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, $\{\omega : \omega \in A$ and $\omega \in B\}$ event that both A and B occurs
$A \cup B$	union, $\{\omega : \omega \in A$ 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, $\{\omega : \omega \in A$ 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:

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

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

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

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

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

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

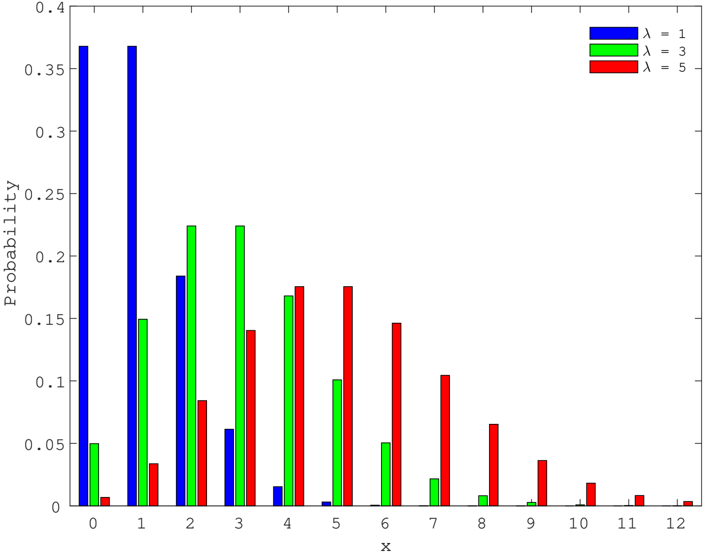

Figure 1: The p.f. of the Poisson distribution

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

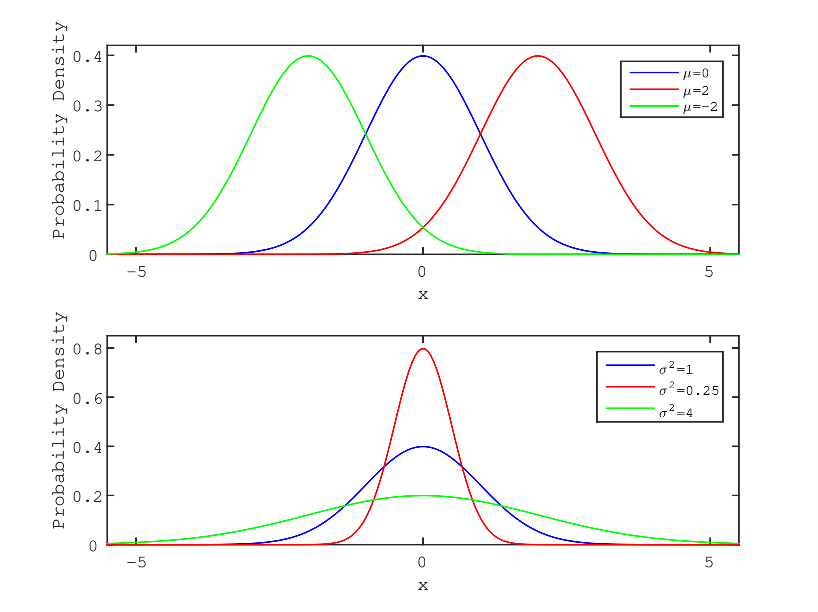

Figure 2: The p.d.f. of the normal distribution

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

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

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

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 μ.

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

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

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.

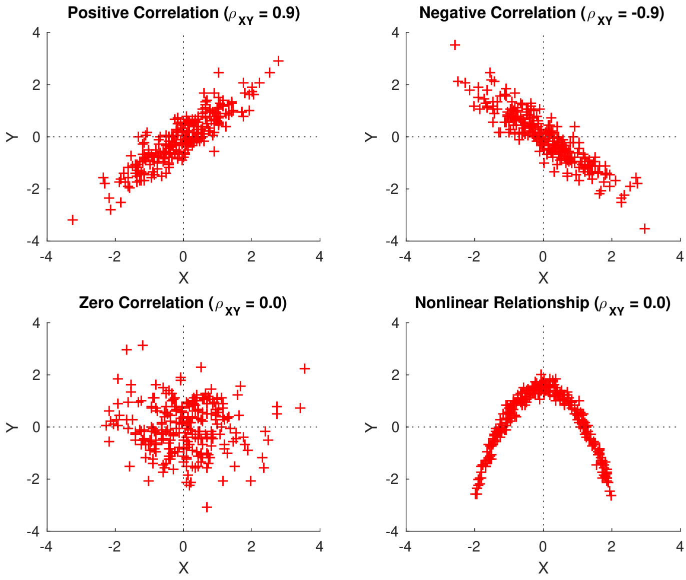

Figure 3: Scatter Plots for Illustration of Correlation

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 $\boldsymbol{D = (x_1, \ldots, x_n)}$ 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 $\boldsymbol{p(x_i|\theta)}$ where each $\boldsymbol{x_i}$ $\boldsymbol{(i = 1, \ldots, n)}$ 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 $\boldsymbol{p(D|\theta)}$ as a function of $\boldsymbol{\theta}$, it is called the likelihood or likelihood function.

