Lecture Note: Bayesian Statistics

PROBABILITY AND STATISTICS A

Teruo Nakatsuma

Spring Semester 2019

Faculty of Economics, Keio University

Aims Of This Course

1. Learn basic principles of Bayesian learning.
2. Learn how to conduct statistical inference (point estimation, interval estimation, model selection) in the Bayesian way.
3. Learn basic principles of Markov chain Monte Carlo methods.
4. Hands-on practice of Python.

Reading List i

1. Introduction to Bayesian statistics
    - Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC.
    - Greenberg, E. (2013). *Introduction to Bayesian Econometrics*, 2nd ed., Cambridge University Press.
2. Advanced topics in Bayesian statistics
    - Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press.

Reading List ii

- Koop, G., Poirier, D.J. and Tobias, J.L. (2007). *Bayesian Econometric Methods*, Cambridge University Press. *The 2nd edition will be publised in 2019.*
- Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*, Chapman & Hall/CRC. *The 2nd edition will be publised in 2019.*
- Rossi, P.E., Allenby, G.E. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley.

3. PyMC
    - Davidson-Pilon, C. (2016). Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference, Addison-Wesley.

Reading List iii

- Martin, O. (2018). Bayesian Analysis with Python, 2nd ed., Packt Publishing.

4. Markov chain Monte Carlo (MCMC)
    - Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed., Springer.
5. Classics
    - Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer.
    - Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley.

Python

- Python is a high-level programming language.
- Designed by Guido van Rossum
- Released in 1991
- Python is popular.
  `https://spectrum.ieee.org/computing/software/the-2018-top-programming-languages`
  `https://www.tiobe.com/tiobe-index/`

Why Python?

- It is free.
- It is slow in execution but highly manageable.
- Python codes are arguably more readable than other languages such as C/C++.
- Numerous packages have been developed for Python.
- Most of them are free and written in faster programming languages such as C/C++.

How To Obtain Python

- The official Python is downloadable at `https://www.python.org`
- Unfortunately, the plain Python does not include any useful tools for statistics / data science.
- Python distributions for scientific computing
  - **Anaconda**
    `https://www.anaconda.com`
  - **ActivePython**
    `https://www.activestate.com/activepython`
  - **Canopy**
    `https://www.enthought.com/product/canopy`

Tools For Python Programming

- REPL (Read-Eval-Print-Loop)
  - Terminal-based REPL – **IPython**, **QtConsole**
  - Browser-based REPL – **Jupyter Notebook**
    `https://jupyter-notebook.readthedocs.io/en/latest/`
- An **integrated development environment (IDE)** is an application that consists of integrates an editor, a debugger, a profiler and other tools for developers.
  - **Spyder**
    `https://www.spyder-ide.org/`
  - **PyCharm**
    `https://www.jetbrains.com/pycharm/`

Basic Packages

- **NumPy** – n-dimensional array and mathematical functions (`https://www.numpy.org`)
- **SciPy** – functions for scientific computing (`https://www.scipy.org`)
- **Matplotlib** – 2D/3D plotting (`https://matplotlib.org`)
- **Pandas** – data structure (`https://pandas.pydata.org`)

PyMC

PyMC (`https://docs.pymc.io/index.html`) is a Python package for Bayesian MCMC computation. Unlike other tools such as Stan (`https://mc-stan.org`), PyMC is specifically designed for Python and is well integrated with Python and NumPy. So you can write a very *Pythonic* code to perform MCMC computation.

Reference: Salvatier J., Wiecki, T.V. and Fonnesbeck, C. (2016). "Probabilistic Programming in Python Using PyMC3," *PeerJ Computer Science*, 2:e55.

Review Of Probability Theory

Before we proceed to learn Bayesian statistics, let us review the probability theory.

- Probability
- Random Variable
- Probability (Density) Function
- Expectation
- Variance
- Covariance and Correlation

Key Concepts In Probability Theory

Experiment

Suppose researchers conduct a scientific experiment in the laboratory. Their purpose is to gather relevant data with which they confirm or repudiate a hypothesis.

