume the random variables  $Y_1, Y_2, \dots, Y_n$  to be independent and be distributed according to some distribution with p.d.f.  $f(y; \theta)$ .
- The likelihood function represents the *p.d.f. for the data, interpreted as a function of the parameter  $\theta$* .

# LIKELIHOOD FUNCTION

- The likelihood function  $\mathcal{L}(\theta; y_1, y_2, \dots, y_n)$  for data  $y_1, y_2, \dots, y_n$ , observations of independent and identically distributed random variables is given by:

$$\mathcal{L}(\theta; y_1, y_2, \dots, y_n) = \prod_{i=1}^n f(y_i; \theta)$$

- In frequentist inference, we perform inference on the parameter  $\theta$  via Maximum Likelihood Estimation by finding the value of  $\theta$  that maximizes the likelihood function and thus maximizes the chance of occurrence for the data that we have observed.
- The value  $\hat{\theta}$  of  $\theta$  that maximized the likelihood function is called the Maximum Likelihood Estimate (MLE) and is taken as the estimate of the parameter  $\theta$ :

$$\hat{\theta} := \max_{\theta} \mathcal{L}(\theta; y_1, y_2, \dots, y_n)$$

# MLE IN PRACTICE

- To find the MLE in practice, we:
  1. take the log of the likelihood function, denoted by  $\ell(\theta; y_1, y_2, \dots, y_n) = \log[\mathcal{L}(\theta; y_1, y_2, \dots, y_n)]$
  2. take the derivative(s) of  $\ell(\theta; y_1, y_2, \dots, y_n)$  with respect to  $\theta$  (partial derivatives if  $\theta$  consists of multiple parameters) and set it/them equal to 0.
- To verify that the solution  $\hat{\theta}$  of  $\frac{\partial \ell(\theta; y_1, y_2, \dots, y_n)}{\partial \theta} = 0$  is a point of maximum, we might want to check that  $\frac{\partial^2 \ell(\theta; y_1, y_2, \dots, y_n)}{\partial \theta^2} \Big|_{\hat{\theta}} < 0$



# EXAMPLE: PROBABILITY OF SIDE EFFECTS

Suppose that we would like to evaluate the side effects of a new drug.

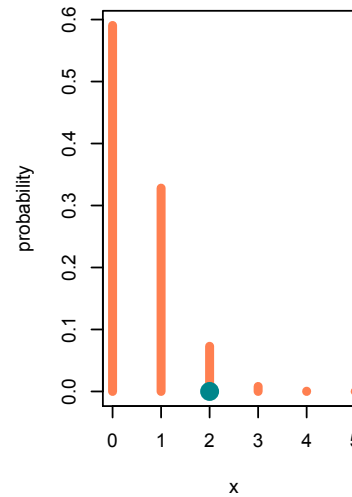
We hire  $n=5$  patients to take the drug and we record their responses: two patients report side effects. We want to estimate the probability that a patient will experience side effects when taking this new drug.

**Stochastic model:** We denote by  $Y_1, Y_2, \dots, Y_5$  the response of the 5 patients with  $Y_i = 1$  if patient  $i$  experiences a side effect,  $Y_i = 0$  otherwise.

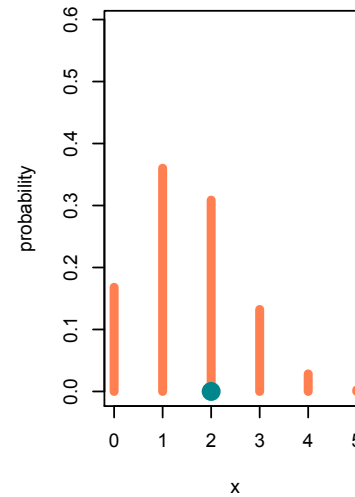
**Likelihood function:** .....