Example: Bernoulli Distribution i

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

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

and

$$\mathrm{P}(X_i = 1) = \theta, \quad \mathrm{P}(X_i = 0) = 1 - \theta.$$

Then X_i follows the Bernoulli distribution and its p.f. is given by

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

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	0.6000	0.7000	0.8000	0.9000
$p(D\|\theta)$	0.0001	0.0013	0.0057	0.0154	0.0312	0.0518	0.0720	0.0819	0.0656

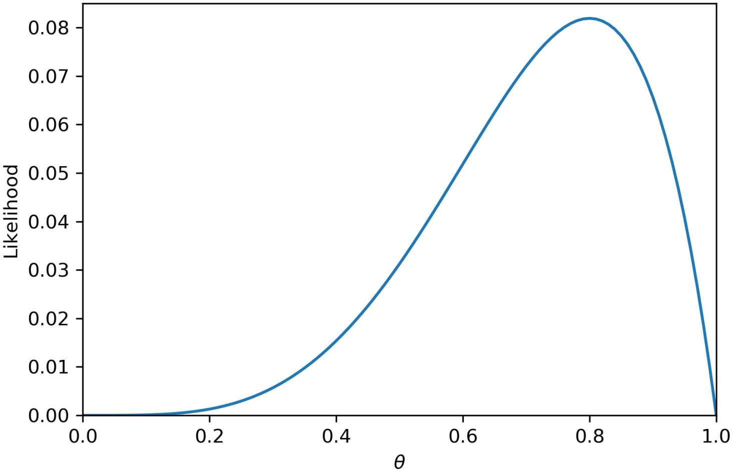

Figure 4: The likelihood of θ in the Bernoulli distribution

Interpretation Of The Likelihood

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

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

To make comparison between two competing values of θ, say θ_0 and θ_1, we introduce the likelihood ratio:

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

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.

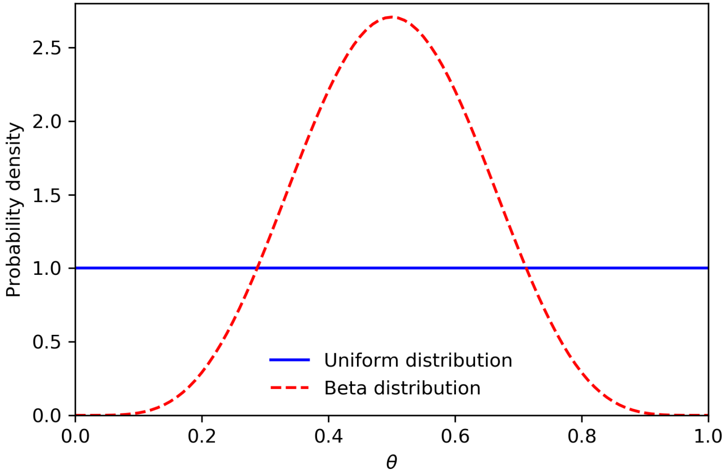

Figure 5: Prior distributions of θ in the Bernoulli distribution

The uniform distribution 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 Beta(α, β) 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:

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

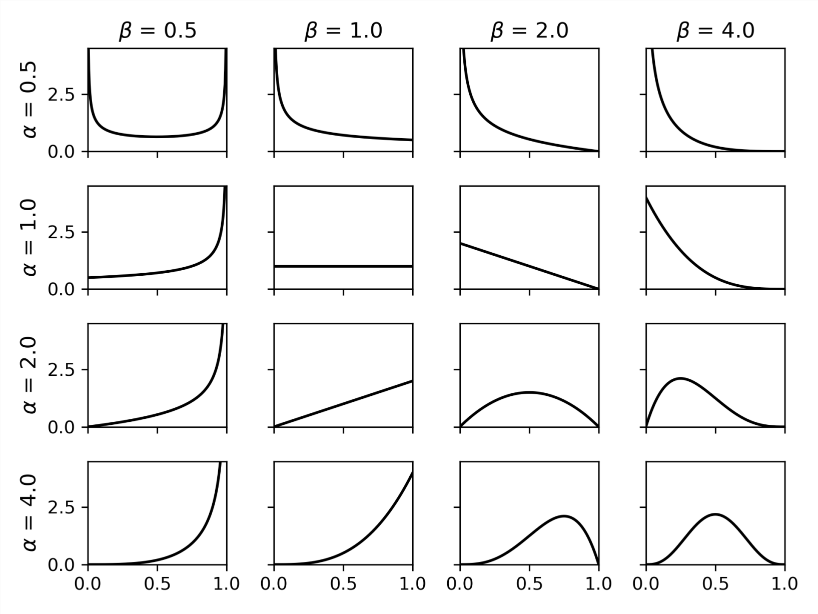

Figure 6: Beta distributions with various (α, β)

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 $\mathit{Beta}(\alpha_\star, \beta_\star)$.