Data

Once a data set is obtained through the experiment, it is regarded as a realization of possible outcomes of the experiment.

Probability

Probability of an event is a number between zero and one that represents a degree of chance that they observe this particular event in the experiment.

Sample Space And Events i

- Let ω denote such a state of the world we are interested in, and Ω denote the set of all conceivable states which is called the sample space.
- When we conduct a scientific study with a series of experiments, we will observe certain outcomes of the experiments.
- Since all states in our world are summarize in the sample space Ω, those outcomes are characterized by a single state $\omega \in \Omega$ or their combination $\{\omega_1, \omega_2, \omega_3, \ldots\}$. We call them events.

Sample Space And Events ii

- Formally speaking, An event is a subset of the sample space $\mathbf{\Omega}$, and will be denoted by uppercase alphabets, e.g., $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\ldots$ in this note.
- The sample space, $\mathbf{\Omega}$, itself can be regarded as the event that at lease one state will be realized.
- As a complement of the sample space, we define the empty event, denoted by $\varnothing$, the one that nothing occurs.
- Since events are mathematically equivalent to subsets of the sample space, we can apply ordinary set operations to them.

Set Operations i

$A \cap B$	intersection, $\{\omega : \omega \in A \text{ and } \omega \in B\}$
	event that both A and B occurs
$A \cup B$	union, $\{\omega : \omega \in A \text{ or } \omega \in B\}$
	event that A and/or B occurs
A^c	complement of A, $\{\omega : \omega \notin A\}$
	event that A does not occurs
$A \setminus B$	difference, $\{\omega : \omega \in A \text{ and } \omega \notin B\} = A \cap B^c$
	A occurs but B does not
$A \subseteq B$	A is a subset of B, $^\forall\omega \in A$, $\omega \in B$
	A occurs, then B occurs
$A = B$	A and B are equivalent, i.e., $A \subseteq B$ and $B \subseteq A$
$A \subset B$	$^\forall\omega \in A$, $\omega \in B$ but $^\exists\omega \in B$ such that $\omega \notin A$

Set Operations ii

The intersection and the union of a sequence of events $\{A_i\}_{i=1}^n$ are defined as follows:

$$\bigcap_{i=1}^{n} A_i = \{\omega \in \Omega \, : \, {}^{\forall} i \in \{1, \dots, n\}, \, \omega \in A_i\},$$
$$\bigcup_{i=1}^{n} A_i = \{\omega \in \Omega \, : \, {}^{\exists} i \in \{1, \dots, n\}, \, \omega \in A_i\}.$$

The famous de Morgan's law

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c, \quad \left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c,$$

is also applicable to events.

Definition Of Probability

A mathematically more rigorous definition of probability is given as follows:

Definition of Probability

Suppose Ω is a sample space.

Axiom 1. For any event $A \subseteq \Omega$, $\mathrm{P}(A) \geqq 0$.

Axiom 2. $\mathrm{P}(\Omega) = 1$.

Axiom 3. For any events $A_1, \ldots, A_n \subseteq \Omega$ such that $A_i \cap A_j = \varnothing$ $(i \neq j)$, we have

$$\mathrm{P}\left(A_1 \cup \cdots \cup A_n\right) = \mathrm{P}(A_1) + \cdots + \mathrm{P}(A_n),$$

where n can be infinite.

Properties

A and B are events, and $\{A_n\}_{n=1}^{\infty}$ is a sequence of events.

1. $\mathrm{P}(A) \leqq \mathrm{P}(B)$ if $A \subseteq B$.
2. $\mathrm{P}(A) \leqq 1$.
3. $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.
4. $\mathrm{P}(\varnothing) = 0$.
5. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
6. $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$ if $A \subseteq B$.
7. $\mathrm{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leqq \sum_{n=1}^{\infty} \mathrm{P}(A_n)$.

