

# Session 2

## Bayesian Methods

Yuri Balasanov

University of Chicago, MScA

© Y. Balasanov, iLykei 2016

© Yuri Balasanov, iLykei 2016

All Rights Reserved

No part of this lecture notes document or any of its contents may be reproduced, copied, modified or adapted without the prior written consent of the author, unless otherwise indicated for stand-alone materials.

The content of these lectures, any comments, opinions and materials are put together by the author especially for the course Linear and Nonlinear Statistical Models, they are sole responsibility of the author, but not of the author's employers or clients.

The author cannot be held responsible for any material damage as a result of use of the materials presented in this document or in this course.

For any inquiries contact the author, Yuri Balasanov, at  
ybalasan@uchicago.edu or yuri.balasanov@iLykei.com .

# Outline of the Session

- Conditional probability and Bayes theorem
- Examples of application of Bayes theorem
- General framework of Bayesian analysis
- Role of advanced calculations in Bayesian analysis
- Likelihood function for binomial distribution
- Concept of conjugate distributions
- Conjugate prior for binomial data: beta distribution
- Compromise between the prior and the data in binomial case

# Bayes Theorem I

## Definition

Conditional probability: For two events  $A$  and  $B$

$$P\{A|B\} \doteq \frac{P\{A \cap B\}}{P\{B\}} = \frac{P\{AB\}}{P\{B\}}$$

From the definition it follows that

$$P\{AB\} = P\{B\} P\{A|B\} = P\{A\} P\{B|A\}$$

and if  $\bar{A}$  is the complimentary event to  $A$  ( $P\{A \cup \bar{A}\} = 1$ ) then

$$P\{A|B\} = \frac{P\{A\} P\{B|A\}}{P\{B\}} = \frac{P\{A\} P\{B|A\}}{P\{B|A\} P\{A\} + P\{B|\bar{A}\} P\{\bar{A}\}}$$

## Fact

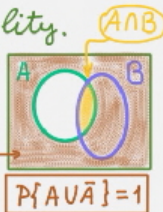
**Bayes Theorem.** If  $H_1, \dots, H_k$  are such that  $P\{H_1 \cup \dots \cup H_k\} = 1$ ;  $H_i \cap H_j = \emptyset$  for  $i \neq j$ , then  $P\{H_i|B\} = \frac{P\{H_i\}P\{B|H_i\}}{\sum_{i=1}^k P\{B|H_i\}P\{H_i\}}$

# Bayes Theorem II

## Bayes Theorem.

Definition of conditional probability.

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}} \quad \leftarrow \text{Chance for a point from } B \text{ to belong to } A$$



$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}} \quad P\{A \cap B\} = P\{B\}P\{A|B\}$$

$$P\{B|A\} = \frac{P\{A \cap B\}}{P\{A\}} \quad P\{A \cap B\} = P\{A\}P\{B|A\}$$

$$P\{B\}P\{A|B\} = P\{A\}P\{B|A\} \Rightarrow P\{A|B\} = \frac{P\{A\}P\{B|A\}}{P\{B\}}$$

$$P\{B\} = P\{A\}P\{B|A\} + P\{\bar{A}\}P\{B|\bar{A}\}$$

$$P\{A|B\} = \frac{P\{A\}P\{B|A\}}{P\{A\}P\{B|A\} + P\{\bar{A}\}P\{B|\bar{A}\}}$$

Bayes Theorem

$$\left. \begin{array}{l} \text{General case. } H_1, \dots, H_K: \\ H_i \cap H_j = \emptyset, i \neq j; P\{\bigcup_{i=1}^K H_i\} = 1 \end{array} \right| \Rightarrow P\{H_j|B\} = \frac{P\{H_j\}P\{B|H_j\}}{\sum_{i=1}^K P\{H_i\}P\{B|H_i\}}$$

# Example of Bayesian Thinking

## Example

A rare disease appears in 1 person out of 1000, its presence is coded by  $\theta$ :

$$\theta = \begin{cases} \theta_1, & \text{if a person has the disease,} \\ \theta_0, & \text{if a person has no disease.} \end{cases}$$

Our prior belief: for a random person  $P\{\theta = \theta_1\} = 0.001$ .

There is a test for the disease with 99% hit rate, its result is coded by  $\tau$ :

$$\tau = \begin{cases} \tau_1, & \text{if test result is positive,} \\ \tau_0, & \text{if test result is negative} \end{cases}$$

Then  $P\{\tau_1|\theta_1\} = 0.99$ .

The test has 5% false alarm rate:  $P\{\tau_1|\theta_0\} = 0.05$ .

