

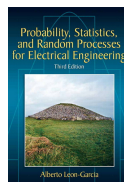
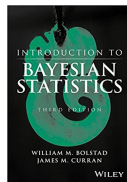
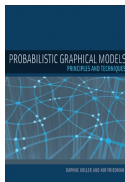
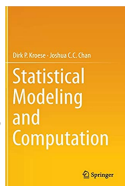
EN007001 Engineering Research Methodology

Statistical Inference: Bayesian Inference

Assoc. Prof. Bhichate Chiewthanakul


Faculty of Engineering, Khon Kaen University
slide:: https://gear.kku.ac.th/~bhichate/Bayesian_inferV5.pdf

- Lecture: 3 hours
- Text:



- Kroese, D. P., & Chan, J. C. (2014). Statistical modeling and computation. New York: Springer.
- Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.
- Bolstad, W. M., & Curran, J. M. (2016). Introduction to Bayesian statistics. John Wiley & Sons.
- Leon-Garcia, A. (2017). Probability, statistics, and random processes for electrical engineering. Pearson Education.

Topic outline

- The conceptual framework for statistical modeling and analysis
- Review of probability \Rightarrow Model for Data by functions
- Statistical inference \Rightarrow Conclusions about Model 
 - Classical Stat.
 - Bayesian Stat.
- Bayesian statistics \Rightarrow Inference about random parameter
- Maximum A Posteriori (MAP) Estimation $\Rightarrow \max_x f_{X|Y}(x|y)$
For more understanding, let us show an example.
- Comparison to ML Estimation $\Rightarrow \max_x l(y_1, \dots, y_n; x)$
- Bayesian Networks \Rightarrow Model for Data by Graph

The conceptual framework for statistical modeling and analysis is sketched in Fig 1

Coin Fairness? $X \sim \text{Bin}(1000, p)$

compute

$p = 0.5?$

$\begin{cases} \text{fair,} & p = 0.5 \\ \text{unfair,} & \text{otherwise} \end{cases}$

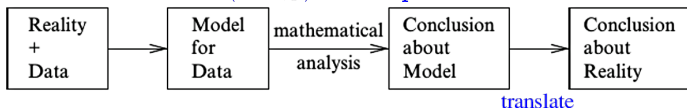


Fig 1. Statistical modeling and analysis

- Data is used to represent the real-life problem
- Probabilistic model for data : Model represent a reality
- Using the model to carry out our calculations and analysis
- Conclusions about the model
- Conclusions about the model are translated into conclusions about the reality.

Definition 1 (Random Variable)

A *random variable* is a function from the sample space Ω to \mathbb{R} .
i.e.

$$X : \Omega \rightarrow \mathbb{R}$$

Symbol \mapsto real

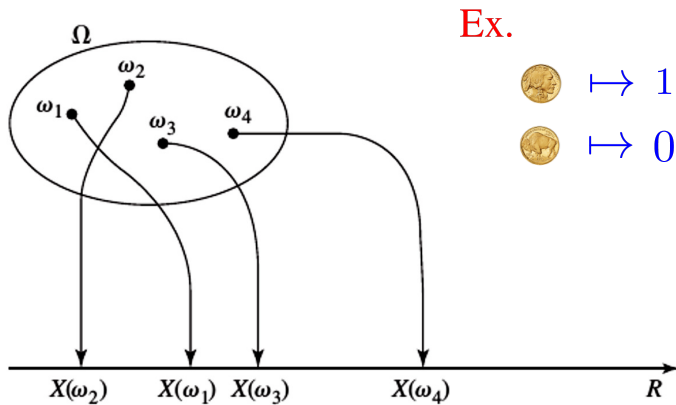


Fig 2. Random variable as a mapping from Ω to \mathbb{R} .

The behavior of the Reality can be model by cdf, pmf/pdf, conditional pmf/pdf and joint cdf/pmf/pdf as follows:

Definition 2 (Cumulative Distribution Function)

The *cumulative distribution function (cdf)* of a random variable X is the function $F : \mathbb{R} \rightarrow [0, 1]$ defined by

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

↑
fixed

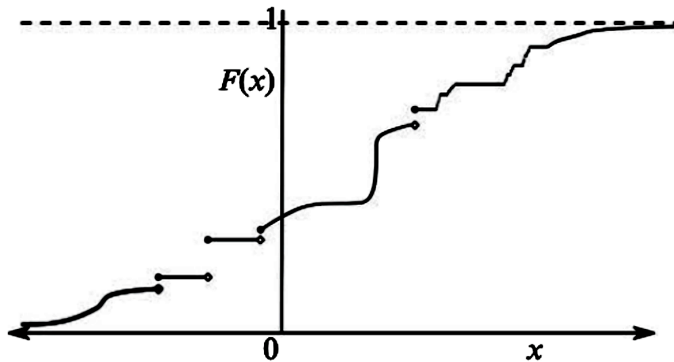


Fig 3. A cumulative distribution function (cdf)

Definition 3 (Discrete Random Variable)

A *discrete random variable* X is defined as a random variable that assume values from a countable set.