Random Variable

Loosely speaking, a random variable (r.v.) is something associated with randomly realized numbers, e.g., a dice, a deck of cards, roulette or lottery. Mathematically, it is a kind of rule or function which matches each outcome in the sample space with a certain number.

Example: Consider an experiment in which a coin is to be tossed 10 times. Then we may define a random variable as

$$\boldsymbol{X(\omega) =} \text{ the number of heads in the sequence for } \boldsymbol{\omega \in \Omega},$$

where $\boldsymbol{\omega}$ stands for a sequence of heads and tails and $\boldsymbol{\Omega}$ is a set of all possible sequences of heads and tails. For example, $\boldsymbol{X(\omega) = 5}$ for the sequence $\boldsymbol{\omega =}$ HTHHTHTHTT. Obviously $\boldsymbol{0 \leqq X(\omega) \leqq 10}$.

Discrete Distribution

It is said that a random variable X has a discrete distribution if X can take only a countable number of different values. The term "countable" means that the number of values is either finite or as many as natural numbers $(1, 2, 3, \ldots)$. Such X is often called a discrete random variable.

If X has a discrete distribution, the probability function or p.f. $f(x)$ is defined as

$$f(x) = \mathrm{P}(X = x).$$

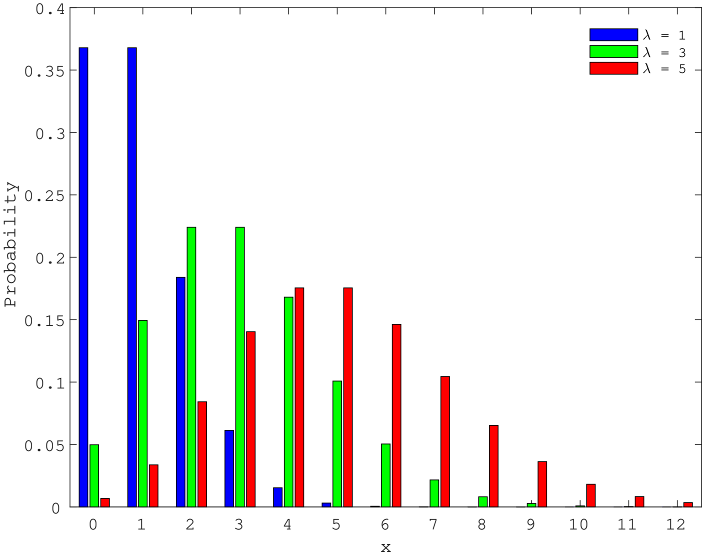

Figure 1: The p.f. of the Poisson distribution

Continuous Distribution

It is said that a random variable X has a continuous distribution if there exists a non-negative function $f(x)$ such that

$$\mathrm{P}(X \in A) = \int_A f(x)dx,$$

where A is any region on $\mathbb{R}$. The function $f(x)$ is called the probability density function or p.d.f. Every p.d.f. must satisfy