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

$$\begin{cases} \dfrac{p(\theta|D)}{p(\theta)} > 1, & \text{plausibility of } \theta \text{ is increased;} \\[2em] \dfrac{p(\theta|D)}{p(\theta)} < 1, & \text{plausibility of } \theta \text{ is decreased.} \end{cases}$$

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

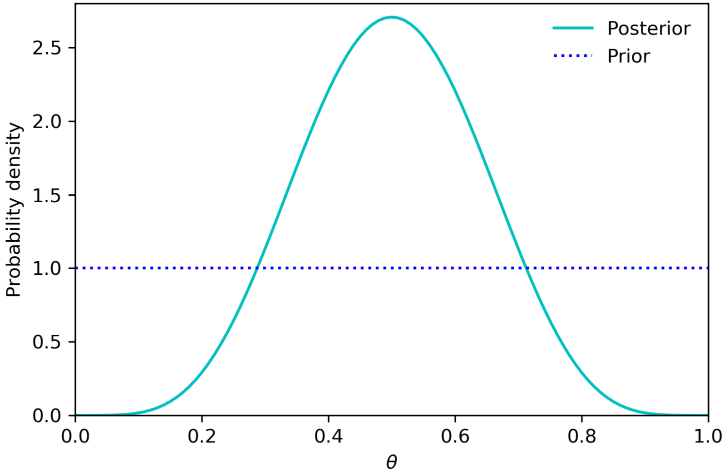

Figure 7: Posterior distributions of the probability of success

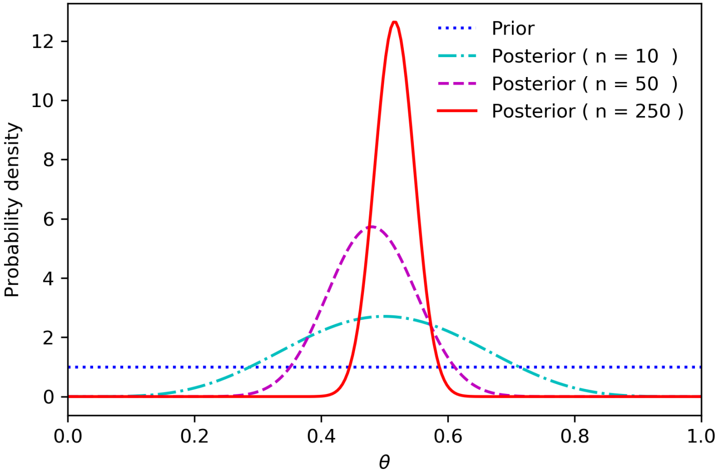

Figure 8: Sequential updating of the posterior distribution

Example: Exit Poll i

On the day of an election, news media often take an exit poll to forecast which candidate/party will win the election. Suppose there are only two candidates on the ballot in this electoral district and one of them must be elected.

One candidate belongs to the ruling party, labeled as R, and the other belongs to the opposition party, labeled as O.

Pollsters ask “Who did you vote for?” to voters who just get out of voting places in this district.

Example: Exit Poll ii

They ask n voters in total and the answer by the i-th voter is recorded as

$$x_i = \begin{cases} 1, & \text{Voted for Candidate R;} \\ 0, & \text{Voted for Candidate O.} \end{cases}$$

Results of this exit poll are collected into the data set $D = (x_1, \ldots, x_n)$.

Suppose R's true vote share is θ ($0 \leqq \theta \leqq 1$).

Since pollsters randomly encounter a voter who voted for R, the probability that the i-th voter did vote for R is θ.

Example: Exit Poll iii

Furthermore $\boldsymbol{x_1, \ldots x_n}$ are supposed to be independent. Therefore each outcome of the exit poll follows a Bernoulli distribution. Hence the likelihood of $\boldsymbol{\theta}$ given $\boldsymbol{D}$ is

$$\begin{aligned} p(D|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{y}(1-\theta)^{n-y}, \quad y = \sum_{i=1}^{n} x_i. \end{aligned}$$

We can derive the posterior distribution of $\boldsymbol{\theta}$ as explained previously.

Example: Exit Poll iv

Alternatively, suppose a researcher is informed that pollsters asked $\boldsymbol{n}$ voters independently and found $\boldsymbol{y}$ voters who voted for R. In this situation, the researcher is not aware of exact composition of $\boldsymbol{D = (x_1, \ldots, x_n)}$ but only knows a summary statistic $\boldsymbol{y}$. Then $\boldsymbol{y}$ follows a binomial distribution:

$$\boldsymbol{p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}},$$

which is also regarded as the likelihood of $\boldsymbol{\theta}$.

The above likelihood is proportional to the one in the previous slide. Thus the posterior distribution of $\boldsymbol{\theta}$ is the same as before.

Likelihood Principle

Likelihood Principle

Information on unknown parameters $\boldsymbol{\theta}$ brought by a sample $\boldsymbol{X}$ is entirely contained in the likelihood $\boldsymbol{p(X|\theta)}$. Furthermore, suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are samples dependent on the same $\boldsymbol{\theta}$ and the likelihood of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are proportional to each other, i.e.,

$$\boldsymbol{p(X|\theta) = \mathcal{K} \cdot p(Y|\theta)} \quad \text{for all } \boldsymbol{\theta} \text{ and some } \boldsymbol{\mathcal{K} > 0}.$$

Then $\boldsymbol{X}$ and $\boldsymbol{Y}$ bring the same information on $\boldsymbol{\theta}$ and the resulting inference should be the same.