EXAMPLE:  
PROBABILITY OF  
SIDE EFFECTS

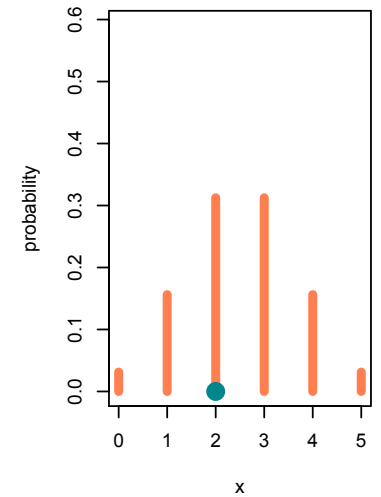
**N=5, p=0.1**



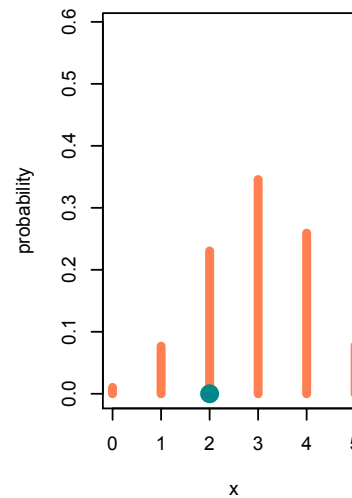
**N=5, p=0.3**



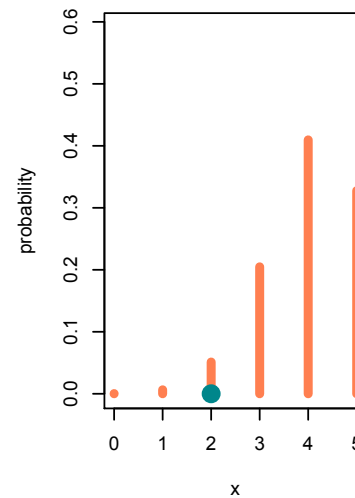
**N=5, p=0.5**



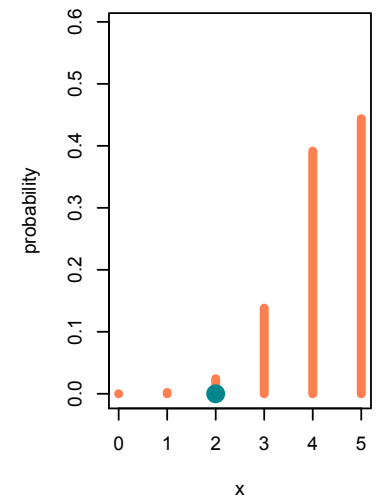
**N=5, p=0.6**



**N=5, p=0.8**



**N=5, p=0.85**





# BAYESIAN INFERENCE

- Another way to perform inference on parameters is statistics is within the Bayesian framework.
- In this case, we:
  1. Start with a sampling model for the data:  $y_1, y_2, \dots, y_n | \theta \stackrel{\text{iid}}{\sim} f(y; \theta)$
  2. Provide a prior distribution for the parameter  $\theta$ :  $\theta \sim p(\theta)$
  3. Derive the posterior distribution for  $\theta$  given the data up to a proportionality constant using Bayes' theorem

$$f(\theta; y_1, y_2, \dots, y_n) \propto \mathcal{L}(\theta; y_1, y_2, \dots, y_n) \cdot p(\theta)$$

# BAYESIAN INFERENCE

- We can summarize the posterior distribution by calculating
  - [Point estimate] The mean or median of the posterior distribution, called posterior mean and posterior median, respectively.
  - [Interval-valued estimate] A  $100 \cdot (1 - \alpha)\%$  equal-tailed credible interval for  $\theta$  by deriving the  $100 \cdot \frac{\alpha}{2}$ -th and the  $100 \cdot \left(1 - \frac{\alpha}{2}\right)$ -th percentiles, respectively, in the posterior distribution.

# Monte Carlo Methods for Posterior Distributions

- Suppose that we have observations  $y_1, y_2, \dots, y_n$  of  $n$  i.i.d. random variables  $Y_1, Y_2, \dots, Y_n$  that we can model as follows:

$$Y_1, Y_2, \dots, Y_n | \theta \stackrel{\text{iid}}{\sim} f(y; \theta)$$
$$\theta \sim p(\theta)$$

- We want to derive the posterior distribution  $f(g(\theta) | y_1, y_2, \dots, y_n)$  of  $g(\theta)$  for some function  $g(\cdot)$ .
- We can use Monte Carlo methods to approximate it!

# MONTE CARLO METHODS: ALGORITHM

- We will generate  $B$  values  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(B)}$  independently from the posterior distribution  $f(\theta|y_1, y_2, \dots, y_n)$ .
- For each  $i$ , we will evaluate  $\gamma^{(i)} = g(\theta^{(i)})$ .
- Then:  $\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(B)}$  is a sample from the posterior distribution  $f(g(\theta)|y_1, y_2, \dots, y_n)$ .
- Empirical distribution of  $\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(B)}$  is an approximation to  $f(g(\theta)|y_1, y_2, \dots, y_n)$ .
- Mean of  $\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(B)}$  is an approximation to the posterior mean of  $g(\theta)$ .
- Variance of  $\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(B)}$  is an approximation to the posterior variance of  $g(\theta)$ .

## EXAMPLE: RISK OF BREAST CANCER

The most important factor in early breast cancer is the number of axillary lymph nodes that test positive for breast cancer. There is no standard number of lymph nodes to sample. Sometimes surgeon remove 3 nodes and sometimes they remove 30. The proportion  $\theta$  of lymph nodes that are positive has approximately a  $p(\theta) = \text{Beta}(0.1, 5)$  density. So, a priori, the probability that a sampled lymph node is positive is:  $E(\theta) = \frac{\alpha}{\alpha + \beta} = \frac{0.1}{5.1} = 0.02$ , approximately 2%.