A random person is tested positive. Find posterior probability  $P\{\theta_1|\tau_1\}$ .

Answer the question to check your intuition.

# Example of Bayesian Thinking

Example of Bayesian thinking.

$\theta = \{\theta_1, \text{disease} \quad \tau = \{\tau_1, \text{test positive},$   
 $\theta_0, \text{no disease} \quad \tau_0, \text{test negative}.$

$$P\{\theta_1\} = 0.001$$

$$P\{\tau_1|\theta_1\} = 0.99 \quad P\{\tau_1|\theta_0\} = 0.05$$

Joint Distribution of  $\theta, \tau$ .

	$\theta_1$	$\theta_0$	
$\tau_1$	$P\{\tau_1 \theta_1\}P\{\theta_1\} = 0.99 \cdot 0.001$	$P\{\tau_1 \theta_0\}P\{\theta_0\} = 0.05(1-0.001)$	$\sum_{\theta} P\{\tau_1 \theta\}P\{\theta\}$
$\tau_0$	$P\{\tau_0 \theta_1\}P\{\theta_1\} = (1-0.99) \cdot 0.001$	$P\{\tau_0 \theta_0\}P\{\theta_0\} = (1-0.05)(1-0.001)$	$\sum_{\theta} P\{\tau_0 \theta\}P\{\theta\}$
	$P\{\theta_1\} = 0.001$	$P\{\theta_0\} = 1-0.001$	

Use Bayes theorem.

$$P\{\theta_1|\tau_1\} = \frac{P\{\tau_1|\theta_1\}P\{\theta_1\}}{\sum_{\theta} P\{\tau_1|\theta\}P\{\theta\}}$$

$$= \frac{0.99 \cdot 0.001}{P\{\tau_1|\theta_1\}P\{\theta_1\} + P\{\tau_1|\theta_0\}P\{\theta_0\}} = \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05(1-0.001)} = 0.019$$

# Summary of the Example

- The probability calculated in the example may seem unintuitive: the test hit rate is  $P\{\tau_1|\theta_1\} = 0.99$ , false alarm probability is  $P\{\tau_1|\theta_0\} = 0.05$ . Yet for a randomly selected person with positive test result posterior probability of disease is low:  $P\{\theta_1|\tau_1\} = 0.019$ . Such posterior may cast doubt in preventive screening, especially if there are significant negative consequences of false positive

## Example

There is a discussion of controversy of mammogram screening of breast cancer. In question is significance of the effect of screening in comparison with the consequences of false positive test result. See the workshop

- The reasons for unexpectedly low posterior probability are:
  - Low prior. A person randomly selected for testing rare outcome. If tested person is from higher risk group the posterior will be better
  - The false positive probability in fact is not low: compare  $0.99 \times 0.001 = 0.00099$  and  $0.05 \times (1 - 0.001) = 0.04995$  in the denominator



# Comparison of Approaches


- How do we collect data for the previous example?
- How do we answer the question without using Bayesian analysis, but using only arguments of frequentist approach?
- Remember the main difference between the approaches:
  - Frequentist assumes that there is an unknown parameter  $\theta$  which can be estimated using the sample  
Sample changes - parameter remains unchanged  
In order to find estimate maximize the likelihood with respect to  $\theta$ : find such  $\hat{\theta}$  that maximizes  $L(\theta; Data) = P\{Data|\theta\}$
  - Bayesianist assumes that there is a random variable  $\theta$  with an unknown distribution  $F_{\theta}(x) = P\{\theta \leq x|Data\}$   
This distribution is estimated by combining the likelihood function  $L(\theta; Data) = P\{Data|\theta\}$  with the prior distribution  $P\{\theta \leq x\}$   
Parameter is not an unknown constant, but a random variable, it changes; but sample is not changing  
In order to find  $P\{\theta \leq x|Data\}$  flip it into  $P\{Data|\theta\}$  using Bayes theorem

# Bayes Theorem in Statistics

Bayes Theorem in Statistics.

Model: 

Parameter:  $\theta$  Data

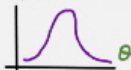
Prior Distribution 

Likelihood

Prior

Bayes Theorem:  
Posterior Distr.

$$P(\theta | \text{Data}) = \frac{P(\text{Data} | \theta) P(\theta)}{P(\text{Data})}$$



Model is a condition in all terms!

Model

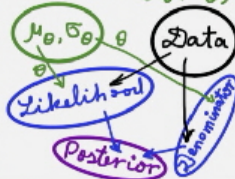
Steps of Bayesian Statistics.

