MONASH University

ETC2420

Statistical methods in Insurance

Week 10.
Bayesian Reasoning: Introduction

22 September 2016

Outline

Week	Topic	Lecturer
1	Randomization & Hypothesis Testing I	Souhaib & Di
2	Hypothesis Testing II & Decision Theory	Souhaib
3	Statistical Distributions	Di
4	Model fitting & Linear regression	Di
5	Linear models	Di
6	Bootstrap, Permutation and Linear models	Di
	Multilevel models	Di
7	Generalized Linear models	Di
8	Compiling data for problem solving	Di
9	Bayesian Reasoning I & II	Souhaib
10	Bayesian Reasoning III & Time series models I	Souhaib
10	Time series models II & III	Souhaib
11	Project presentation	Souhaib

References

- Berger, J. O. 2013. **Statistical Decision Theory and Bayesian Analysis**. Springer Series in Statistics. Springer New York.
- Blitzstein, Joseph K., and Jessica Hwang. 2014. **Introduction to Probability** (Chapman & Hall/CRC Texts in Statistical Science). Chapman and Hall/CRC.
- Wasserman, Larry. 2004. **All of Statistics: A Concise Course in Statistical Inference** (Springer Texts in Statistics). 1st Corrected ed. 20 edition. Springer.

Frequentist philosopy

- ”Probability refers to **limiting relative frequencies**. Probabilities are objective properties of the real world.”
- ”Parameters are **fixed, unknown constants**. Because they are not fluctuating, no useful probability statements can be made about parameters.”
- ”Statistical procedures should be designed to have **well-defined long run frequency properties**. For example, a 95 percent confidence interval should trap the value of the parameter with limiting frequency at least 95 percent.”

Bayesian philosopy

- Probability describes **degree of belief**, not limiting frequency. As such, we can make probability statements about anything that is subject to random variation. For example, *the probability that it will rain on Sunday is* .30. This does not refer to any limiting frequency. It reflects my strength of belief that the proposition is true.
- We can make **probability statements about parameters**, even though they are fixed constants.
- We make inferences about a parameter by producing **a probability distribution for the parameter**. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Conditional probability

If A and B are events with $P(B) > 0$, then the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

For any events A and B with positive probabilities

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(or Bayes' theorem)

Example

A patient name Bob is tested for a disease that **afflicts** 1% **of the population**. The test result is **positive**, i.e., the test claims that the Bob has the disease. Let D be the event that Bob has the disease and T be the event that he tests positive. Suppose that the test is 95% accurate (we suppose $P(T|D) = 0.95$ and $P(T|D^c) = 0.95$). What is $P(D|T)$?

$$P(D) = 0.01 \rightarrow P(D|T) = ?$$

1. P(D|T) = 0.02
2. P(D|T) = 0.16
3. P(D|T) = 0.99

Example

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)}$$
$$= \frac{0.95 \cdot 0.01}{P(T)}$$

(Law of total probability) Let $A_1, \ldots, A_n$ be a partition of the sample space S (i.e. the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Example

$$
\begin{aligned}
P(D|T) &= \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \\
&= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \\
&\approx 0.16
\end{aligned}
$$

Why such a small chance to have the disease given that he tested positive, and the test is reliable ($P(T|D) = 0.95$)?

Example

$$
\begin{aligned}
P(D|T) &= \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \\
&= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \\
&\approx 0.16
\end{aligned}
$$

Why such a small chance to have the disease given that he tested positive, and the test is reliable ($P(T|D) = 0.95$)?

- Two factors: the **evidence** from the test and the **prior information** about the disease.
- Although the test provides evidence in favour of disease, the disease is rare.
- The conditional probability reflects a balance between these two factors

The Bayesian method

We are interested in the unknown parameter θ.

- We choose a probability density $\pi(\theta)$, called the **prior distribution**, that expresses our beliefs about the parameter θ before we see any data.
- We choose a probability distribution $f(x|\theta)$ that reflects our beliefs about the random variable X given θ.
- After observing data point x, we update our beliefs and calculate the **posterior** distribution $\pi(\theta|x)$:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} \propto f(x|\theta)\pi(\theta)$$

where

$$f(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

Comparison with MLE

Suppose we use an "uninformative" prior ($\pi(\theta) = 1$), then

$$\begin{aligned}
\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{f(x)} \\
&= f(x|\theta)k(x) \\
&\propto f(x|\theta) = f(x;\theta) \\
\mathcal{L}(\theta|x) &= f(x;\theta)
\end{aligned}$$

$$\implies (MLE)\,\text{maximize}_\theta\, \mathcal{L}(\theta|x) \equiv (MAP)\,\text{maximize}_\theta\, \pi(\theta|x)$$

Note:

- Any prior includes information, including priors that state that no information is known or that do not favour some values over others.
- $\hat{\theta}_{\text{MLE}} \neq E[\theta|x]$

Example

Consider estimating the probability θ that a coin will turn up heads.