This statement is called the likelihood principle and is regarded as the pillar of statistical inference by Bayesians.

Example: Lindley and Phillips (1976)

Here we keep using the exit poll as an example. Consider two pollsters who used different polling methods:

1. Pollster A questioned 12 voters and 9 voters replied that they voted for Candidate R.
2. Pollster B questioned voters until she got 3 voters who voted for Candidate O. She questioned 12 voters to get the result.

Two Likelihoods, One Information

For A, the number of voters who voted for R follows a binomial distribution:

$$p(X_A|\theta) = \binom{12}{9}\theta^9(1-\theta)^3.$$

For B, on the other hand, it follows a negative binomial distribution. Thus we have

$$p(X_B|\theta) = \binom{11}{2}(1-\theta)^3\theta^9.$$

Since $p(X_A|\theta) \propto p(X_B|\theta) \propto \theta^9(1-\theta)^3$, the likelihood principle implies that the inference on θ should be the same.

In fact, the posterior distribution of θ is identical for both pollsters when we apply the same prior distribution.

Paradox On The P-Value

- The probability that the pollster finds 9 or more voters who voted for R, $\mathbf{P}(\boldsymbol{X} \geqq \mathbf{9})$, does depend on how she collects the data.
- Suppose $\boldsymbol{\theta} = \mathbf{1/2}$. For A, $\mathbf{P}(\boldsymbol{X} \geqq \mathbf{9})$ is 7.30%. For B, $\mathbf{P}(\boldsymbol{X} \geqq \mathbf{9})$ is 3.27%.
- If the significance level is set at 5%, B would conclude that R has won while A would conclude the opposite. This is a classic example why the use of the P-value such as $\mathbf{P}(\boldsymbol{X} \geqq \mathbf{9})$ is sometimes misleading.

$\mathbf{P}(\boldsymbol{X} \geqq \boldsymbol{x})$	$\boldsymbol{X_A}$	$\boldsymbol{X_B}$
0	1.0000	1.0000
1	0.9998	0.8750
2	0.9968	0.6875
3	0.9807	0.5000
4	0.9270	0.3438
5	0.8062	0.2266
6	0.6128	0.1445
7	0.3872	0.0898
8	0.1938	0.0547
9	0.0730	0.0327
10	0.0193	0.0193
11	0.0032	0.0112
12	0.0002	0.0065

Bayesian Inference With The Posterior Distribution

The posterior distribution $\boldsymbol{p(\theta|D)}$ embodies all available information about unknown parameter(s), $\boldsymbol{\theta}$. When the number of parameters to be analyzed is relatively small, displaying graphs of all (marginal) posterior distributions may be sufficient to convey useful insights on the parameters to readers.

However, when we need to analyze many parameters, it is impractical and pointless to show all graphs on the parameters in an article or report. In practice, we calculate and report several "summary statistics" that show us key characteristics of the posterior distribution. We call them the posterior statistics.

Point Estimation

On many occasions, we need to report one particular value of the parameter we regard as the most plausible guess. This type of value is called an estimate and a procedure to obtain an estimate is called point estimation.

In Bayesian statistics, an estimate of the parameter is defined as a value that minimize the expected loss.

$$\begin{aligned} \delta_\star &= \arg\max_\delta \mathrm{E}_\theta[L(\theta,\delta)|D] \\ &= \arg\max_\delta \int_\Theta L(\theta,\delta)p(\theta|D)d\theta, \end{aligned}$$

where L is the loss function and Θ is a set of all possible values of θ (parameter space). In case of the Bernoulli probability, $\Theta = \{\theta : \ 0 \leqq \theta \leqq 1\}$.

Examples Of Loss Functions

loss function	$\boldsymbol{L(\theta, \delta)}$	point estimate
quadratic loss	$\boldsymbol{(\theta - \delta)^2}$	posterior mean
absolute loss	$\boldsymbol{\lvert\theta - \delta\rvert}$	posterior median
0–1 loss	$\boldsymbol{1 - 1_\theta(\delta)}$	posterior mode

where $\boldsymbol{1_q(\delta)}$ is the indicator function such that

$$\boldsymbol{1_\theta(\delta) = \begin{cases} 1, & (\delta = \theta), \\ 0, & (\delta \neq \theta). \end{cases}}$$

