STAT2006 Basic Concepts in Statistics and Probability II
Philosophy in Statistics: Believe it or not

Benjamin Chun Ho Chan[*][†][‡][§]

January 4, 2019

Abstract

From a high-level point of view, there are at least two kinds of ways to do statistical inference, namely Bayesian and frequentist. Here it should be emphasized that Bayesian inference is unrelated to STAT2006 examinations. The target audience is mathematically mature students and those who are curious about deeper understanding of Statistics.

Some materials are extracted from STAT3006 Statistical Computing lecture note and do credit to Prof Wei Yingying. It is great to learn about statistical computing and algorithms, in particular Markov chain Monte Carlo algorithms, with application of Bayesian Statistics. RMSC4001 Simulation Techniques in Financial Risk Management is another course to learn simulation with application to risk management.

1 Two Schools of Statisticians

- **Frequentist approach** is to view that samples are random and parameters are fixed. Topics: most of the undergraduate techniques
- **Bayesian approach** is to view that samples are random and parameters are random. Topics: Bayes' rule, prior distributions, posterior distributions, etc.
- You have been learning the classical frequentist techniques throughout the course. You should have learnt Bayes' rule. A naive way to distinguish between frequentist approach and Bayesian approach is whether Bayes' rule is used or not. If Bayes' rule is to be used, it can be considered as Bayesian inference, if not, frequentist inference.
- Although the name has suggested this point. The more fundamental question is whether we have any prior belief/ information or not and whether we can incorporate prior knowledge in analyzing problems, in which the word "prior distribution" comes from. In fact, there is philosophy underneath every subject.
- There is a series of Youtube videos from Brian Caffo, a professor at the Department of Biostatistics at Johns Hopkins University, introducing the puzzle about Statistics:

  A rambling rant about Bayes versus frequentist statistics
  (https://www.youtube.com/watch?v=qAFBsUVUtp8)

  How can I get started in Bayesian data analysis?
  (https://www.youtube.com/watch?v=ACq0KAL5_2Q)

[*] For enquiry, please email to 1155049861@link.cuhk.edu.hk.

[†] Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho

[‡] GitHub repository: https://github.com/BenjaminChanChunHo/CUHK-STAT-or-RMSC-Tutorial-Note

[§] RPubs: http://rpubs.com/Benjamin_Chan_Chun_Ho

2 Frequentist

- Frequentist approach is to view that samples are random and parameters are fixed.
- Maximum Likelihood Estimation (MLE) is the value $\hat{\theta}$ at which the likelihood function $L(\theta;\boldsymbol{X})$ is maximized, where $\boldsymbol{X}=(X_1,\dots,X_n)^T$. In general, $\boldsymbol{\theta}$ can be a vector as well. In other words,

  $$\hat{\theta} := \arg\max_{\theta} L(\theta;\boldsymbol{X}).$$

  Since ln transformation is monotone, to be specific, strictly increasing,

  $$\hat{\theta} = \arg\max_{\theta} l(\theta;\boldsymbol{X}).$$

Example (Normal): Assume that $X \sim N(\mu, 1)$.

$$\begin{aligned} l(\mu;x_1,\dots,x_n) &= ln\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\text{exp}\Big\{-\frac{(x_i-\mu)^2}{2}\Big\} \\ &= \sum_{i=1}^{n}\Big[-\frac{1}{2}ln(2\pi)-\frac{(x_i-\mu)^2}{2}\Big] \\ &= -\frac{n}{2}ln(2\pi)-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} \end{aligned}$$

Taking derivative with respect to μ, we set that

$$\frac{\partial l(\mu;x_1,\dots,x_n)}{\partial\mu} = \sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n}x_i}{n} = \bar{x}.$$

Please refer to Tutorial 5: Point Estimation for verification of global maximum.