Definition 4 (Probability Mass Function)

The *probability mass function (pmf)* of a discrete random variable X is defined as:

$$p_X(x) = P(X = x) = P(\{\zeta : X(\zeta) = x\}), \quad x \in \mathbb{R}$$

Definition 5 (Conditional Probability Mass Function)

Let X be a discrete random variable with pmf $p_X(x)$, and let C be an event that has nonzero probability, $P(C) > 0$.

The conditional probability mass function of X is defined by the conditional probability:

$$p_X(\overset{\text{second}}{\downarrow} x \mid \overset{\text{first}}{\downarrow} C) = \frac{P(\{X = x\} \cap C)}{P(C)}$$

See: **conditional event:** $A|B$

Definition 6 (Continuous Random Variable)

A *continuous random variable* is defined as a random variable whose cdf $F_X(x)$ is continuous everywhere and it can be written as an integral of some nonnegative function $f(x)$:

$$F_X(x) = \int_{-\infty}^x f(t)dt,$$

where

$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

Definition 7 (Probability Density Function)

The *probability density function* of X (pdf), if it exist, is defined as the derivative of cdf $F_X(x)$:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Definition 8 (Joint Cumulative Distribution Function)

The *joint cumulative distribution function* of X and Y is defined by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

Definition 9 (Jointly continuous random variables)

Two random variables X and Y are *jointly continuous* with *joint density* $f_{X,Y}(x,y)$ if

$$P((X,Y) \in A) = \iint_A f_{X,Y}(x,y) dx dy,$$

where

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

Definition 10 (Condition pdf)

The *conditional pdf* of X given C is defined by

$$f_X(x | C) = \frac{dF_X(x | C)}{dx}$$

Fact 1

If X and Y are independent then

- the joint cdf is

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$$

- the joint pdf is

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

Important Random Variables. The most commonly used random variables in communications are:

Bernoulli Random Variable.

$$p_X(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

It is a good model for a binary data generator and a channel errors.

Binomial Random Variable. It is a r.v. giving the number of 1's

Let $\underbrace{X_1, \dots, X_n}_{\text{iid Bernoulli RVs.}} \Rightarrow$ in a sequence of n independent Bernoulli trials.

$$P\left(\sum_{i=1}^n X_i = k\right) : P(X = k) = \begin{cases} p^k(1-p)^{n-k}, & 0 \leq k \leq n \\ 0, & \text{otherwise} \end{cases}$$

<https://www.geogebra.org/m/EmPmvCTc>

This r.v. model, for example, the total number of bits received in error when a sequence of n bits is transmitted over a channel with bit-error probability of p .

Uniform Random Variable. The pdf is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

<https://www.geogebra.org/m/EmPmvCTc>

This r.v. model, for example, when the phase of a sinusoid is random it is usually modeled as a uniform random variable between 0 and 2π .

The most important distribution in the study of statistics: the normal (or Gaussian) distribution.

Definition 11 (Normal Distribution)

A random variable X is said to have a *normal distribution* with parameters μ and σ^2 , $N(\mu, \sigma^2)$ if its pdf is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}$$

If $\mu = 0$ and $\sigma^2 = 1$ or $N(0, 1)$, then

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is known as the *standard normal distribution*.

<https://www.geogebra.org/m/fXww9z9S>

Example 1

Show that

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-(2x^2 - 2xy + y^2)/2}$$

is a valid joint probability density. : see def 09 p.13

Solution. Since $f_{X,Y}(x,y) > 0$, all we have to do is show that it integrates to one. By factor the exponent, we obtain

$$f_{X,Y}(x,y) = \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}} \cdot \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Then

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = \int_{-\infty}^{\infty} \underbrace{\frac{e^{-x^2/2}}{\sqrt{2\pi}}}_{N(0,1)} \left(\int_{-\infty}^{\infty} \underbrace{\frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}}_{N(x,1)} dy \right) dx = 1$$

#

Sagemath Program 1

```
#Proof
reset()
x,y=var("x,y")
f(x,y)=1/2/pi*e^( -(2*x^2-2*x*y+y^2)/2 )

#Proof f(x,y)>0
print bool(f(x,y)>0)

#Proof volume below the surface is equal to 1
print integral(integral(f,(x,-oo,oo)),y,-oo,oo)
```

Julia Program 1

```
#Proof valid of the joint pdf
using PyCall
@pyimport sympy as sm
x,y=sm.symbols("x y")
oo=sm.oo

f=sm.Function("f")
f=1/2/(sm.pi)*sm.exp(-(2*x^2-2x*y+y^2)/2)
sm.integrate(f, (x,-oo,oo), (y,-oo,oo))
```