1. $f(x) \geqq 0$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

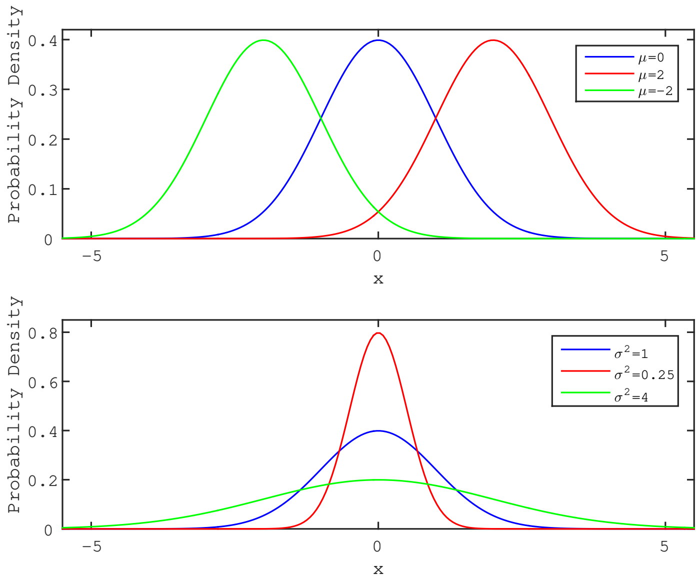

Figure 2: The p.d.f. of the normal distribution

Expectation

The expectation or expected value of a random variable X is defined as

Definition: Expectation of a Random Variable

$$\mathrm{E}[X] = \begin{cases} \sum_{i=1}^{\infty} x_i f(x_i) & \text{for discrete r.v.'s;} \\ \int_{-\infty}^{\infty} x f(x)\,dx & \text{for continuous r.v.'s,} \end{cases}$$

where $f(x)$ is the p.f. or p.d.f. of X. $\mathrm{E}[X]$ is often referred to as the mean of the distribution. This is due to the fact that in the discrete case $\mathrm{E}[X]$ is a weighted average of all possible values that X would take.

Properties

$X, Y, X_1, \ldots, X_n$: random variables
$a, b, c, a_1, \ldots, a_n$: real numbers

1. $\mathrm{E}[X + c] = \mathrm{E}[X] + c$.
2. $\mathrm{E}[aX] = a\mathrm{E}[X]$.
3. $\mathrm{E}[aX + c] = a\mathrm{E}[X] + c$.
4. $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
5. $\mathrm{E}[aX + bY + c] = a\mathrm{E}[X] + b\mathrm{E}[Y] + c$.
6. $\mathrm{E}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} \mathrm{E}[X_i]$.
7. $\mathrm{E}[a_1 X_1 + \cdots + a_1 X_n + c] = \sum_{i=1}^{n} a_i \mathrm{E}[X_i] + c$.

Variance

The variance of a random variable (r.v.) $\boldsymbol{X}$ is defined as

Definition: Variance of a Random Variable

$$\mathrm{Var}[\boldsymbol{X}] = \mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})^2]$$
$$= \begin{cases} \sum_{i=1}^{\infty} (x_i - \mu)^2 f(x_i) & \text{for discrete r.v.'s;} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{for continuous r.v.'s,} \end{cases}$$

where $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{X}]$.

The square root of the variance is called the standard deviation. The variance of a random variable is interpreted as a measurement of spread or dispersion of the distribution around the mean $\boldsymbol{\mu}$.

Properties

1. $\mathrm{Var}[X] = 0$ when $\Pr(X = c) = 1$ for a constant number c.
2. $\mathrm{Var}[X + c] = \mathrm{Var}[X]$.
3. $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$.
4. $\mathrm{Var}[aX + c] = a^2\mathrm{Var}[X]$.
5. $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mu^2$.

Covariance And Correlation

The covariance of two random variables X and Y is

$$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)],$$
$$\mu_X = \mathrm{E}[X], \quad \mu_Y = \mathrm{E}[Y].$$

The correlation (coefficient) of X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}, \quad \sigma_X^2 = \mathrm{Var}[X], \quad \sigma_Y^2 = \mathrm{Var}[Y].$$

Note that $-1 \leqq \rho_{XY} \leqq 1$ is true for any X and Y as long as $\mathrm{Var}[X]$ and $\mathrm{Var}[Y]$ are well defined.

