

# 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 published in 2019.*
- Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*, Chapman & Hall/CRC. *The 2nd edition will be published 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.

- 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 (<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  $\omega$  denote such a **state** of the world we are interested in, and  $\Omega$  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  $\Omega$ , those outcomes are characterized by a single state  $\omega \in \Omega$  or their combination  $\{\omega_1, \omega_2, \omega_3, \dots\}$ . We call them **events**.

## Sample Space And Events ii

- Formally speaking, An event is a subset of the sample space  $\Omega$ , and will be denoted by uppercase alphabets, e.g.,  $A$ ,  $B$ ,  $C$ , ... in this note.
- The sample space,  $\Omega$ , itself can be regarded as the event that at least one state will be realized.
- As a complement of the sample space, we define the empty event, denoted by  $\emptyset$ , 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

$A \subset B$   $\forall \omega \in A, \omega \in B$  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  $\Omega$  is a sample space.

**Axiom 1.** For any event  $A \subseteq \Omega$ ,  $P(A) \geq 0$ .

**Axiom 2.**  $P(\Omega) = 1$ .

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

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n),$$

where  $n$  can be infinite.

# Properties

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

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

# 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

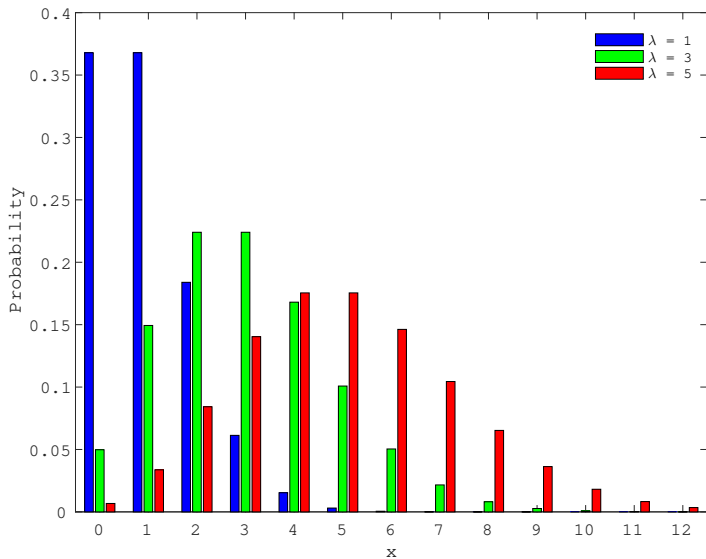
$X(\omega)$  = the number of heads in the sequence for  $\omega \in \Omega$ , where  $\omega$  stands for a sequence of heads and tails and  $\Omega$  is a set of all possible sequences of heads and tails. For example,  $X(\omega) = 5$  for the sequence  $\omega = \text{HTHHTHTHTT}$ . Obviously  $0 \leq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

# 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  $(1, 2, 3, \dots)$ . 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