Julia Program 1 (cont.)

```
In [136]: using PyCall
          @pyimport sympy as sm
          x,y=sm.symbols("x y")
          oo=sm.oo
```

Out[136]: PyObject oo

```
In [139]: f=sm.Function("f")
          f=1/2/(sm.pi)*sm.exp(-(2*x^2-2x*y+y^2)/2)
          sm.integrate(f, (x, -oo, oo),(y,-oo,oo))
```

Out[139]: PyObject 1.0000000000000000

Definition 12 (Random Vector and Random Matrix)

A vector \mathbf{X} whose entries are random variables is called *a random vector*, and a matrix \mathbf{Y} whose entries are random variables is called *a random matrix*. i.e.

$$\mathbf{X} = [X_1, X_2, \dots, X_n]^T,$$
$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1p} \\ Y_{21} & Y_{22} & \dots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{np} \end{pmatrix}.$$

Definition 13 (Statistical inference)

Statistical inference is a collection of methods that deal with drawing conclusions about the model on the basis of the observed data. (See fig 1)

The two main approaches to statistical inference are:

- Classical statistics.
- Bayesian statistics.

Definition 14 (Classical statistics)

Let \mathbf{x} be outcome of a random vector \mathbf{X} described by a probabilistic model that depend on unknown parameter θ . Let θ is assumed to be fixed. Then *classical statistic* is the method for estimating and for drawing inferences about a parameter θ .

The model is specified up to a (multidimensional) parameter θ ; that is, $X \sim f(\cdot; \theta)$.

Definition 15 (Bayesian statistics)

Let \mathbf{x} be outcome of a random vector \mathbf{X} described by a probabilistic model that depend on unknown parameter θ . Let θ is assumed to be random. Then *Bayesian statistic* is the method for estimating and for drawing inferences about a parameter θ , such that θ is carried out by analyzing the conditional pdf $f(\theta | \mathbf{x})$.

$f(\theta | \mathbf{x})$ is called *posterior pdf* of the parameter θ .

Example 2 (Bias coin)

We throw a coin 1000 times and observe 570 Heads. Using this information, what can we say about the “fairness” of the coin? The data (or better, datum) here is the number $x = 570$. Suppose we view x as the outcome of a random variable X which describes the number of Heads in 1000 tosses. Our statistical model (See Fig 1) is then

$$X \sim \text{Bin}(1000, p)$$

where $p \in [0, 1]$ is unknown.

Remark 1

- *Any statement about the fairness of the coin is expressed in terms of p and is assessed via this model.*
- *It is important to understand that p will never be known.*
- *A common-sense estimate of p is simply the proportion of Heads, $x/1000 = 0.570$.*

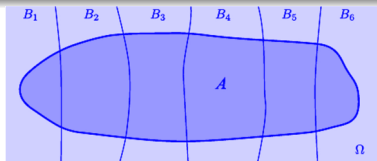
How accurate is this estimate? Is it possible that the unknown p could in fact be 0.5? One can make sense of these questions through detailed analysis of the statistical model. (i.e. by using Classical statistics)

Bayesian statistics is a branch of statistics that is centered around Bayes' formula.

Theorem 1 (Bayes' Rule)

Let A be an event with $P(A) > 0$ and let B_1, B_2, \dots, B_n be a partition of Ω . Then,

$$P(B_j | A) = \frac{P(A | B_j)P(B_j)}{\sum_{i=1}^n P(A | B_i)P(B_i)}. \quad (1)$$



$$\Leftrightarrow \bigcup_{i=1}^n B_i = \Omega, B_i \cap B_j = \emptyset \text{ if } i \neq j$$

Corollary 1.1

For continuous random variables X and Y , Bayes' Theorem is formulated in terms of densities:

$$\underbrace{f(y | x)}_{\text{posterior}} = \frac{f(x | y)f(y)}{\underbrace{\int f(x | y)f(y)dy}_{\text{ind. of } Y}} \propto \underbrace{f(x | y)}_{\text{likelihood}} \cdot \underbrace{f(y)}_{\text{prior}}, \quad (2)$$

where

$f(y) := f_Y(y)$, $f(x | y) := f_{X|Y}(x | y)$, $f(y | x) := f_{Y|X}(y | x)$
and likelihood function, $l(x | y) = f(x | y)$.

Definition 16 (Prior, Likelihood, and Posterior)

Let \mathbf{x} and θ denote the data and parameters in a Bayesian statistical model.

- The pdf of θ , $f(\theta)$ is called the *prior* pdf.
- The conditional pdf $f(\mathbf{x} \mid \theta)$ is called the Bayesian *likelihood* function.
- The central object of interest is the *posterior* pdf $f(\theta \mid \mathbf{x})$ which, by Bayes theorem, is proportional to the product of the prior and likelihood:

$$f(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)f(\theta).$$

$$\text{From (2) } \overbrace{f(y | x)}^{\text{posterior}} \propto f(x | y) \cdot \overbrace{f(y)}^{\text{prior}}$$

Remark 2 (Bayesian Statistical Inference)

The goal is to draw inferences about an unknown variable Y by observing a related random variable X . The unknown variable is modeled as a random variable Y , with prior distribution

$f_Y(y)$, if Y is continuous,

$P_Y(y)$, if Y is discrete.