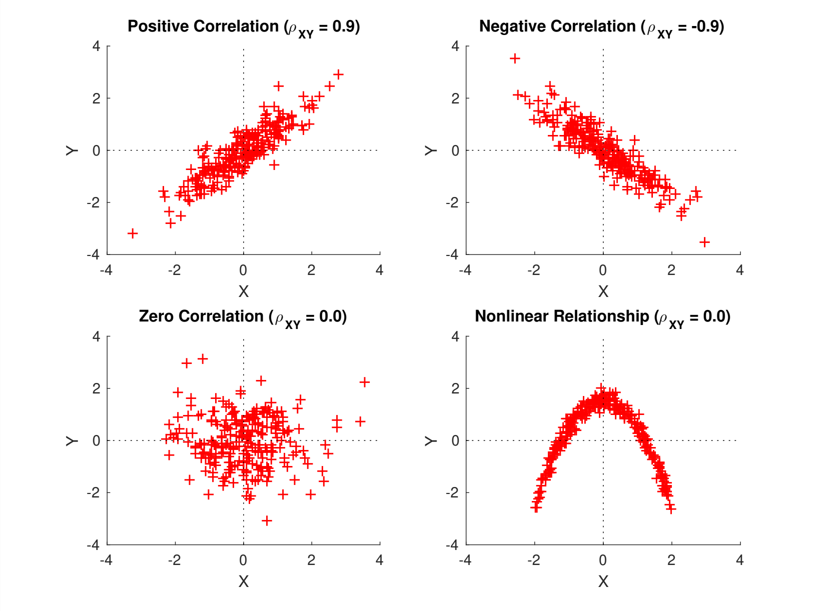

Figure 3: Scatter Plots for Illustration of Correlation

Properties

1. $|\rho_{XY}| \leqq 1$.
2. $\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$.
3. If X and Y are independent, $\mathrm{Cov}[X, Y] = \rho_{XY} = 0$.
4. $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$.
5. $\mathrm{Var}[aX + bY + c] =$
   $a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}[X, Y]$.

Population, Sample, Parameter

1. In statistics, the population is any subject (not necessarily a group) which a researcher try to analyze.
2. The sample is a collection of data related to the population. In a typical situation, the sample is assumed to be randomly and independently extracted from the population.
3. The parameter represents a property of the population to be analyzed. The parameter is unknown to the researcher.
4. The goal of statistics is to obtain useful insights about the parameter of the population with the sample extracted from it.

Population Distribution

We regard the population as a probability distribution and call it the population distribution. Then we can interpret the sample as a set of random variables following the population distribution, and the parameters as variables which determine the "shape" of the population distribution.

Let $\boldsymbol{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ denote the sample, and $\boldsymbol{\theta}$ denote the parameter of the population distribution. To indicate that the shape of the population distribution depends on $\boldsymbol{\theta}$, the population p.f. or p.d.f. is denoted by $\boldsymbol{p}(\boldsymbol{x}_i|\boldsymbol{\theta})$ where each $\boldsymbol{x}_i$ $(\boldsymbol{i} = 1, \ldots, \boldsymbol{n})$ is called an observation and supposed to be a realized value of the random variable following the population distribution. $\boldsymbol{n}$ is often called the sample size.

Likelihood

Suppose the sample $\boldsymbol{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are taken from a population distribution where the parameter $\boldsymbol{\theta}$ is a set of unknown parameters. The joint p.f. or the joint p.d.f of $\boldsymbol{D}$ is denoted by

$$p(D|\theta) = p(x_1, \ldots, x_n|\theta).$$

In particular, if observations are independent of each other,

$$p(D|\theta) = p(x_1|\theta) \times \cdots \times p(x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

When we regard $p(D|\theta)$ as a function of θ, it is called the likelihood or likelihood function.

Example: Bernoulli Distribution i

Let us define a random variable $\boldsymbol{X_i}$ $(\boldsymbol{i = 1, \ldots, n})$ corresponding to tossing a coin such that

$$\boldsymbol{X_i} = \begin{cases} \mathbf{1}, & \text{Head is obtained;} \\ \mathbf{0}, & \text{Tail is obtained,} \end{cases}$$

and

$$\mathbf{Pr}(\boldsymbol{X_i} = 1) = \theta, \quad \mathbf{Pr}(\boldsymbol{X_i} = 0) = 1 - \theta.$$

Then $\boldsymbol{X_i}$ follows the Bernoulli distribution and its p.f. is given by

$$\boldsymbol{p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}}, \quad \boldsymbol{x_i = 0, 1.}$$

Example: Bernoulli Distribution ii

Then the joint p.f. of $D = (x_1, \ldots, x_n)$ is

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{y}(1-\theta)^{n-y}, \quad y = \sum_{i=1}^{n} x_i.$$

Suppose we have $(x_1, x_2, x_3, x_4, x_5) = (1, 0, 1, 1, 1)$. The value of $p(D|\theta)$ depends on the value of θ.