$$f(x) = P(X = x).$$



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

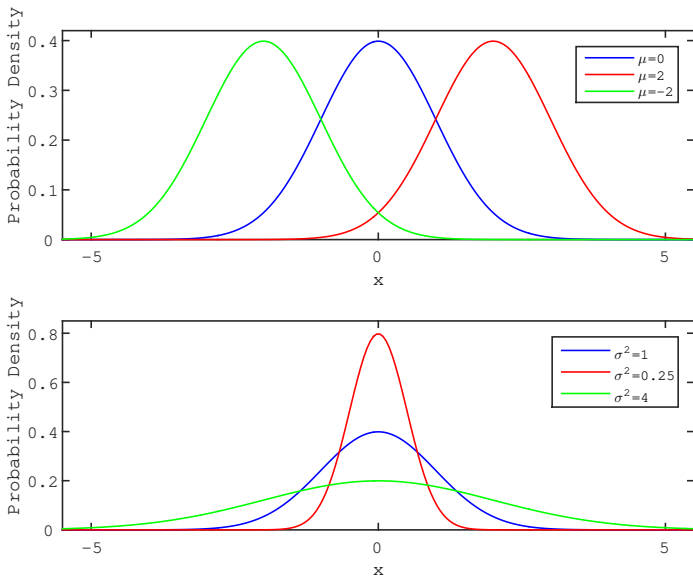
# Continuous Distribution

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

$$\mathbf{P(X \in A)} = \int_{\mathbf{A}} \mathbf{f(x)} d\mathbf{x},$$

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

1.  $\mathbf{f(x)} \geq 0$ .
2.  $\int_{-\infty}^{\infty} \mathbf{f(x)} d\mathbf{x} = 1$ .



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



# Expectation

The **expectation** or **expected value** of a random variable  $\mathbf{X}$  is defined as

## Definition: Expectation of a Random Variable

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

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

# Properties

$X, Y, X_1, \dots, X_n$ : random variables

$a, b, c, a_1, \dots, a_n$ : real numbers

1.  $E[X + c] = E[X] + c.$
2.  $E[aX] = aE[X].$
3.  $E[aX + c] = aE[X] + c.$
4.  $E[X + Y] = E[X] + E[Y].$
5.  $E[aX + bY + c] = aE[X] + bE[Y] + c.$
6.  $E[X_1 + \dots + X_n] = \sum_{i=1}^n E[X_i].$
7.  $E[a_1X_1 + \dots + a_nX_n + c] = \sum_{i=1}^n a_iE[X_i] + c.$

# Variance

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

## Definition: Variance of a Random Variable

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

where  $\mu = \text{E}[\mathbf{X}]$ .

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  $\mu$ .

# Properties

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

# Covariance And Correlation

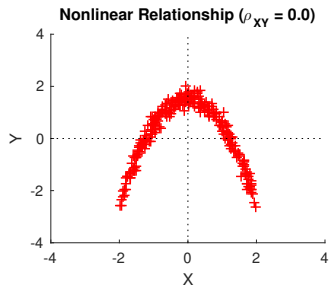
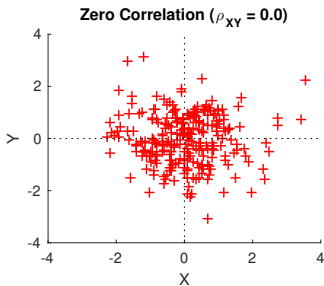
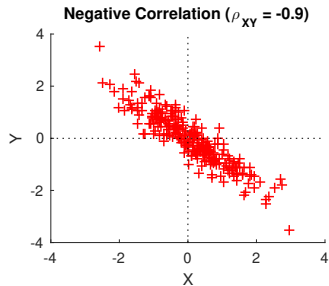
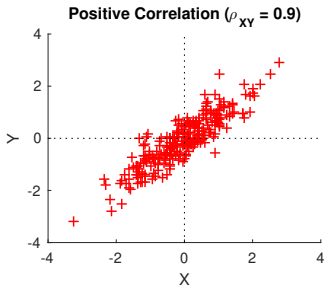
The **covariance** of two random variables  $\mathbf{X}$  and  $\mathbf{Y}$  is

$$\text{Cov}[\mathbf{X}, \mathbf{Y}] = \text{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})],$$
$$\mu_{\mathbf{X}} = \text{E}[\mathbf{X}], \quad \mu_{\mathbf{Y}} = \text{E}[\mathbf{Y}].$$

The **correlation (coefficient)** of  $\mathbf{X}$  and  $\mathbf{Y}$  is

$$\rho_{\mathbf{XY}} = \frac{\text{Cov}[\mathbf{X}, \mathbf{Y}]}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}, \quad \sigma_{\mathbf{X}}^2 = \text{Var}[\mathbf{X}], \quad \sigma_{\mathbf{Y}}^2 = \text{Var}[\mathbf{Y}].$$

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



**Figure 3:** Scatter Plots for Illustration of Correlation

# Properties

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