From (2) $\overbrace{f(y | x)}^{\text{posterior}} \propto f(x | y) \cdot \overbrace{f(y)}^{\text{prior}}$

Remark 2 (Cont.)

After observing the value of the random variable X , we find the posterior distribution of Y . This is the conditional PDF (or PMF) of Y given $X = x$,

$$f_{Y|X}(y | x) \text{ or } P_{Y|X}(y | x).$$

<https://www.geogebra.org/m/EmPmvCTc>

Fact 2

- If $X \sim \text{Uniform}[a, b]$, then

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

- If $X \sim \text{Geometric}(p)$, then

$$P_X(x) = (1-p)^{x-1} \cdot p, \quad x = 1, 2, 3, \dots$$



Trial iid Bernoulli RVs until outcome is one = $0, 0, \dots, 0, 1$
 $\underbrace{\hspace{1.5cm}}_{x-1}$

Fact 3 (Law of Total Probability)

Let A be an event and let B_1, B_2, \dots, B_n be a partition of Ω .
Then,

$$P(A) = \sum_{i=1}^n P(A \mid B_i)P(B_i).$$

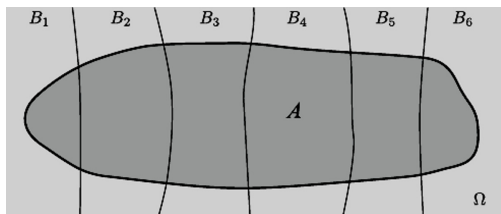


Fig 4. A partition B_1, \dots, B_6 of the sample space Ω

Example 3

cont.RV.

Let $X \sim \text{Uniform}(0, 1)$. Suppose that we know

discrete

$(Y|X = x) \sim \text{Geometric}(x)$. Find the posterior density of X given $Y = 2, f_{X|Y}(x|2)$.

Solution. Using Bayes'rule we have

$$f_{X|Y}(x | 2) = \frac{\overset{\text{Geo}(p)}{\underbrace{P(Y=y)}_{p \leftarrow x}} \overset{1}{P_{Y|X}(2 | x)} f_X(x)}{\overset{2}{P_Y(2)}}.$$

¹ Since $(Y|X = x) \sim \text{Geometric}(x)$. We obtain

$$P_{Y|X}(y | x) = x(1 - x)^{y-1}, \quad y = 1, 2, \dots$$

and

$$P_{Y|X}(2 | x) = x(1 - x).$$

Example 3 (cont.)

To find $P_Y(2)$, we can use the law of total probability

$$\begin{aligned} P_Y(2) &= \int_{-\infty}^{\infty} P_{Y|X}(2 | x) f_X(x) dx \\ &= \int_0^1 x(1 - x) \cdot 1 dx \\ &= \frac{1}{6}. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} f_{X|Y}(x | 2) &= \frac{x(1 - x) \cdot 1}{\frac{1}{6}} \\ &= 6x(1 - x), \quad \text{for } 0 \leq x \leq 1. \end{aligned}$$

#

The posterior distribution, $f_{X|Y}(x | y)$ (or $P_{X|Y}(x | y)$), contains all the knowledge about the unknown quantity X . Therefore, we can use the posterior distribution to find point or interval estimates of X .

Definition 17 (MAP)

Let $f_{X|Y}(x | y)$ be a posterior distribution. Then the *Maximum A Posteriori (MAP) Estimation* is defined as

$$\hat{x}_{MAP} = \max_x f_{X|Y}(x | y) \quad (3)$$

Since $f_{X|Y}(x | y) = \frac{f_{Y|X}(y|x)f_X(x)}{\underbrace{f_Y(y)}_{\text{ind } x}}$ and $f_Y(y)$ does not depend on x . Therefore,

$$\hat{x}_{MAP} = \max_x f_{Y|X}(y | x)f_X(x) \quad (4)$$

Find \hat{x}_{MAP}

To find the MAP estimate of X given that we have observed $Y = y$, we find the value of x that maximizes

$$f_{Y|X}(y | x)f_X(x)$$

If either X or Y is discrete, we replace its PDF in the above expression by the corresponding PMF.

Example 4

Let X be a continuous random variable with the following PDF:

$$f_X(x) = \begin{cases} 2x, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Also, suppose that $(Y | X = x) \sim \text{Geometric}(x)$. Find the MAP estimate of X given $Y = 3$. $= \max_x f_{Y|X}(y | x) \cdot f_X(x)$ (4)

Solution. Since $(Y | X = x) \sim \text{Geometric}(x)$, Then

$$P_{Y|X}(y | x) = x(1 - x)^{y-1}, \quad \text{for } y = 1, 2, \dots$$

Example 4 (cont.)