θ	0.1000	0.2000	0.3000	0.4000	0.5000
$p(D\|\theta)$	0.0001	0.0013	0.0057	0.0154	0.0312
θ	0.6000	0.7000	0.8000	0.9000	
$p(D\|\theta)$	0.0518	0.0720	0.0819	0.0656	

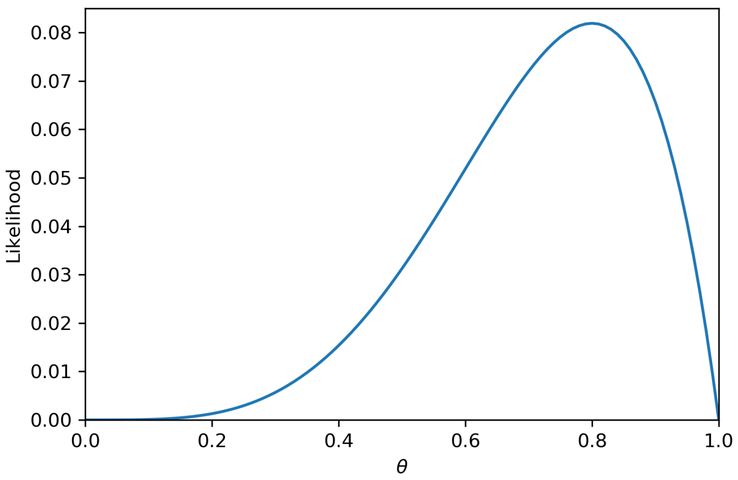

Figure 4: The likelihood of θ in the Bernoulli distribution

Interpretation Of The Likelihood

Given the sample $\boldsymbol{D}$, the likelihood $\boldsymbol{p}(\boldsymbol{D}|\boldsymbol{\theta})$ is regarded as a kind of "plausibility" of a specific value of $\boldsymbol{\theta}$.

For example, the likelihood of $\boldsymbol{\theta} = \mathbf{0.9}$ is $\mathbf{0.656}$ while that of $\boldsymbol{\theta} = \mathbf{0.4}$ in the previous example is $\mathbf{0.0154}$. We may say that $\mathbf{0.9}$ is about 4 times more plausible than $\mathbf{0.4}$ as the true value of $\boldsymbol{\theta}$.

To make comparison between two competing values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, we introduce the likelihood ratio:

$$\text{likelihood ratio} = \frac{\boldsymbol{p}(\boldsymbol{D}|\boldsymbol{\theta_0})}{\boldsymbol{p}(\boldsymbol{D}|\boldsymbol{\theta_1})}.$$

Prior Knowledge On Parameters

In practice, researchers often have information on unknown parameters before they start analysis. For example,

- $\boldsymbol{\theta}$ must take a value between 0 and 1 because it is probability;
- in case of tossing a coin, $\boldsymbol{\theta}$ is supposed to be 50% if the coin is fair.

In Bayesian statistics, we construct a distribution of unknown parameters that reflect our prior knowledge on their true values. This is call the prior distribution. Let $\boldsymbol{p(\theta)}$ denote the prior distribution.