1. Specify the model  $m_\theta$
2. Define prior  $F_\theta(x)$
3. From Data and model  $m_\theta$  find likelihood
4. Find denominator by integrating over all  $\theta$
5. Find posterior  $F_\theta$

Example

$\text{Norm}(\theta, 1)$

$\text{Norm}(\mu_\theta, \sigma_\theta)$



# Bayes Theorem in Statistics

## Numerator

- Numerator in Bayes theorem expression for statistical analysis consists of 2 terms:

$$P\{\theta | Data, Model\} = \frac{P\{Data | \theta, Model\} P\{\theta | Model\}}{P\{Data | Model\}}$$

- The first term is likelihood function  $P\{Data | \theta, Model\}$ . For each value of the parameter  $\theta$  it shows the probability of the observed data
- The second term is marginal distribution of the parameter. This distribution appears in the right-hand side, so it is known before the analysis, and even before the experiment is conducted. Thus the name prior distribution
- Both terms can only be calculated after the model generating the data is specified. For example, in the mammography experiment the model is binomial: the data are generated by binomial distribution with parameter  $\theta = p$  equal to probability of disease

# Bayes Theorem in Statistics

## Denominator

- In the central relationship of Bayesian analysis

$$P\{\theta | Data, Model\} = \frac{P\{Data | \theta, Model\} P\{\theta | Model\}}{P\{Data | Model\}}$$

the trickiest part for calculation is the denominator

- Note that if  $\theta = \theta_i, i = 1, \dots, n$  then

$$P\{Data | Model\} = \sum_{i=1}^n P\{\theta = \theta_i | Model\} P\{Data | \theta = \theta_i, Model\}.$$

This means that the denominator is the sum (integral) of the numerator values for all possible values of the parameter  $\theta$

- Easy to say, but how to calculate it?
- Fortunately this denominator is the same regardless choice of  $\theta$ . This fact is reflected in the common proportionality notation:

$$P\{\theta | Data, Model\} \propto P\{Data | \theta\} P\{\theta, Model\}.$$

- This is the foundation for all modern Bayesian computational methods

# General Framework of Bayesian Analysis II

## Computations in Bayesian analysis

- In early days the use Bayesian analysis was limited by only few possible ways of calculating the posterior distribution with special conjugate pairs of prior and likelihood
- Examples of conjugate distributions: normal-normal, beta-binomial, gamma-Poisson, etc.
- Since technology and new efficient algorithms were developed for sampling of posterior distributions the use of Bayesian approach has expanded
- The basic idea of sampling is:
  - Simulate values of the parameter  $\theta$  from the prior distribution  $F_{\theta}(x) = P\{\theta \leq x\}$
  - Using the sample, calculate the likelihood and the numerator for each simulated  $\theta$ . This makes a sample of the numerator
  - From the sample of the numerator estimate the posterior distribution
  - If new data arrive the posterior distribution is used as a new prior and the whole sequence of steps is repeated again

# Likelihood Function of Binomial distribution I

## Definition

Random variable  $X$  has binomial distribution with parameters size  $s$  and probability of success  $p$  if it can take values  $k = 0, 1, 2, 3, \dots$  and

$$\mathbb{P}\{X = k\} = \binom{s}{k} p^k (1 - p)^{s-k}$$

Binomial distribution describes series of independent Bernoulli experiments with two outcomes ("success", "failure") with corresponding probabilities  $p$  and  $1 - p$ . The size  $s$  is the number of Bernoulli experiments and  $X$  is the number of successes in the series.

When there is only one Bernoulli experiment ( $s = 1$ ) then  $k = 0, 1$  and  $p(k|p) = \mathbb{P}\{X = k\} = p^k (1 - p)^{1-k}$ .

Function  $p(k|p)$  is called **likelihood function of binomial distribution**. It shows the probability (likelihood) of the observed sample given the probability of success.