For $Y = 3$, it follows that

$$P_{Y|X}(3 | x) = x(1 - x)^2.$$

For $x \in [0, 1]$ one has,

$$P_{Y|X}(3 | x)f_X(x) = x(1 - x)^2 \cdot 2x \quad (5)$$

To find the value of x that maximizes Eq. (5), one need

Example 4 (cont.)

$$\frac{d}{dx}(x^2(1-x)^2) = 2x(1-x)^2 - 2(1-x)x^2 = 0. \quad (6)$$

Solve for x , one obtain

$$\hat{x}_{MAP} = \frac{1}{2}.$$

#

Julia Program 2 (Find solution of (6) and \hat{x}_{MAP})

```
In [4]: using Pycall  
        @pyimport SymPy as sm
```

```
In [13]: x=sm.symbols("x")  
         f(x)=x^2*(1-x)^2  
         eq=sm.diff(f(x),x)  
         sol=sm.solve(eq,x)
```

```
Out[13]: 3-element Array{Pycall.PyObject,1}:  
          PyObject 0  
          PyObject 1/2  
          PyObject 1
```

```
In [12]: map(x->f(x),sol)
```

```
Out[12]: 3-element Array{Pycall.PyObject,1}:  
          PyObject 0  
          PyObject 1/16  
          PyObject 0
```

We now consider a classical statistic inference called maximum likelihood method for finding a point estimator that maximizes the probability of the observed data $\mathbf{Y}_n = (Y_1, Y_2, \dots, Y_n)$.

Definition 18 (Likelihood function)

Let $\mathbf{Y}_n = (y_1, y_2, \dots, y_n)$ be the observed values of a random sample for the random variable Y and let X be the parameter of interest. Then *likelihood function* of the sample is a function of X defined as follows:

$$\begin{aligned} l(\mathbf{y}_n; x) &= l(y_1, y_2, \dots, y_n; x) \\ &= \begin{cases} P_{Y|X}(y_1, y_2, \dots, y_n | x), & Y \text{ discrete r.v.} \\ f_{Y|X}(y_1, y_2, \dots, y_n | x), & Y \text{ cont. r.v.,} \end{cases} \quad (7) \end{aligned}$$

Definition 18 (cont.)

where $P_{Y|X}(y_1, y_2, \dots, y_n | x)$ and $f_{Y|X}(y_1, y_2, \dots, y_n | x)$ are the joint pmf and joint pdf evaluated at the observation values if the parameter value is x .

Definition 19 (MLE: maximum likelihood estimator)

Let $\mathbf{Y}_n = (y_1, y_2, \dots, y_n)$ be the observed values of a random sample for the random variable Y and let x be the parameter of interest. Then *maximum likelihood estimator* of x , denoted by \hat{x}_{ML} is the parameter value that maximizes the likelihood function, that is,

$$l(y_1, y_2, \dots, y_n; \hat{x}_{ML}) = \max_x l(y_1, y_2, \dots, y_n; x)$$

Remark 3

*The maximum likelihood estimate (MLE), answers the question:
“For which parameter value of x does the observed data
 (y_1, y_2, \dots, y_n) have the biggest probability?”*

Example 5

Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency θ in the population. That is, a random copy of the gene is A with probability θ and a with probability $1 - \theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

genotype	AA	Aa	aa
probability	θ^2	$2\theta(1 - \theta)$	$(1 - \theta)^2$

Suppose we test a random sample of people and find that k_1 are AA , k_2 are Aa , and k_3 are aa . Find the MLE of θ .

Example 5 (cont.)

Solution. The likelihood function is given by

$$P(k_1, k_2, k_3 \mid \theta) = \overbrace{\binom{k_1 + k_2 + k_3}{k_1} \binom{k_2 + k_3}{k_2} \binom{k_3}{k_3}}^{\text{constant}} \times \quad (8)$$
$$\theta^{2k_1} (2\theta(1 - \theta))^{k_2} (1 - \theta)^{2k_3}.$$

monotonic fn.

So the \log likelihood is given by

$$\text{constant} + 2k_1 \ln(\theta) + k_2 \ln(2\theta) + k_2 \ln(1 - \theta) + 2k_3 \ln(1 - \theta).$$

```
k1,k2,k3,t=var("k1,k2,k3,t")
2*k1*log(t)+k2*log(2*t)+k2*log(1-t)+2*k3*log(1-t)
```

Example 5 (cont.)