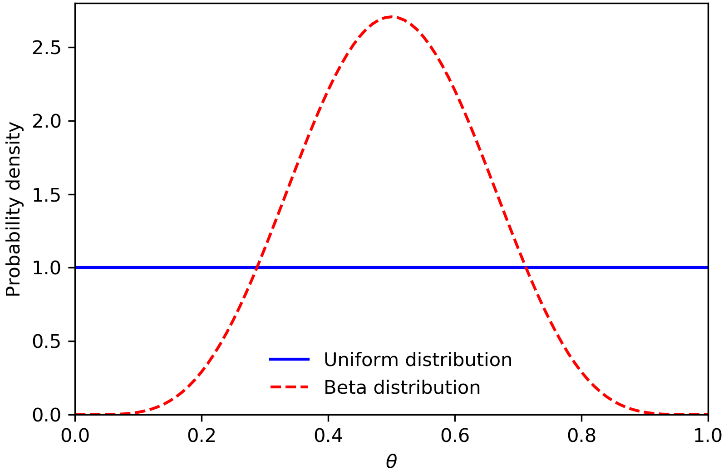

Figure 5: Prior distributions of θ in the Bernoulli distribution

The uniform distribution $\mathrm{Uniform}(a, b)$ is

$$p(x|a,b) = \begin{cases} \frac{1}{b-a}, & (a \leqq x \leqq b); \\ 0, & (\text{otherwise}). \end{cases}$$

In the above figure, we set $a = 0$ and $b = 1$.

The beta distribution $\mathrm{Beta}(\alpha, \beta)$ is

$$p(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ 0 \leqq x \leqq 1.$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

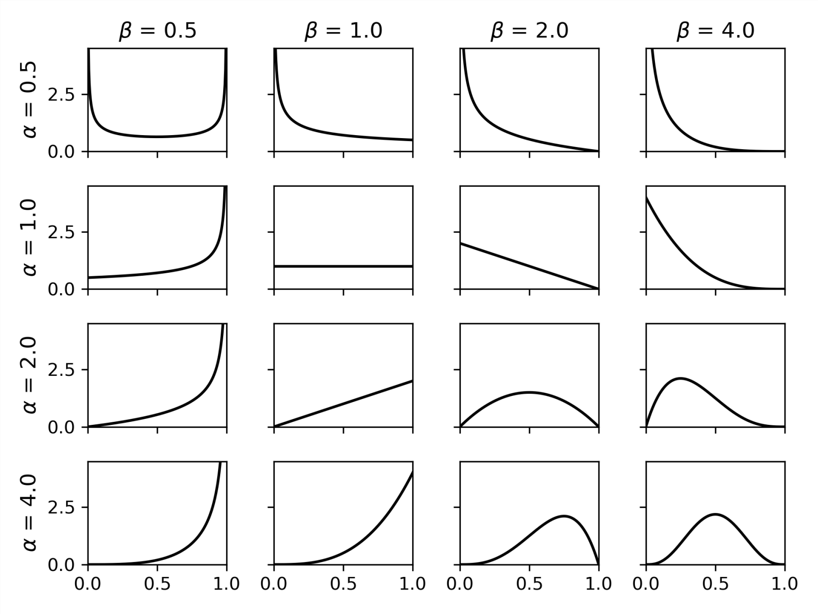

Figure 6: Beta distributions with various (α, β)

Bayes' Theorem And Posterior Distribution

Suppose $p(\theta, D)$ is the joint distribution θ and D. Using the definition of the conditional distribution, we have

$$p(\theta, D) = p(\theta|D)p(D) = p(D|\theta)p(\theta).$$

Arranging the middle and the right-hand side of the above equation, we have

Bayes' Theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.$$

- The above formula is called Bayes' theorem.
- $p(\theta|D)$ is called the posterior distribution.
- $p(D)$ is called the normalizing constant.

Marginal Likelihood

By marginalizing the joint distribution $\boldsymbol{p(\theta, D)}$, $\boldsymbol{p(D)}$ is given as

$$\boldsymbol{p(D) = \int p(\theta, D) d\theta = \int p(D|\theta) p(\theta) d\theta.}$$

Thus $\boldsymbol{p(D)}$ is interpreted as "averaged likelihood" in terms of the prior $\boldsymbol{p(\theta)}$. In this sense, $\boldsymbol{p(D)}$ is called the marginal likelihood.