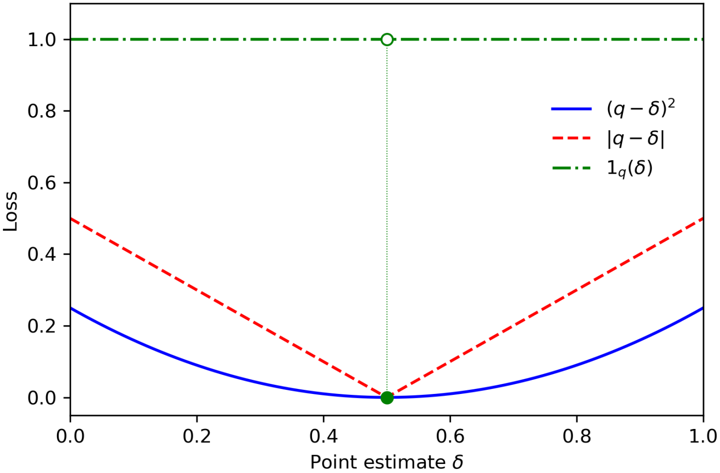

Figure 9: Examples of loss functions

Mean, Median, Mode

The mean of the distribution is the weighted average of all possible values θ may take, i.e.,

$$\mathrm{E}_{\theta}[\theta|D] = \int_{\Theta} \theta p(\theta|D) d\theta.$$

The median of the distribution is a point that divides the distribution in half, i.e.,

$$\mathrm{P}(\theta \leqq \mathrm{Median}_{\theta}|D) = 50\%.$$

The mode of the distribution is the highest point of the density, i.e.,

$$\mathrm{Mode}_{\theta} = \arg\max_{\Theta} p(\theta|D).$$

Remarks On Point Estimation

1. A point estimate (mean, median, mode) is merely a representative point in the posterior distribution. So it is by no means the true value of the parameter in any sense.
2. The mode is not necessarily located in the center of the posterior distribution. You must be careful about using the mode as a point estimate.
3. A Bayesian point estimate is not a random variable.
4. An estimator in the frequentist approach, on the other hand, is supposed to be a random variable.
5. Some researchers use the formula of a Bayesian point estimate as a frequentist estimator. This is often called a Bayes(ian) estimator.

Posterior Probability

The probability that the true value of $\boldsymbol{\theta}$ is within a region in the parameter space, $\boldsymbol{S_0} \subset \boldsymbol{\Theta}$, is given by

$$\mathrm{P}(\theta \in \boldsymbol{S_0}) = \int_{\boldsymbol{S_0}} \boldsymbol{p}(\theta|\boldsymbol{D})\boldsymbol{d}\theta.$$

Such a probability is often called the posterior probability.

When the region is an interval, $\boldsymbol{S_0} = \{\theta : \ \boldsymbol{a} \leqq \theta \leqq \boldsymbol{b}\}$, we have

$$\mathrm{P}(\boldsymbol{a} \leqq \theta \leqq \boldsymbol{b}|\boldsymbol{D}) = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{p}(\theta|\boldsymbol{D})\boldsymbol{d}\theta.$$

Credible Interval (CI)

It is tempting to state that the true value of the parameter must be within an interval with very high posterior probability (say 95%). However, there exist infinitely many intervals with 95% probability because the posterior distribution of the parameter is continuous. Thus we need extra conditions to pin down a unique interval with high posterior probability.

The confidence interval of θ is an interval $[a_c, b_c]$ such that

$$\mathrm{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c,$$
$$\mathrm{P}(\theta < a_c | D) = \frac{c}{2} \quad \text{and} \quad \mathrm{P}(\theta > b_c | D\} = \frac{c}{2}.$$

Set $c = 0.05$ for the 95% CI.

Highest Posterior Density Interval (HPDI)

The highest posterior density interval of θ is an interval $[a_c, b_c]$ such that

1. $\mathrm{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c$,
2. for any pair (θ, θ') such that $\theta \in [a_c, b_c]$ and $\theta' \notin [a_c, b_c]$, $p(\theta|D) > p(\theta'|D)$ must hold.

In particular, if the distribution is unimodal (it has the unique mode), the HPDI must satisfy

$$\begin{aligned} \mathrm{P}(a_c \leqq \theta \leqq b_c | D) = 1 - c, \\ \mathrm{P}(a_c|D) = p(b_c|D). \end{aligned}$$