A woman with breast cancer had a mastectomy and subsequently the surgeon removed 20 lymph nodes. Eight tested positive.

What is the updated probability of positive lymph nodes? What about the risk that the woman has positive lymph nodes?

# EXAMPLE: RISK OF BREAST CANCER

- **Sampling model:** Let  $Y_i$  denote the negative/positive (0/1) test result for the  $i$ -th removed lymph node. Then,

$$y_1, y_2, \dots, y_{20} \overset{\text{iid}}{\sim} \text{Binom}(1, \theta)$$

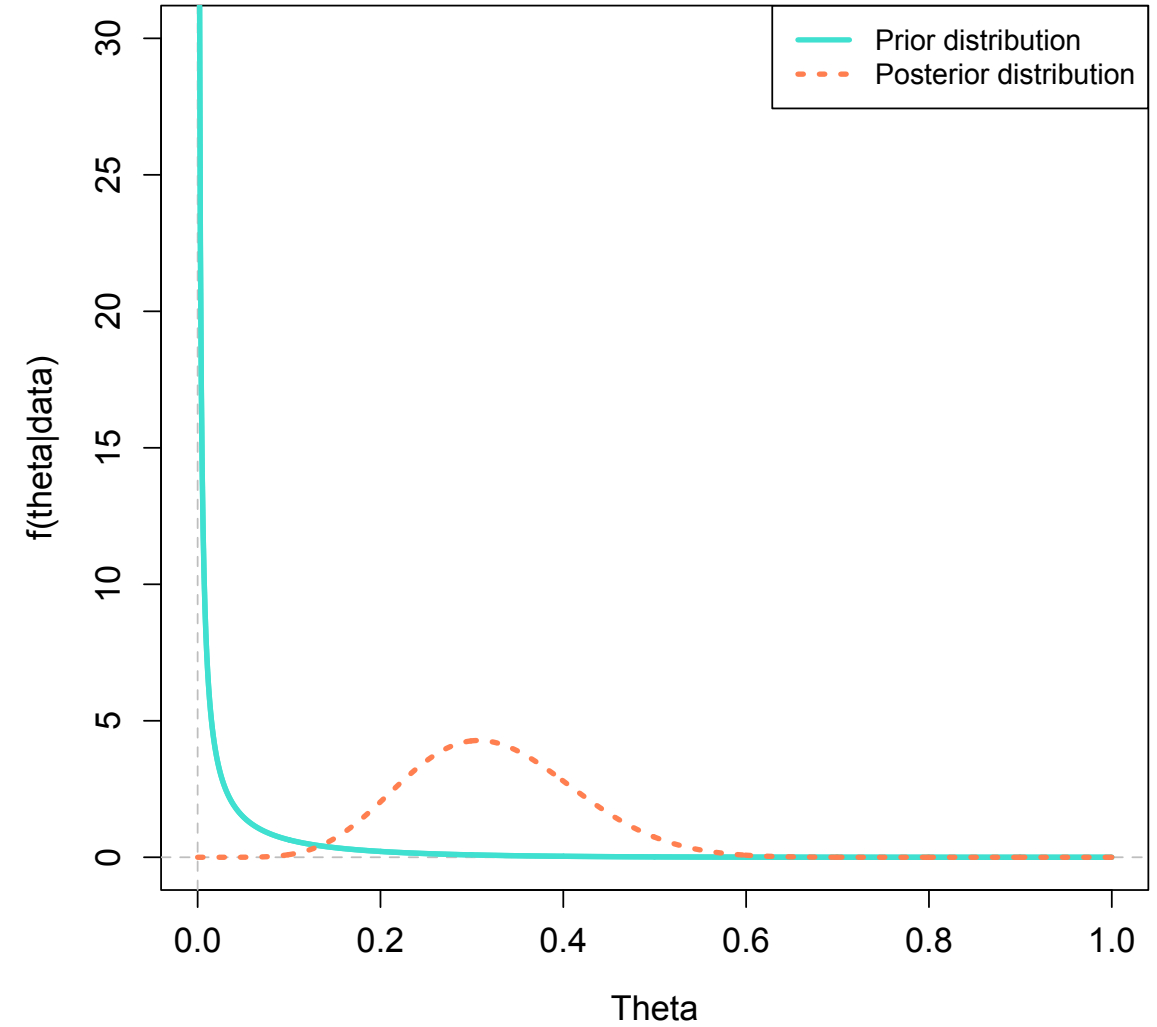
- **Prior distribution:** the probability  $\theta$  that a lymphnode tests positive is at priori  
 $\theta \sim p(\theta) = \text{Beta}(0.1, 5)$