Then Bayes' theorem is rewritten as

Bayes' Theorem (Alternative Forms)

$$\boldsymbol{p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \propto p(D|\theta)p(\theta).}$$

We can ignore $\boldsymbol{p(D)}$ since it does not depend on $\boldsymbol{\theta}$.

Example: Bernoulli Distribution

Suppose the prior distribution is Beta(α_0, β_0).

The posterior distribution of θ is given by

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) \\ &\propto \theta^{y}(1-\theta)^{n-y} \times \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\ &\propto \theta^{y+\alpha_0-1}(1-\theta)^{n-y+\beta_0-1} \\ &\propto \theta^{\alpha_\star-1}(1-\theta)^{\beta_\star-1}, \\ \alpha_\star = y + \alpha_0, &\quad \beta_\star = n - y + \beta_0. \end{aligned}$$

This is the beta distribution $\textit{Beta}(\alpha_\star, \beta_\star)$.

Bayesian Learning

Bayes' theorem is rearranged as

$$\frac{p(\theta|D)}{p(\theta)} = \frac{p(D|\theta)}{p(D)} \propto p(D|\theta).$$

The left-hand side is

$$\begin{cases} \dfrac{p(\theta|D)}{p(\theta)} > 1, & \text{plausibility of } \theta \text{ is increased;} \\ \\ \dfrac{p(\theta|D)}{p(\theta)} < 1, & \text{plausibility of } \theta \text{ is decreased.} \end{cases}$$

In other words, a Bayesian update in belief is proportional to the likelihood $p(D|\theta)$.

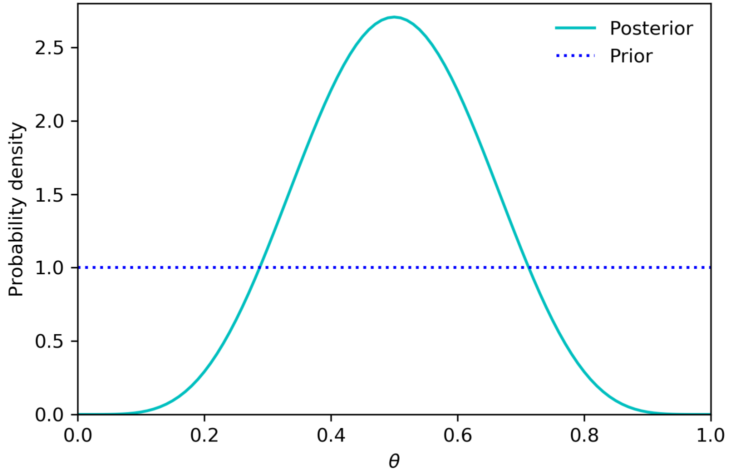

Figure 7: Posterior distributions of the probability of success

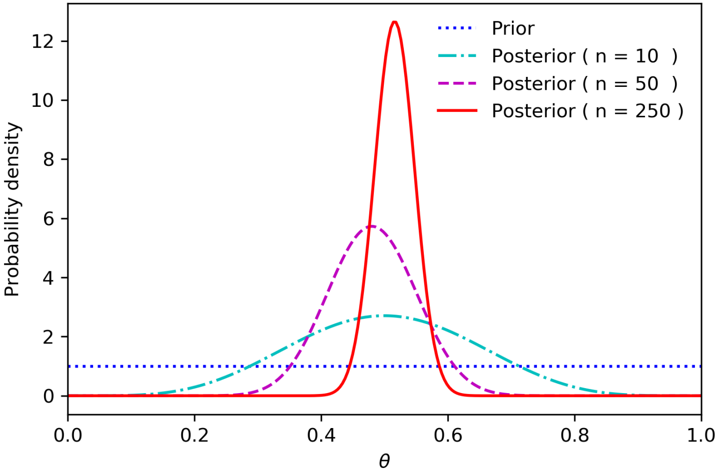

Figure 8: Sequential updating of the posterior distribution