1. Before tossing the coin, what is θ?
2. We toss the coin one time, and we see tails. What is θ?

The Bayesian method

If we have n i.i.d. observations $x_1, \ldots, x_n$, we can replace $f(x|\theta)$ with

$$f(x_1, \ldots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta) = \mathcal{L}_n(\theta)$$

and compute

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{f(x_1, \ldots, x_n)}$$

where

$$f(x_1, \ldots, x_n) = \int_\Theta \mathcal{L}_n(\theta)\pi(\theta)d\theta = c_n$$

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

Since c_n does not depend on θ, Posterior is proportional to Likelihood times Prior.

Comparison with MLE

Let $X_1, \ldots, X_n$ be i.i.d. with PDF $f(x; \theta)$. The likelihood function is defined by

$$\mathcal{L}_n(\theta) = \Pi_{i=1}^{n} f(X_i; \theta).$$

The log-likelihood is defined by $l_n(\theta) = log\mathcal{L}_n(\theta)$.

The likelihood is the joint density of the data, *treated as a function of the parameter* θ. Thus, $\mathcal{L}_n : \Theta \to [0, \infty]$. $\mathcal{L}_n$ is not a density function: in general, $\int_\Theta \mathcal{L}_n(\theta) \neq 1$.

The maximum likelihood estimator (MLE) is defined by

$$\hat{\theta}_n = \text{argmax}_\theta \, \mathcal{L}_n(\theta).$$

Example with the Bayesian method

Suppose that $X_1, \ldots, X_n \sim \text{Bernouilli}(p)$. We want to estimate p.

- $\pi(p) = 1; 0 \leq p \leq 1$ (uniform prior distribution)
- We observe $x_1, \ldots, x_n$. Compute $\pi(p|x_1, \ldots, x_n)$.

If $s = \sum_{i=1}^{n} x_i$ is the number of successes, then

$$\pi(p|x_1, \ldots, x_n) = \frac{\mathcal{L}_n(p)\pi(p)}{c_n} = \frac{p^s(1-p)^{n-s} \cdot 1}{c_n},$$

where

$$c_n = \int_0^1 p^s(1-p)^{n-s} dp = \frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)}$$

and $\Gamma(n) = (n-1)!$

Example with the Bayesian method

The pdf of the **beta distribution**, for $0 \leq x \leq 1$, is given by

$$f(x;\alpha;\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

with shape parameters $\alpha, \beta > 0$.

- We can see that $p|x_1, \ldots, x_n \sim \text{Beta}(s+\alpha, n-s+\beta)$ with $\alpha = \beta = 1$.
- It is also possible to consider different values for α and β which will lead to different prior assumptions.
- From the posterior, we can compute multiple posterior quantities (analytically or via simulation): mean, standard deviation, credible intervals, etc.

Example with MLE

Suppose that $X_1, \ldots, X_n \sim$ Bernouilli(p). We want to compute the Maximum Likelihood Estimate (MLE) of p. The likelihood function is defined by

$$\mathcal{L}_n(p) = \Pi_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

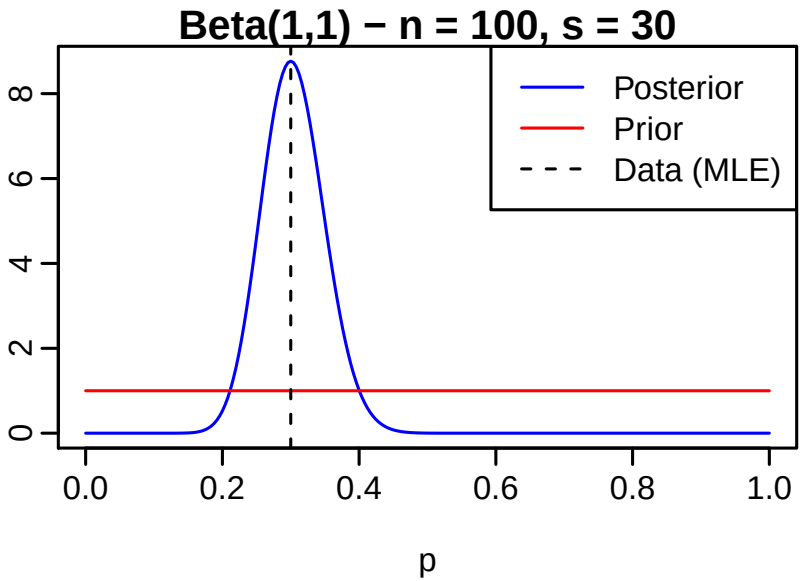
where $S = \sum_i X_i$.
The log-likelihood is given by