We set the derivative equal to zero:

$$\frac{2k_1 + k_2}{\theta} - \frac{k_2 + 2k_3}{1 - \theta} = 0.$$

Solving for θ , we find the MLE is

$$\hat{\theta} = \frac{2k_1 + k_2}{2k_1 + 2k_2 + 2k_3}.$$

#

Julia Program 3 (Find $\hat{\theta}$ that MLE of (8))

```
In [6]: using Pycall
        @pyimport SymPy as sm

In [24]: k1,k2,k3,theta,c=sm.symbols("k1 k2 k3 theta c")
        lexpr=c+2*k1*sm.log(theta)+k2*sm.log(theta)+k2*sm.log(1-theta)+2*k3*sm.log(1-theta)

Out[24]: PyObject c + 2*k1*log(theta) + k2*log(theta) + k2*log(-theta + 1) + 2*k3*log(-theta + 1)

In [15]: expr=sm.diff(lexpr,theta)

Out[15]: PyObject 2*k1/theta - k2/(-theta + 1) + k2/theta - 2*k3/(-theta + 1)

In [23]: sm.solve(expr,theta)

Out[23]: 1-element Array{Pycall.PyObject,1}:
        PyObject (k1 + k2/2)/(k1 + k2 + k3)
```

Example 6

Suppose that the signal $X \sim N(0, \sigma_X^2)$ is transmitted over a communication channel. Assume that the received signal is given by

$$Y = X + W,$$

where $W \sim N(0, \sigma_W^2)$ is independent of X .

- 1 Find the ML estimate of X , given $Y = y$ is observed.
- 2 Find the MAP estimate of X , given $Y = y$ is observed.

Example 6 (cont.)

Solution. The PDF for r.v. X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{x^2}{2\sigma_X^2}}.$$

Since $(Y \mid X = x) \sim N(x, \sigma_W^2)$, thus the conditional PDF is

$$f_{Y|X}(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-\frac{(y-x)^2}{2\sigma_W^2}}.$$

Example 6 (cont.)

- 1 The ML estimate of X , given $Y = y$, is the value of x that maximizes

$$f_{Y|X}(y | x) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-\frac{(y-x)^2}{2\sigma_W^2}}.$$

To maximize the above function, we should minimize $(y - x)^2$. Therefore, we conclude

$$\hat{x}_{ML} = y.$$

- 2 The MAP estimate of X , given $Y = y$, is the value of x that maximizes

$$f_{Y|X}(y | x)f_X(x) = c \exp \left\{ - \left[\frac{(y - x)^2}{2\sigma_W^2} + \frac{x^2}{2\sigma_X^2} \right] \right\},$$

Example 6 (cont.)

where c is a constant. To maximize the above function, we should minimize

$$\frac{(y - x)^2}{2\sigma_W^2} + \frac{x^2}{2\sigma_X^2}. \quad (9)$$

By differentiation, we obtain the MAP estimate of x as

$$\hat{x}_{MAP} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} y.$$

#

Julia Program 4 (Find the MAP estimate of x from (9))

```
In [3]: using PyCall
        @pyimport SymPy as sm
```

```
In [9]: x,y,sig_w,sig_x=sm.symbols("x y sig_w sig_x")
        expr=(y-x)^2/(2*sig_w^2)+x^2/(2*sig_x^2)
        dexpr=sm.diff(expr,x)
        sm.solve(dexpr,x)
```

```
Out[9]: 1-element Array{PyObject,1}:
        PyObject sig_x**2*y/(sig_w**2 + sig_x**2)
```

Example 7 (Bayesian Inference for Coin Toss Experiment)

Consider the basic random experiment where we toss a biased coin n times. Suppose that the outcomes are x_1, \dots, x_n , with $x_i = 1$ if the i th toss is Heads and $x_i = 0$ otherwise, $i = 1, \dots, n$. Let θ denote the probability of Heads. We wish to obtain information about θ from the data $\mathbf{x} = (x_1, \dots, x_n)$. For example, we wish to construct a confidence interval.

Solution.

Example 7 (cont.)

Let prior pdf $f(\theta)$ is given by a uniform prior $f(\theta) = 1, 0 \leq \theta \leq 1$. We assume that conditional on θ the $\{x_i\}$ are independent and $\text{Ber}(\theta)$ distributed. Thus, the Bayesian likelihood is

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^s (1 - \theta)^{n-s},$$

where $s = x_1 + \cdots + x_n$ represents the total number of successes. Using a uniform prior gives the posterior pdf

$$f(\theta \mid \mathbf{x}) = c \theta^s (1 - \theta)^{n-s}, \quad 0 \leq \theta \leq 1.$$

Example 7 (cont.)

This is the pdf of the $\text{Beta}(s + 1, n - s + 1)$ distribution. The normalization constant is $c = (n + 1) \binom{n}{s}$. The graph of the posterior pdf for $n = 100$ and $s = 1$ is given in Fig 3.

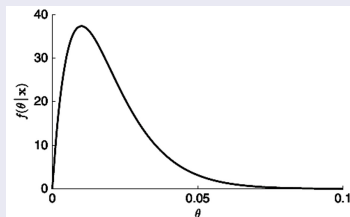


Fig 5. Posterior pdf for θ , with $n = 100$ and $s = 1$

Example 7 (cont.)