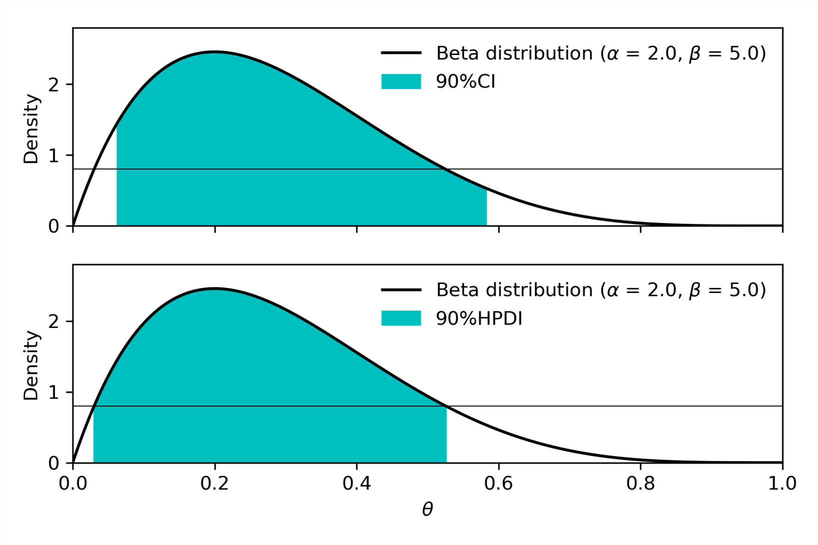

Figure 10: Comparison between CI and HPDI

Remarks On Interval Estimation

1. Both ends of CI or HPDI are not random.
2. The confidence interval in the frequentist approach, on the other hand, is a randomly shifting interval.
3. In Bayesian statistics, the posterior probability is your degree of belief. So you can say “I am 95% certain that the true value of $\boldsymbol{\theta}$ is located in the 95% HPDI.”
4. The 95% confidence interval may capture the true value of the parameter with probability 95%. This probability “95%” is interpreted as the frequency with which the confidence interval will succeed in capturing the true value.

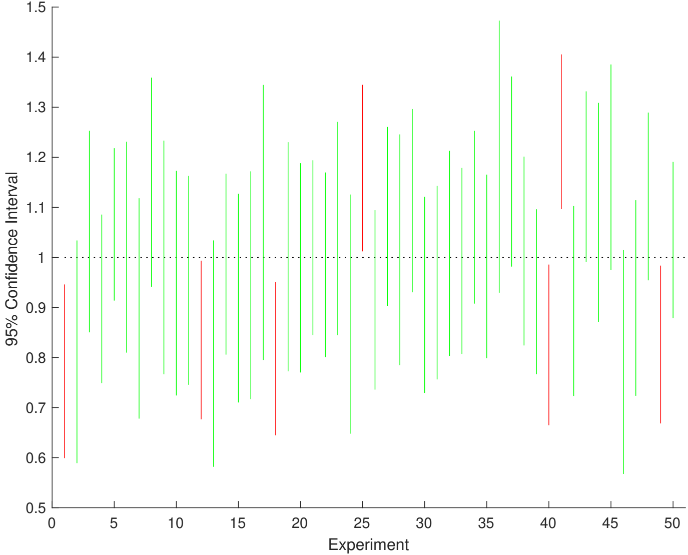

Figure 11: Confidence interval as a randomly shifting interval

Hypotheses On Parameters

In statistics, either Bayesian or frequentist, a hypothesis on the parameter(s) is a region or interval where the true value of the parameter is supposed to be located. For example,

- $\{\theta : 0.5 \leqq \theta \leqq 1\}$,
- $\theta = 0.5$,
- $\theta \neq 0.5$
  $\Leftrightarrow \{\theta : 0 \leqq \theta < 0.5\} \cup \{\theta : 0.5 < \theta \leqq 1\}$.

In general, a hypothesis H_i under which the true value of θ is located in a region $S_i \subset \Theta$ is expressed as

$$H_i : \theta \in S_i, \quad i = 0, 1, 2, \ldots$$

Hypothesis Testing

In Bayesian statistics, plausibility of a hypothesis is measured by the posterior probability that the true value of $\boldsymbol{\theta}$ is located in $\boldsymbol{S_i}$, that is,

$$\mathrm{P}(H_i|D) = \mathrm{P}(\theta \in S_i|D) = \int_{S_i} p(\theta|D)d\theta.$$

Competing hypotheses can be compared by using the posterior odds ration:

$$\text{Posterior odds ratio} = \frac{\mathrm{P}(H_i|D)}{\mathrm{P}(H_j|D)}, \quad i \neq j.$$

Unlike the frequentist approach, Bayesian hypothesis testing does not involves the level of significance, the power and that dreaded P-value!

Bayes Factor

One catch of the posterior odds ratio is that it is affected by the prior information. If the prior information is biased in favor of one hypothesis, the posterior odds ratio is also biased for that hypothesis.

To control the impact of the prior information, the Bayes factor is often used. It is defined as

$$\text{Bayes factor} = \mathrm{B}_{ij} = \frac{\mathrm{P}(H_i|D)}{\mathrm{P}(H_j|D)} \div \frac{\mathrm{P}(H_i)}{\mathrm{P}(H_j)},$$

where $\mathrm{P}(H_i) = \int_{S_i} p(\theta) d\theta$ and $\mathrm{P}(H_i)/\mathrm{P}(H_j)$ is called the prior odds ratio. Note that the Bayes factor is equivalent to the posterior odds ratio if the prior odds ratio is one.

Scale Of Bayes Factor By Jeffreys (1961)

We compare $\boldsymbol{H_i}$ against $\boldsymbol{H_j}$ $(\boldsymbol{i \neq j})$. We suppose $\boldsymbol{H_i}$ is the hypothesis we keep unless we have no strong evidence against it. In the frequentist approach, such a hypothesis is called the null hypothesis. $\boldsymbol{H_j}$, on the other hand, is the hypothesis we want to check whether it is supported by the evidence. In the frequentist approach, it is called the alternative hypothesis.