# Likelihood Function of Binomial Distribution II

- Likelihood  $p(k|p)$  is interpreted in statistics as a function of  $p$ , given the sample  $k$ .
- Even though  $p(k|p)$  is a probability for each  $p, k$ , it is not a probability distribution of  $p$ : for example, given that  $k = 1$

$$\int_0^1 p(1|p) dp = \int_0^1 p dp = \frac{1}{2} \neq 1$$

- In Bayesian context the sample  $k$  is fixed while the parameter  $p$  is a random variable. Following common notations rename  $p$  into  $\theta$ .
- When  $s > 1$  and the sample is  $k = \langle k_1, k_2, \dots, k_s \rangle$ ,  $k_i = \{0, 1\}$ , definition of likelihood function is

$$L(k|\theta) = \prod_{i=1}^s p(k_i|\theta) = \prod_{i=1}^s \theta^{k_i} (1-\theta)^{1-k_i} = \theta^K (1-\theta)^{s-K},$$

where  $K = \sum k_i$  is the number of successes

# Conjugate Distributions

- Recall Bayes theorem

$$P\{\theta | Data, Model\} = \frac{P\{Data | \theta, Model\} P\{\theta | Model\}}{P\{Data | Model\}}$$

or  $P\{\theta | Data, Model\} \propto P\{Data | \theta\} P\{\theta, Model\}$ , where the left-hand side is the posterior and the right-hand side is the product of likelihood and the prior

- It would be convenient to have posterior of the same parametric family as the prior. A few pairs of distributions for data and for parameter have such property, they are called **conjugate distributions**.
- Using conjugate distributions is convenient because posterior distribution replaces prior distribution for a new portion of data
- Using conjugate distributions also allows analytical solution, there is no need in expensive computations
- For conjugate distributions the denominator of Bayes theorem is also obtained by formula



# Choice of Prior Distribution for Binomial Parameter

- For binomial parametric family of data distributions the conjugate parametric family of priors is beta distribution
- Likelihood function is of the form  $L(k|\theta) = \theta^K (1 - \theta)^{s-K}$ . If prior is of the form  $\theta^a (1 - \theta)^b$  then the product is of the same form  $\theta^{K+a} (1 - \theta)^{s-K+b}$ . This gives a hint that conjugate prior distribution should include  $\theta^a (1 - \theta)^b$ .
- Distribution of the form conjugate with binomial is called beta distribution

$$p(\theta|a, b) = \text{Beta}(\theta|a, b) = \frac{\theta^{a-1} (1 - \theta)^{b-1}}{B(a, b)},$$

$$B(a, b) = \int_0^1 \theta^{a-1} (1 - \theta)^{b-1} d\theta$$

- Check interactive demonstration of beta distribution to build some intuition about it

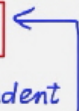
# Beta Distribution as a Prior

Posterior Beta Distribution.

$$p(\theta | \text{data}) = \frac{p(\text{data} | \theta) p(\theta)}{\int_{\theta^*} p(\text{data} | \theta^*)}$$

← Bayes theorem.

$$p(\text{data} | \theta) = \theta^K (1-\theta)^{S-K}$$



$$p(\theta | a, b) = \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1}$$

Likelihood of  $S$  independent Bernoulli trials with  $K$  successes

↑ Beta prior for  $\theta$  with parameters  $a, b$ .

$$p(\theta | K, S) = \frac{p(K, S | \theta) p(\theta | a, b)}{p(K, S)} = \frac{\theta^K (1-\theta)^{S-K} \theta^{a-1} (1-\theta)^{b-1}}{B(a, b) p(K, S)} = \frac{\theta^{(K+a)-1} (1-\theta)^{(S-K+b)-1}}{B(a, b) p(K, S)}$$

$$\int_0^1 \theta^{K+a-1} (1-\theta)^{S-K+b-1} d\theta = B(K+a, S-K+b) \text{ Posterior Beta distribution}$$

$$p(\theta | K, S) = \frac{\theta^{(K+a)-1} (1-\theta)^{(S-K+b)-1}}{B(K+a, S-K+b)} = \text{Beta}(K+a, S-K+b)$$

↓ Denominator

$$p(K, S) = \frac{B(K+a, S-K+b)}{B(a, b)}$$

# Compromise Between Prior and Data

Compromise Between prior and likelihood.

Prior:

$\text{Beta}(a, b)$

Likelihood

$\text{Binom}(s, k)$

$\text{Beta}(K+a, S-K+b)$

posterior

Prior mean:  $\frac{a}{a+b}$

Data:  $\frac{K}{S}$

Posterior mean:  $\frac{K+a}{S+a+b}$

1  $\mu$  between  $\mu$  and  $\frac{K}{S}$

$$\frac{K+a}{S+a+b} = \frac{K}{S+a+b} + \frac{a}{S+a+b} = \frac{K}{S} \frac{S}{S+a+b} + \frac{a}{a+b} \frac{a+b}{S+a+b}$$

2 More data  
- less influence  
of the prior

Posterior  $\mu = \text{Data } q + \text{prior } \mu(1-q)$

3  $S = a + b \Rightarrow q = 1 - q = 0.5$   
Tipping point:  
Concentration  $\lambda = S = a + b$