- Then, the **posterior distribution** (see lab in the afternoon) of  $\theta$  given 8 positive lymphnodes is:

$$f(\theta | y_1, y_2, \dots, y_{20}) \sim \text{Beta}(8.1, 17)$$

# PRIOR AND POSTERIOR DISTRIBUTION

**Prior and posterior distribution**  
**Probability of positive lymphnode**





# EXAMPLE: RISK OF BREAST CANCER

- A point estimate of the probability that the woman has a positive lymphnode given the 8 positive sampled ones can be the posterior mean:  $\hat{\theta} = E(\theta | y_1, y_2, \dots, y_{20}) =$ .
- What about the risk that she has a positive lymph node, given then 8 positive sampled lymphnodes?
- By definition, risk is:

Risk := Probability of success / Probability of failure.

Hence:  $g(\theta) = \frac{\theta}{1-\theta}$ . We want to derive the posterior distribution of  $\frac{\theta}{1-\theta}$ .

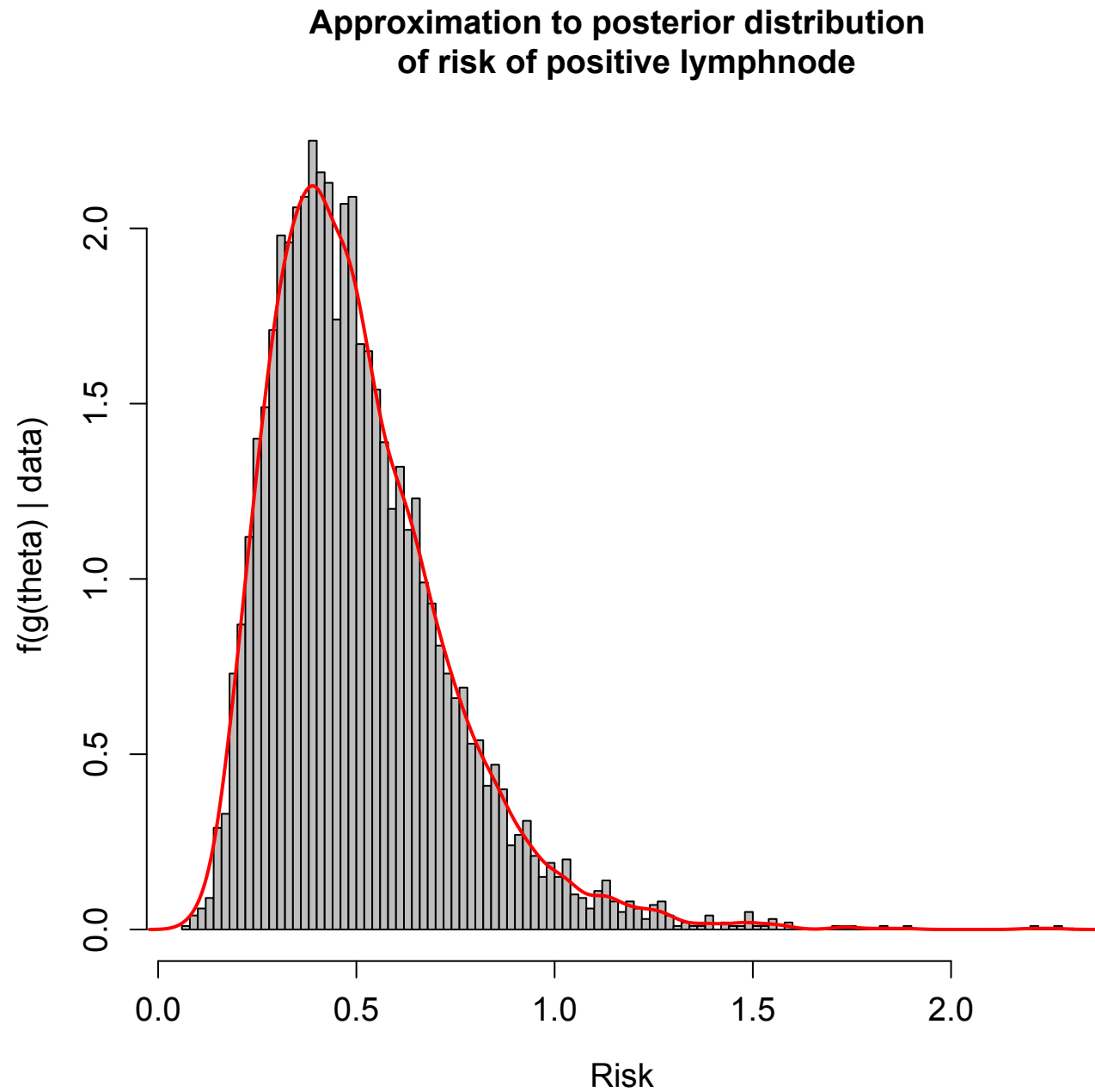
## EXAMPLE: RISK OF BREAST CANCER

- We sample  $B=5,000$  values  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(B)}$  independently from  $f(\theta|y_1, y_2, \dots, y_{20}) \sim \text{Beta}(8.1, 17)$  and we compute