Rank	Bayes factor B_{ij}	Support for $\boldsymbol{H_j}$
0	$\mathbf{0} < \log_{10}(\mathrm{B}_{ij})$	Rejected
1	$-\frac{1}{2} < \log_{10}(\mathrm{B}_{ij}) < \mathbf{0}$	Barely worth mentioning
2	$-\mathbf{1} < \log_{10}(\mathrm{B}_{ij}) < -\frac{1}{2}$	Substantial
3	$-\frac{3}{2} < \log_{10}(\mathrm{B}_{ij}) < -\mathbf{1}$	Strong
4	$-\mathbf{2} < \log_{10}(\mathrm{B}_{ij}) < -\frac{3}{2}$	Very strong
5	$\log_{10}(\mathrm{B}_{ij}) < -\mathbf{2}$	Decisive

Two-Sided Test

On some occasions, we need to check whether the true value of θ is exactly equal to a particular value, say $\mathbf{0.5}$ (θ must be $\mathbf{0.5}$ if the coin is fair). For this purpose, we need to compare $H_0 : \ \theta = \mathbf{0.5}$ against $H_1 : \ \theta \neq \mathbf{0.5}$.

As a general setup, we consider two competing hypotheses:

$$\begin{cases} H_0 : & \theta = \theta_0, \\ H_1 : & \theta \neq \theta_0. \end{cases}$$

For these hypothesis, however, it is meaningless to construct the Bayes factor because

$$\mathrm{P}(\theta = \theta_0) = \mathrm{P}(\theta = \theta_0 | D) = 0,$$

and both prior and posterior odds ratio are identical to zero.

Spike-And-Slab Prior

To avoid this problem, we introduce a spike-and-slab prior:

$$p(\theta) = p_0\delta(\theta - \theta_0) + (1 - p_0)f(\theta), \quad 0 < p_0 < 1,$$

where $f(\cdot)$ is a continuous distribution of θ and $\delta(\cdot)$ is the Dirac delta function such that

- for any continuous function $g(x)$, $\int_{-\infty}^{\infty} g(x)\delta(x)dx = g(0)$;
- $\int_{-\infty}^{\infty} \delta(x)dx = 1$;
- $\delta(x) = 0$ only if $x \neq 0$.

Savage-Dickey Density Ratio

With the spike-and-slab prior, the prior odds ratio is $\frac{p_0}{1-p_0}$ and the posterior odds ratio is

$$\text{Posterior odds ratio} = \frac{p_0 f(\theta_0|D)}{(1-p_0) f(\theta_0)},$$

where $f(\theta|D)$ is the posterior distribution when $\theta \neq \theta_0$, i.e.,

$$f(\theta|D) = \frac{p(D|\theta) f(\theta)}{\int_\Theta p(D|\theta) f(\theta) d\theta}.$$

Then the Bayes factor is given by

$$\mathrm{B}_{01} = \frac{f(\theta_0|D)}{f(\theta_0)},$$

which is called the Savage-Dickey density ratio (SDDR).

Remarks On Hypothesis Testing i

1. The Bayes factor measures how much our degree of belief on the hypothesis is reinforced by the data.
2. Hypothesis testing in the frequentist approach is based on the Neyman-Pearson lemma. Although this lemma uses the likelihood ratio for hypothesis testing, it treats the likelihood ratio as a random variable (test statistic).
3. In the frequentist approach, we reject the null hypothesis if the test statistic takes a value in the rejection region.
4. We may wrongly reject a correct null hypothesis (type I error) or may wrongly accept a wrong alternative hypothesis (type II error).

Remarks On Hypothesis Testing ii

5. Type I error = false positive;
   Type II error = false negative.
6. We set a ceiling on the probability of type I error, which is called the level of significance, and try to minimize the probability of type II error.
7. The optimal test that can achieve the minimal probability of type II error is called the most powerful test, though it is not always available in practice.

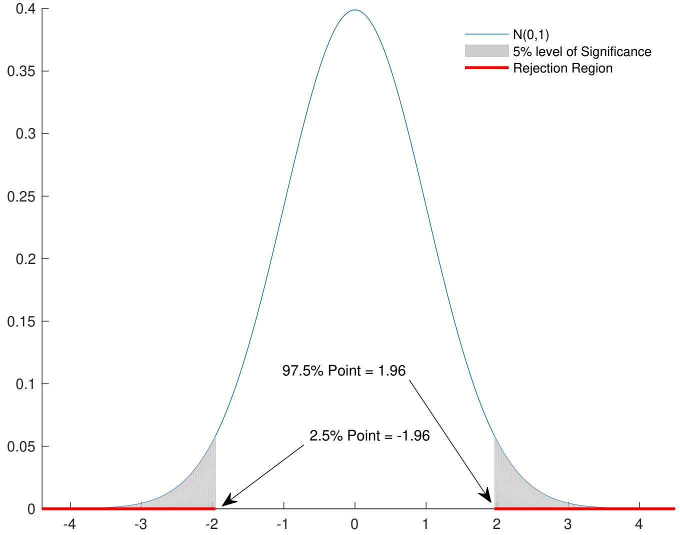

Figure 12: Rejection region of the two-sided Z test

Misunderstanding On The P-Value

1. The P-value is the probability to reject the null hypothesis when
   - the null hypothesis is true,
   - the rejection region is set by the current value of the test statistic.
2. If the P-value is below the level of significance, we reject the null hypothesis. It is a convenient tool to check whether the null hypothesis should be rejected.
3. The P-value is not a degree of trust in either null hypothesis or alternative hypothesis. Therefore we should not regard a lower P-value itself as supportive evidence for the alternative hypothesis.
4. The level of significance (5% or even 1%) is possibly so high that a wrong alternative may be falsely supported.