A Bayesian confidence interval, called a *credible interval*, for θ is formed by taking the appropriate quantiles of the posterior pdf. As an example, suppose that $n = 100$ and $s = 1$. Then, a left one-sided 95% credible interval for θ is $[0, 0.0461]$, where 0.0461 is the 0.95 quantile of the $\text{Beta}(2, 100)$ distribution.

#

Definition 20 (Bayesian Network)

Mathematically, a *Bayesian network* is a directed acyclic graph, that is, a collection of vertices (nodes) and arcs (arrows between nodes) such that arcs, when put head-to-tail, do not create loops.

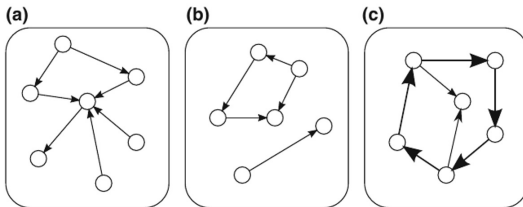


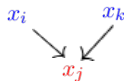
Fig 6. Graph of Network

Remark 4

The directed graphs in (a) and (b) are acyclic. Graph (c) has a (directed) cycle and can therefore not represent a Bayesian network

Bayesian networks can be used to graphically represent the joint probability distribution of a collection of random variables. In particular, consider a Bayesian network with vertices labeled x_1, \dots, x_n . Let P_j denote the set of parents of x_j , that is, the vertices x_i for which there exists an arc from x_i to x_j in the graph. We can associate with this network a joint pdf

$$P_j = \{x_i, x_k\} \Leftarrow$$



$$f(x_1, \dots, x_n) = \prod_{j=1}^n f(x_j \mid \mathcal{P}_j).$$

By the product rule, we obtain

$$f(x_1, \dots, x_n) = f(x_1)f(x_2 \mid x_1) \cdots f(x_n \mid x_1, \dots, x_{n-1}).$$

Example 8

Consider

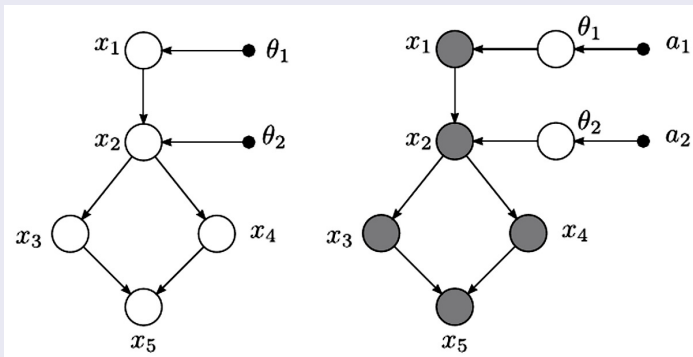


Fig 7. Left: a classical statistical model Right: corresponding Bayesian model with observed (i.e., fixed) data x_1, \dots, x_n , indicated by shaded nodes.

Example 8 (cont.)

The left plane of Fig. 5 shows a classical statistical model with random variables x_1, \dots, x_5 and fixed parameters θ_1, θ_2 :

$$f(x_1, \dots, x_n) = f(x_1; \theta_1) f(x_2 | x_1; \theta_2) f(x_3 | x_2) \\ f(x_4 | x_2) f(x_5 | x_3, x_4).$$

The right plane of Fig. 5 shows the corresponding Bayesian model.

$$f(x_1, \dots, x_n) = \prod_{j=1}^n f(x_j | \mathcal{P}_j) \\ f(x_1, \dots, x_n) = f(x_1) f(x_2 | x_1) f(x_3 | x_2) f(x_4 | x_2) f(x_5 | x_3, x_4).$$

It represents the situation where the “data” x_1, \dots, x_n have been observed. The aim is to find the posterior pdf of θ_1 and θ_2 given the data.

#

Example 9 (Applied Bayesian Networks)

Graph representations:

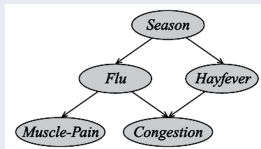


Fig 8. A sample Bayesian networks

Independencies:

no arc between X and Y

Let random variable X is independent of Y given Z
denoted by $(X \perp Y \mid Z)$. Then $\perp :=$ perpendicular

$$(F \perp H \mid S), (C \perp S \mid F, H), (M \perp H, C \mid F), (M \perp C \mid F)$$

$$f(x_1, \dots, x_n) = \prod_{j=1}^n f(x_j \mid \mathcal{P}_j).$$

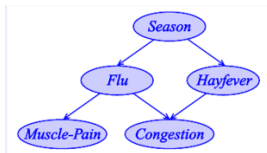
Example 9 (cont.)