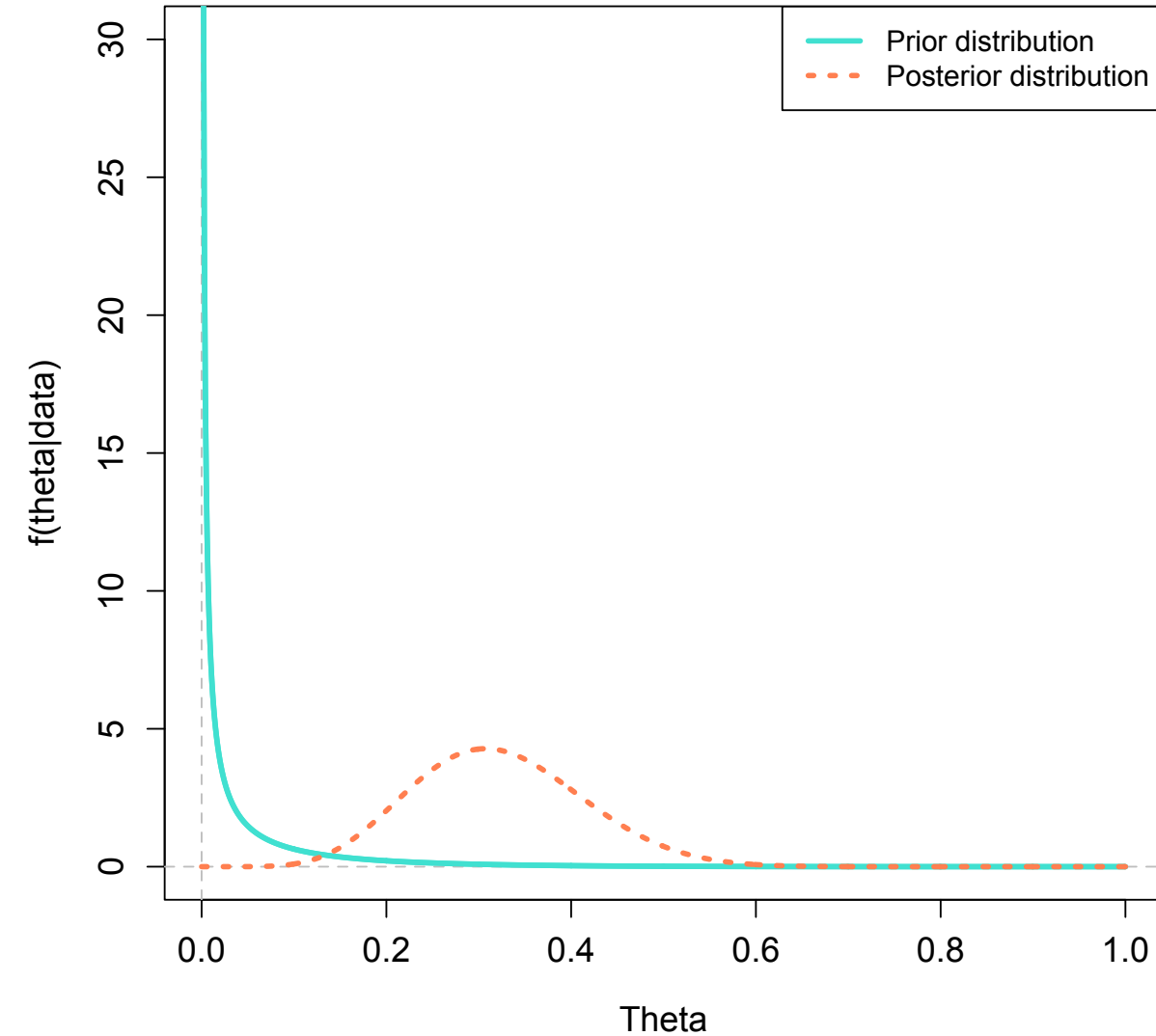
$$\gamma^{(1)} = \frac{\theta^{(1)}}{1-\theta^{(1)}}, \gamma^{(2)} = \frac{\theta^{(2)}}{1-\theta^{(2)}}, \dots, \gamma^{(5,000)} = \frac{\theta^{(5,000)}}{1-\theta^{(5,000)}}.$$

- The histogram of  $\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(5,000)}$  is an approximation of the posterior distribution of the risk of a positive lymphnode given the 8 observed positive lymphnodes.