---

3 Bayesian

- Bayesian approach is to view that samples are random and parameters are random.
- Prior distribution $\pi(\theta)$: some prior belief (e.g. previous knowledge, expert advice) about the parameters.
- Likelihood function $L(\theta;x_1,\dots,x_n)=L(\theta;\boldsymbol{x})$: the same as the one in the frequentist statistics.
- Posterior distribution $p(\theta;\boldsymbol{x})$: update your belief about the parameters after observing the data.
- Bayes' rule: $p(A|B)=\dfrac{p(B|A)p(A)}{p(B)}$.
- The posterior distribution is given by

  $$p(\theta;\boldsymbol{x}) = \frac{p(\theta,\boldsymbol{x})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{\int p(x,\theta)d\theta} = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{\int p(\boldsymbol{x}|\theta)\pi(\theta)d\theta}$$

Example (Normal): $\pi(\mu) = N(a, b^2)$.

$$p(\mu|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mu)\pi(\mu)}{\int p(\boldsymbol{x}|\mu)\pi(\mu)d\mu}$$

The denominator is constant as a function of μ, so we focus on the "kernel" part: $p(\boldsymbol{x}|\mu)\pi(\mu)$.

$$\begin{aligned} p(\boldsymbol{x}|\mu)\pi(\mu) &= \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\text{exp}\left\{-\frac{(x_i-\mu)^2}{2}\right\}\right] \cdot \frac{1}{\sqrt{2\pi}b}\text{exp}\left\{-\frac{(\mu-a)^2}{2b^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1}\frac{1}{b}\text{exp}\left\{-\frac{\sum_{i=1}^{n}(\mu-x_i)^2}{2}-\frac{(\mu-a)^2}{2b^2}\right\} \end{aligned}$$

Again, focus on the kernel part. Then

$$\begin{aligned} p(\boldsymbol{x}|\mu)\pi(\mu) &\propto \text{exp}\left\{-\frac{(b^2n+1)\mu^2-2(a+\sum_{i=1}^{n}x_ib^2)\mu}{2b^2}\right\} \\ &= \text{exp}\left\{-\frac{\mu^2-2\left(\frac{\frac{a}{b^2}+n\bar{x}}{\frac{1}{b^2}+n}\right)\mu}{2\frac{1}{n+1/b^2}}\right\} \\ &\propto \text{exp}\left\{-\frac{\left(\mu-\frac{\frac{a}{b^2}+n\bar{x}}{\frac{1}{b^2}+n}\right)^2}{2\frac{1}{n+1/b^2}}\right\}, \end{aligned}$$

which is the kernel of the normal distribution. The posterior distribution of μ (the distribution of μ after observing data $\boldsymbol{x}$) is also a normal distribution. Specifically, assume η, τ^2 are the mean and the variance of μ's posterior distribution, respectively, then we have

$$\eta = \frac{\frac{a}{b^2}+n\bar{x}}{\frac{1}{b^2}+n} \qquad \text{and} \qquad \tau^2 = \frac{1}{n+\frac{1}{b^2}}.$$

Note that $\mu \sim N(a, b^2)$ and $\bar{x} \sim N(\mu, \frac{1}{n})$. If we call $\frac{1}{\text{variance}}$ as precision, then the precision of the posterior ($\frac{1}{\tau^2}$) equals the summation of the precision of the likelihood ($\frac{1}{1/n} = n$) and the precision of the prior ($\frac{1}{b^2}$). Furthermore, the mean of the posterior (η) is the weighted average of sample mean ($\bar{x}$, MLE) and the prior mean (a), and the weights are exactly the precisions.

As shown, the posterior distribution $N\left(\frac{\frac{a}{b^2}+n\bar{x}}{\frac{1}{b^2}+n}, \frac{1}{n+\frac{1}{b^2}}\right)$, it contains not only the information from the sample $\bar{x}$ but also the information from the prior a and b^2. When n is relatively small, the prior information play a big role for the estimation of μ. When n is large, the information from samples dominates the estimation of μ.

In Bayesian analysis, we only concern the kernel part because of the following:

Theorem. If $f(x) = c_0 ker(x)$ is a density function, and $h(x) = d_0 ker(x)$ is also a density function, then $c_0 = d_0$, $f(x) = h(x)$ for all x.

Proof. On one hand, $f(x)$ is a pdf. So $\int f(x)dx = 1 \Rightarrow c_0 \int ker(x)dx = 1 \Rightarrow c_0 = \frac{1}{\int ker(x)dx}$. On the other hand, $h(x)$ is a pdf. So $\int h(x)dx = 1 \Rightarrow d_0 \int ker(x)dx = 1 \Rightarrow d_0 = \frac{1}{\int ker(x)dx}$. Therefore, $c_0 = d_0$ and hence $f(x) = h(x)$. □