Factorization:

$$P(S, F, H, C, M) = P(S) \cdot P(F \mid S) \cdot P(H \mid S) \cdot$$

$$P(C \mid F, H) \cdot P(M \mid F)$$

#



Example 10 (Belief Nets)

The purpose of this belief net is to determine if a patient is to be diagnosed with heart disease, based on several factors and symptoms. Two important factors in heart disease are smoking and age, and two main symptoms are chest pains and shortness of breath. The belief net in Fig. 8 shows the prior probabilities of smoking and age, the conditional probabilities of heart disease given age and smoking, and the conditional probabilities of chest pains and shortness of breath given heart disease.

Example 10 (cont.)

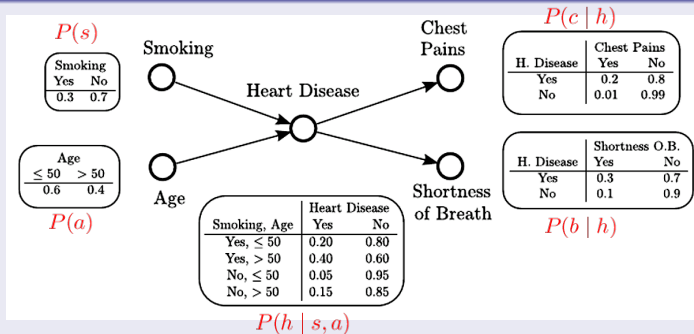


Fig 9. A Bayesian belief net for the diagnosis of heart disease

Example 10 (cont.)

Solution.

- The belief net in Fig. 8 shows the prior probabilities of smoking and age, the conditional probabilities of heart disease given age and smoking, and the conditional probabilities of chest pains and shortness of breath given heart disease.
- Suppose a person experiences chest pains^{*c*} and shortness of breath^{*b*}, but we do not know her/his age and if she/he is smoking. How likely is it that she/he has a heart disease?

$$f(h = \text{yes} | b = \text{yes}, c = \text{yes}) = ?$$

Example 10 (cont.)

Define the variables s (smoking), a (age), h (heart disease), c (chest pains), and b (shortness of breath). We assume that s and a are independent. Let “Yes” denoted by “Y” and “No” denoted by “N”. We wish to calculate

$$P(h = \text{Yes} \mid b = \text{Yes}, c = \text{Yes}) \triangleq P(h = Y \mid b = Y, c = Y).$$

From the Bayesian network structure, we see that the joint pdf of s, a, h, c and b can be written as

$$f(s, a, h, c, b) = f(s)f(a)f(h \mid s, a)f(c \mid h)f(b \mid h).$$

It follows that

$$f(h \mid b, c) \propto \underbrace{f(c \mid h)f(b \mid h)}_{f(h)^\dagger} \sum_{a, s} \underbrace{f(h \mid s, a)f(s)f(a)}_{s \perp a}.$$

Annotations in the original image:
- A blue bracket above $f(b, c \mid h)$ is labeled $b, c : \text{ind}$.
- The term $s \perp a$ is written in red above the summation.

Example 10 (cont.)

We have

$$\begin{aligned} P(h = Y) &= P(h = Y \mid s = Y, a \leq 50) \cdot P(s = Y) \cdot P(a \leq 50) + \\ &\quad P(h = Y \mid s = Y, a > 50) \cdot P(s = Y) \cdot P(a > 50) + \\ &\quad P(h = Y \mid s = N, a \leq 50) \cdot P(s = N) \cdot P(a \leq 50) + \\ &\quad P(h = Y \mid s = N, a > 50) \cdot P(s = N) \cdot P(a > 50) \\ &= 0.2 \times 0.3 \times 0.6 + 0.4 \times 0.3 \times 0.4 \\ &\quad + 0.05 \times 0.7 \times 0.6 + 0.15 \times 0.7 \times 0.4 = 0.147. \end{aligned}$$

Example 10 (cont.)

Consequently,

$$P(h = Y \mid b = Y, c = Y) = \beta \cdot P(c = Y \mid h = Y) \cdot$$

$$P(b = Y \mid h = Y) \cdot P(h = Y)$$

$$= \beta \times 0.2 \times 0.3 \times 0.147 = \beta 0.00882$$

and

$$P(h = N \mid b = Y, c = Y) = \beta \cdot P(c = Y \mid h = N) \cdot$$

$$P(b = Y \mid h = N) \cdot P(h = N)$$

$$= \beta \times 0.01 \times 0.1 \times (1 - 0.147)$$

$$= \beta 0.000853$$

Example 10 (cont.)

for some normalization constant β . Thus,

$$\begin{aligned} f(h = \text{Yes} \mid b = \text{Yes}, c = \text{Yes}) &= \frac{0.00882}{0.0882 + 0.000853} \\ &= 0.911816 \approx 0.91. \end{aligned}$$

#

So, we are sure he will have a heart attack.