RISK OF POSITIVE  
LYMPHNODE



**Prior and posterior distribution  
Probability of positive lymphnode**



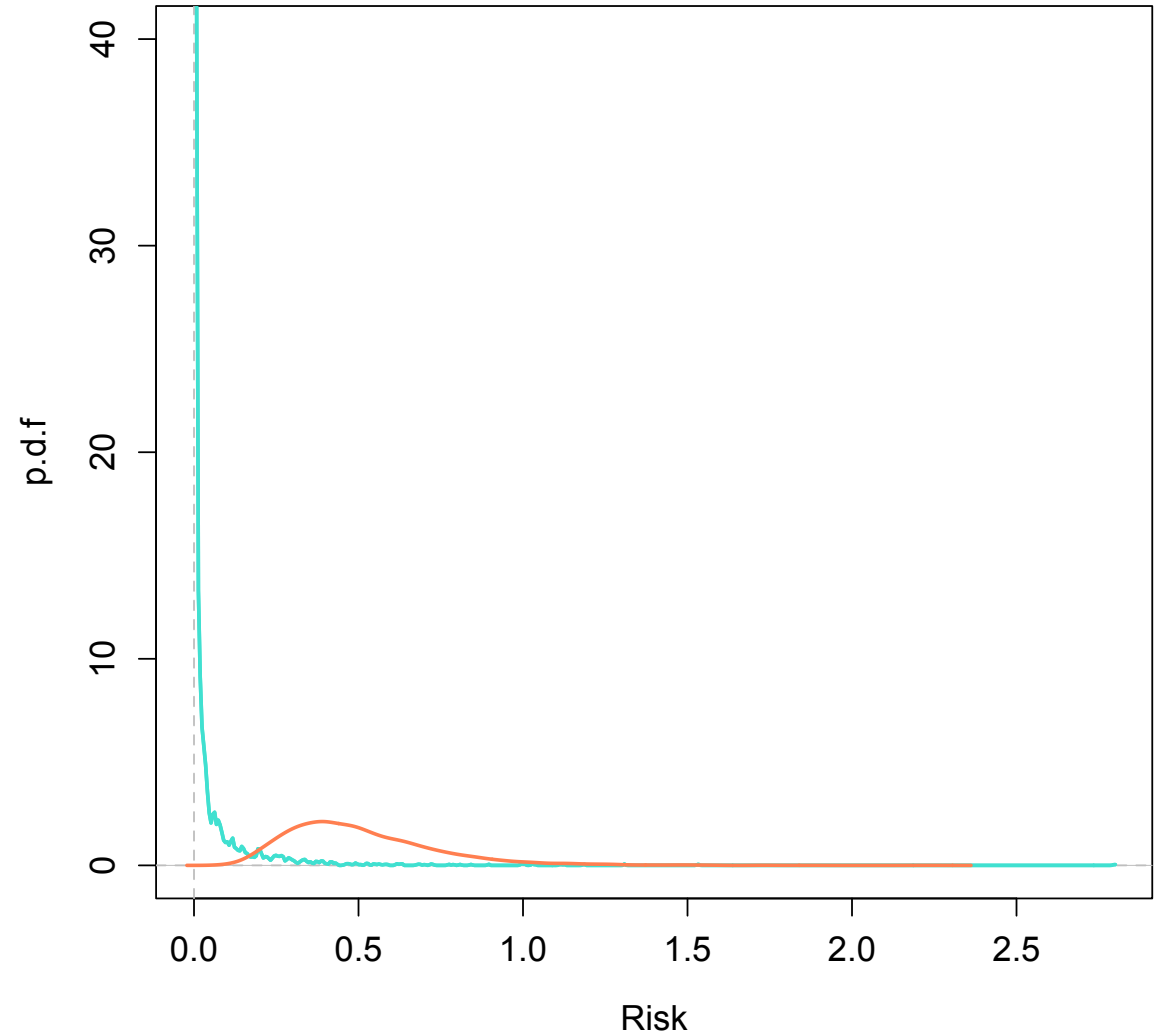
## EXAMPLE: RISK OF BREAST CANCER

What if we want to do a plot of the prior  
vs posterior distribution of the risk of  
positive lymphnode like we did here?

How can we do it?

EXAMPLE: RISK  
OF BREAST  
CANCER

Prior and posterior distribution  
of risk of positive lymphnode



# A PARENTHESIS: THE GAMMA DISTRIBUTION

- Remember the Gamma distribution we saw previously.
- A positive random variable  $Y$  is said to follow a Gamma distribution with shape parameter  $\alpha$  and rate parameter  $\beta$  if its p.d.f. is given by:

$$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot y^{\alpha-1} \cdot e^{-\beta y}$$

with  $\alpha > 0$  and  $\beta > 0$ . Then:  $E(Y) = \frac{\alpha}{\beta}$  and  $Var(Y) = \frac{\alpha}{\beta^2}$ .