Predictive Distribution i

Let $\tilde{x}$ denote an unrealized/future value of the population distribution $p(x|\theta)$. Since it is a random variable, we can consider the joint distribution of $\tilde{x}$ and the previous data $D = (x_1, \ldots, x_n)$:

$$p(\tilde{x}, x_1, \ldots, x_n) = p(\tilde{x}, D),$$

Then, from the definition of the conditional probability, we have

$$p(\tilde{x}, D) = p(\tilde{x}|D)p(D) \quad \Rightarrow \quad p(\tilde{x}|D) = \frac{p(\tilde{x}, D)}{p(D)}.$$

Predictive Distribution ii

Furthermore, both $p(D)$ and $p(\tilde{x}, D)$ are regarded as the marginal likelihood given D and $(\tilde{x}, D)$ respectively, that is,

$$p(D) = \int_\Theta p(\tilde{x}, D|\theta)p(\theta)d\theta,$$

$$p(\tilde{x}, D) = \int_\Theta p(\tilde{x}, D|\theta)p(\theta)d\theta.$$

In sum, we have

$$p(\tilde{x}|D) = \frac{\int_\Theta p(\tilde{x}, D|\theta)p(\theta)d\theta}{\int_\Theta p(D|\theta)p(\theta)d\theta}.$$

Predictive Distribution iii

This is called the predictive distribution of $\tilde{\boldsymbol{x}}$. In particular, if $\tilde{\boldsymbol{x}}$ and $\boldsymbol{D}$ are independent, we have

$$p(\tilde{x}, D|\theta) = p(\tilde{x}|\theta)p(D|\theta).$$

Thus the predictive distribution of $\tilde{\boldsymbol{x}}$ is rearranged as

$$\begin{aligned} p(\tilde{x}|D) &= \frac{\int_\Theta p(\tilde{x}|\theta)p(D|\theta)p(\theta)d\theta}{\int_\Theta p(D|\theta)p(\theta)d\theta} \\ &= \int_\Theta p(\tilde{x}|\theta)\frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)d\theta}d\theta \\ &= \int_\Theta p(\tilde{x}|\theta)p(\theta|D)d\theta. \end{aligned}$$

Predictive Distribution (Bernoulli Model) i

Let us derive the predictive distribution for the Bernoulli distribution.

$$
\begin{aligned}
p(\tilde{x}|D) &= \int_{\Theta} p(\tilde{x}|\theta)p(\theta|D)d\theta \\
&= \int_0^1 \theta^{\tilde{x}}(1-\theta)^{1-\tilde{x}} \frac{\theta^{\alpha_\star-1}(1-\theta)^{\beta_\star-1}}{B(\alpha_\star,\beta_\star)} d\theta \\
&= \frac{\int_0^1 \theta^{\tilde{x}+\alpha_\star-1}(1-\theta)^{\beta_\star-\tilde{x}} d\theta}{B(\alpha_\star,\beta_\star)} \\
&= \frac{B(\alpha_\star+\tilde{x},\beta_\star-\tilde{x}+1)}{B(\alpha_\star,\beta_\star)},
\end{aligned}
$$

Predictive Distribution (Bernoulli Model) ii

Using

$$
\begin{aligned}
B(\alpha+1,\beta) &= \frac{\alpha}{\alpha+\beta}B(\alpha,\beta),\\
B(\alpha,\beta+1) &= \frac{\beta}{\alpha+\beta}B(\alpha,\beta),
\end{aligned}
$$

we have

$$
\begin{aligned}
p(\tilde{x}=1|D) &= \frac{B(\alpha_\star+1,\beta_\star)}{B(\alpha_\star,\beta_\star)} = \frac{\alpha_\star}{\alpha_\star+\beta_\star},\\
p(\tilde{x}=0|D) &= \frac{B(\alpha_\star,\beta_\star+1)}{B(\alpha_\star,\beta_\star)} = \frac{\beta_\star}{\alpha_\star+\beta_\star}.
\end{aligned}
$$

Predictive Distribution (Bernoulli Model) iii

Finally

$$\boldsymbol{p(\tilde{x}|D) = \left(\frac{\alpha_\star}{\alpha_\star + \beta_\star}\right)^{\tilde{x}} \left(\frac{\beta_\star}{\alpha_\star + \beta_\star}\right)^{1-\tilde{x}}}.$$

This is the Bernoulli distribution with $\boldsymbol{\theta = \frac{\alpha_\star}{\alpha_\star + \beta_\star}}$.