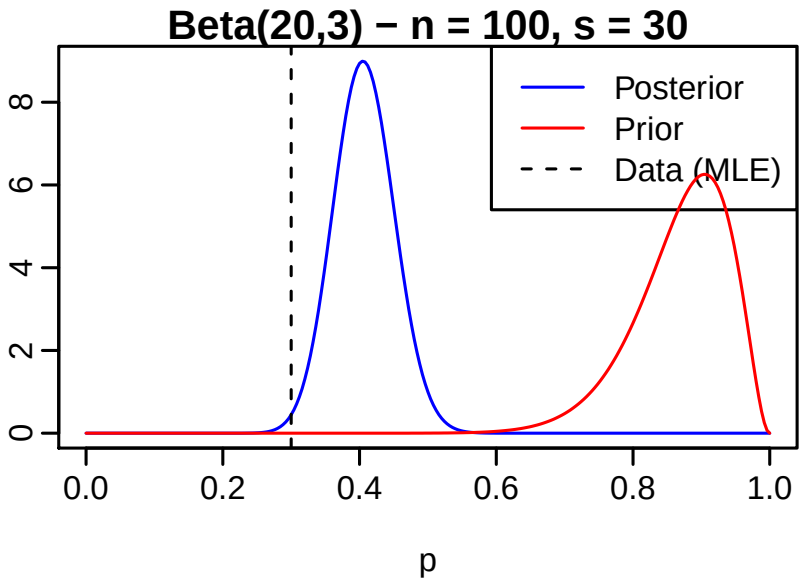
$$l_n(p) = Slog(p) + (n-S)log(1-p)$$

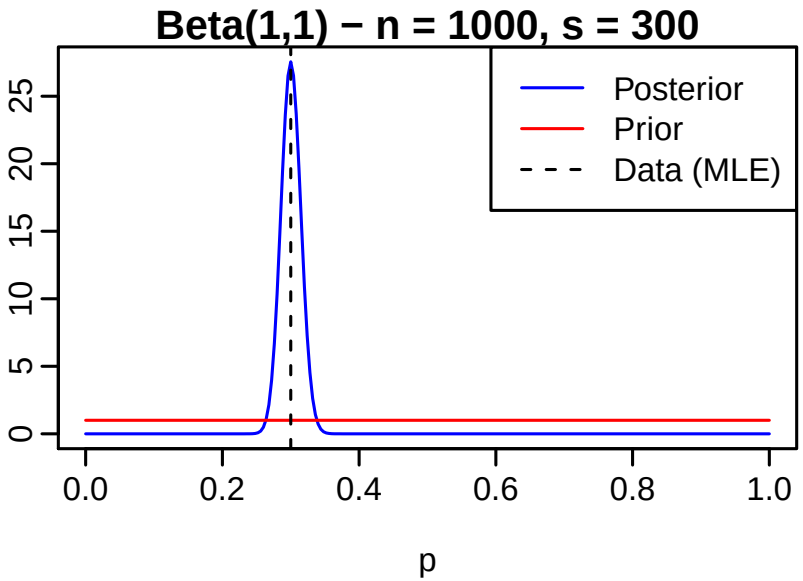
The MLE estimate is then given by

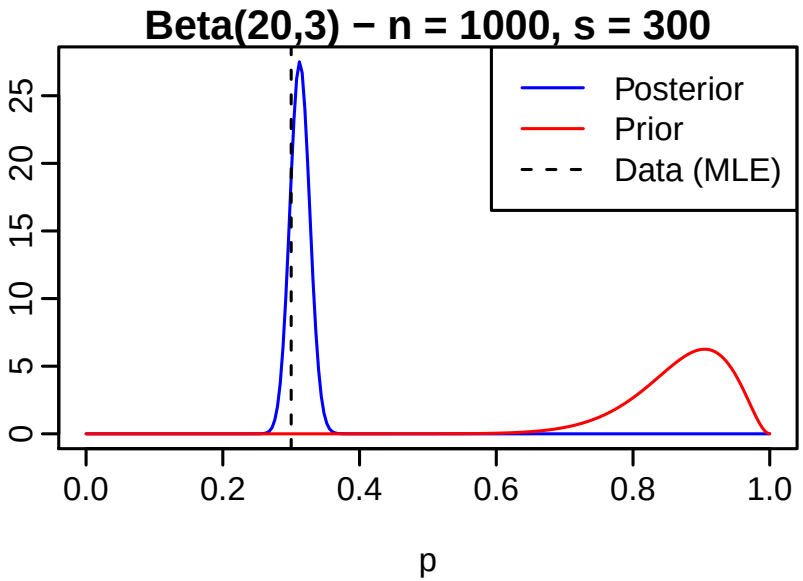
$$\hat{p}_n = \text{argmax}_p \, l_n(p) = \frac{S}{n}$$

Bayesian method vs MLE









Conjugate priors

In the previous example, we had

$$\text{Beta}(s+\alpha, n-s+\beta) = \frac{\mathcal{L}_n(p) \times \text{Beta}(\alpha, \beta)}{c_n}$$

- When the posterior distributions are in the same distribution family as the prior distribution, they are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function
- Conjugate priors are convenient since we have a closed-form expression for the posterior (we avoid numerical integration).
- In practice, we have conjugacy only for very simple models. In most cases, the posterior distribution has to be found numerically via MCMC.