# A PARENTHESIS: THE INVERSE GAMMA DISTRIBUTION

- If  $Y \sim \text{Gamma}(\alpha, \beta)$  then  $X = \frac{1}{Y} \sim \text{InvGamma}(\alpha, \beta)$ .
- The Inverse Gamma distribution with shape parameter  $\alpha$  and scale parameter  $\beta$  has p.d.f. given by

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{x^{\alpha+1}} \cdot e^{-\beta/x}$$

with  $\alpha > 0$  and  $\beta > 0$ . Then:  $E(Y) = \frac{\beta}{\alpha-1}$  and  $\text{Var}(X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ .

- The Inverse Gamma distribution is often used as prior distribution for parameters that represent variances.



# APPROXIMATION TO THE POSTERIOR DISTRIBUTION

- Suppose we have data  $y_1, y_2, \dots, y_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$  and we want to infer upon the mean  $\mu$  and variance  $\sigma^2$  of this distribution.
- We need to place a prior distribution on  $\mu$  and  $\sigma^2$ :  $p(\mu, \sigma^2)$ .
- Let's assume that  $\mu$  and  $\sigma^2$  are independent a priori:  $p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2)$  and that

$$p(\mu) = N(\mu_0, \tau_0^2)$$

$$p(\sigma^2) = \text{InvGamma}(\alpha_0, \beta_0).$$

# APPROXIMATION TO THE POSTERIOR DISTRIBUTION

- The joint prior distribution  $p(\mu, \sigma^2)$  is:

$$p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2) \propto \frac{1}{\tau_0^2} \cdot \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right) \\ \cdot \frac{1}{(\sigma^2)^{\alpha_0+1}} \cdot \exp\left(-\frac{\beta_0}{\sigma^2}\right)$$

is not conjugate!

- The joint posterior distribution  $p(\mu, \sigma^2 | y_1, y_2, \dots, y_n)$  does not follow any standard parametric form.

# APPROXIMATION VIA NUMERICAL METHODS

- We can approximate the joint posterior distribution  $p(\mu, \sigma^2 | y_1, y_2, \dots, y_n)$  using numerical methods.
  1. We choose a grid of values evenly spaced for  $\mu, \sigma^2$ :  $\{\mu_1, \mu_2, \dots, \mu_G\}$  and  $\{\sigma_1^2, \sigma_2^2, \dots, \sigma_H^2\}$ .
  2. We approximate  $p(\mu, \sigma^2 | y_1, y_2, \dots, y_n)$  at  $(\mu_j, \sigma_k^2)$  for  $j = 1, 2, \dots, G$  and  $k = 1, 2, \dots, H$  via:

$$p_D(\mu_j, \sigma_k^2 | y_1, y_2, \dots, y_n)$$

# APPROXIMATION VIA NUMERICAL METHODS

- How do we calculate  $p_D(\mu_j, \sigma_K^2 | y_1, y_2, \dots, y_n)$ ?

# APPROXIMATION VIA NUMERICAL METHODS: REMARKS

- $p_D(\mu, \sigma^2 | y_1, y_2, \dots, y_n)$  is an approximation to the joint posterior distribution  $p(\mu, \sigma^2 | y_1, y_2, \dots, y_n)$  since it sums up to 1.
- We can also derive the approximate discrete marginal posterior distributions  $p_D(\mu | y_1, y_2, \dots, y_n)$  and  $p_D(\sigma^2 | y_1, y_2, \dots, y_n)$  using the grid of values  $\{\mu_1, \mu_2, \dots, \mu_G\}$  and  $\{\sigma_1^2, \sigma_2^2, \dots, \sigma_H^2\}$ :

$$p_D(\mu_j | y_1, y_2, \dots, y_n) = \sum_{k=1}^H p_D(\mu_j, \sigma_k^2 | y_1, y_2, \dots, y_n) \quad j = 1, \dots, G$$

- Note that here is not need to normalize here!

## EXAMPLE: LENGTH OF MIDGE WINGS

Consider data on the length of 9 midge wings:  $y_1 = 1.64$ ,  $y_2 = 1.70$ ,  $y_3 = 1.72$ ,  $y_4 = 1.74$ ,  $y_5 = 1.82$ ,  $y_6 = 1.82$ ,  $y_7 = 1.82$ ,  $y_8 = 1.90$  and  $y_9 = 2.08$ .

We want to infer upon the mean and variance of the normal distribution for the midge wings' length.

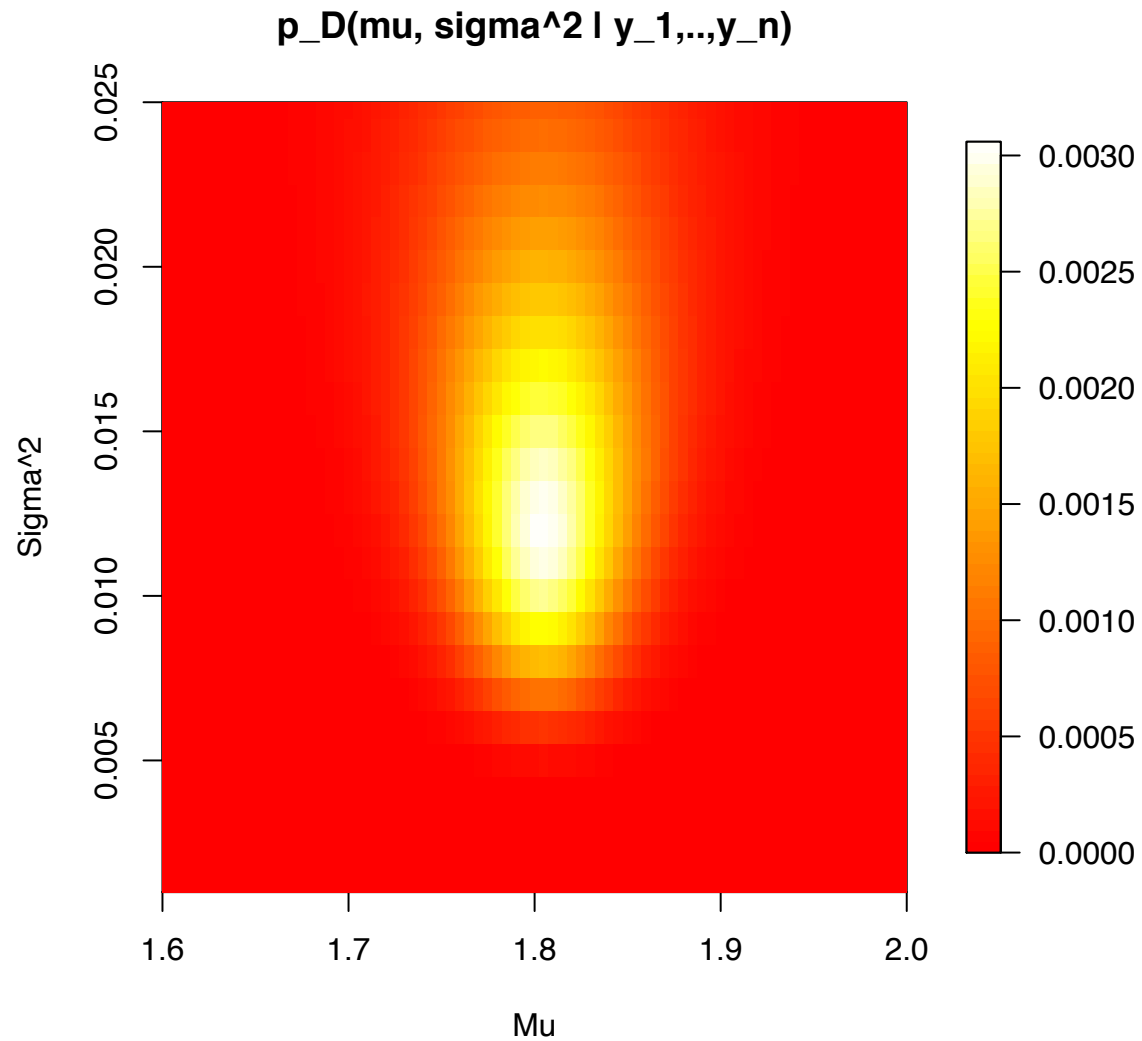
We use the following priors:  $p(\mu, \sigma^2) = p(\mu) \cdot p(\sigma^2) = N(\mu_0 = 1.9, \tau_0^2 = 0.95^2) \cdot \text{InvGamma}(\alpha_0 = 0.5, \beta_0 = 0.005)$ .

## EXAMPLE: LENGTH OF MIDGE WINGS

We approximate the posterior distribution  $p(\mu, \sigma^2 | y_1, y_2, \dots, y_9)$  using the following grid of 100 values for  $\mu$  and  $\sigma^2$ :

- $\{\mu_1, \mu_2, \dots, \mu_G\} = \{1.505, 1.510, \dots, 2.00\}$
- $\{\sigma_1^2, \sigma_2^2, \dots, \sigma_H^2\} = \{0.001, 0.002, \dots, 0.10\}$





APPROXIMATE  
JOINT POSTERIOR  
DISTRIBUTION

# APPROXIMATE MARGINAL DISTRIBUTION

