STAT 3006: Statistical Computing
Lecture 1*

8 January

1 Introduction

1.1 Probability and Statistics

Q1. What is probability? Flipping a coin, the chance of head up is 50%.
Q2. What is statistics? Flipping 100 coins, 48 of them are heads up. Use the 100 samples to estimate the chance of heads up.

Mathematically, we have a distribution (or population) $p(x|\Theta)$ (Θ may be one parameter or a parameter vector) and a set of independent and identically distributed (i.i.d.) samples $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ from F. The probability is that given the population $p(x|\Theta)$, we investigate the properties of $\mathbf{X}$. The statistics is that given the samples $\mathbf{X}$, we estimate the population $p(x|\Theta)$ or equivalently the unknown Θ.

In practice, the population is always unknown, so we accumulate samples to obtain knowledge about the population (Statistics). In the meanwhile, the procedure to gain population knowledge are largely based on the probability theory (e.g. Method of Moments, Central Limit Theory) and optimization techniques (e.g. Maximum Likelihood Estimator).

Likelihood function: $L(\Theta|\mathbf{X}) = p(\mathbf{X}|\Theta) = \prod_{i=1}^{n} p(X_i|\Theta)$.

Example: suppose $X_1, X_2, \ldots, X_n$ (i.i.d.) from the normal distribution $\mathcal{N}(\mu, 1)$,

$$\begin{aligned} L(\mu|x_1, \ldots, x_n) &= \prod_{i=1}^{n} p(x_i|\mu) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\{-\frac{(x_i - \mu)^2}{2}\}. \end{aligned}$$

[*] If you have any question about the note, please send an email to xyluo@link.cuhk.edu.hk

Remark 1. Strictly speaking, X denotes a random variable and its lower case x denotes X's realization (observed sample). Sometimes, we also use X to denote the observed sample (e.g. $p(X|\Theta)$ is the same as $p(x|\Theta)$) in our note.

1.2 Two Schools of Statisticians

Frequentists:

- Samples are random.
- Parameters are fixed.

Definition 1.1. Maximum Likelihood Estimation (MLE) is the value $\hat{\Theta}$ at which the likelihood function $L(\Theta|\mathbf{X})$ is maximized. In other words, $\hat{\Theta} := \arg\max_{\Theta} L(\Theta|\mathbf{X})$.
the log-likelihood function

$$\begin{aligned} l(\Theta|X_1, X_2, \ldots, X_n) &= log L(\Theta|\mathbf{X}) \\ &= log p(\mathbf{X}|\Theta) \\ &= (i.i.d.) log \prod_{i=1}^{n} p(X_i|\Theta) \\ &= \sum_{i=1}^{n} log p(X_i|\Theta). \end{aligned}$$

Since log transformation is monotone, maximizing $l(\Theta|X_1, X_2, \ldots, X_n)$ is equivalent to maximizing $L(\Theta|\mathbf{X})$. Therefore,

$$\hat{\Theta} = \arg\max_{\Theta} l(\Theta|\mathbf{X}).$$

Example (Normal):

$$\begin{aligned} l(\mu|x_1, x_2, \ldots, x_n) &= log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp \left\{ -\frac{(x_i-\mu)^2}{2} \right\} \\ &= \sum_{i=1}^{n} \left[-\frac{1}{2} log(2\pi) - \frac{(x_i-\mu)^2}{2} \right] \\ &= -\frac{n}{2} log(2\pi) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} \end{aligned}$$

Taking derivative with respect to (w.r.t.) μ, we have that

$$\begin{aligned} \frac{\partial l(\mu|x_1, x_2, \ldots, x_n)}{\partial \mu} &= \frac{\sum_{i=1}^{n} 2(x_i-\mu)}{2} = 0 \\ \Rightarrow \hat{\mu} &= \frac{\sum_{i=1}^{n} x_i}{n} = \bar{\mathbf{x}}. \end{aligned}$$

The problem of finding MLE is equivalent to solving an optimization problem.

Bayesian:

- Samples are random.
- Parameters are also random.

Three characteristics of Bayesian statistics:

- Prior distribution $\pi(\Theta)$: some prior belief (e.g. previous knowledge, expert advice) about the parameters.
- Likelihood function $L(\Theta|x_1, \dots, x_n) = L(\Theta|\mathbf{x})$: the same definition as the one in the frequentist statistics.
- Posterior distribution $p(\Theta|\mathbf{x})$: update your belief about the parameters after observing the data.

Q: How to derive the posterior distribution $p(\Theta|\mathbf{x})$? Bayes rule ($p(A|B) = \frac{p(B|A)p(A)}{p(B)}$).

$$\begin{aligned} p(\Theta|\mathbf{x}) &= \frac{p(\Theta, \mathbf{x})}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\Theta)\pi(\Theta)}{\int p(\mathbf{x}, \Theta) d\Theta} \\ &= \frac{p(\mathbf{x}|\Theta)\pi(\Theta)}{\int p(\mathbf{x}|\Theta)\pi(\Theta) d\Theta} \end{aligned}$$

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

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

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

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

which is the kernel of the normal distribution. Based on theorem 1.1 (discussed later), the posterior distribution of μ (the distribution of μ after observing data $\mathbf{x}$) 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

$$\eta = \frac{\frac{a}{b^2} + n\bar{\mathbf{x}}}{\frac{1}{b^2} + n} \tag{1.1}$$

$$\tau^2 = \frac{1}{n + \frac{1}{b^2}}. \tag{1.2}$$

Please notice that $\mu \sim \mathcal{N}(a, b^2)$ and $\bar{\mathbf{x}} \sim \mathcal{N}(\mu, \frac{1}{n})$. If we call $\frac{1}{variance}$ 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 ($\frac{1}{b^2}$) (1.2). Furthermore, the mean of the posterior (η) is the weighted average of sample mean ($\bar{\mathbf{x}}$, MLE!) and the prior mean (a), and the weights are exactly the precisions (1.1).

As you can see from the posterior distribution $\mathcal{N}(\frac{\frac{a}{b^2}+n\bar{\mathbf{x}}}{\frac{1}{b^2}+n}, \frac{1}{n+\frac{1}{b^2}})$, it contains not only the information from the sample $\bar{\mathbf{x}}$ 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 (e.g. n goes to infinity), the information from samples dominates the estimation of μ.

Theorem 1.1. *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 $\forall x$.*

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

1.3 A complex situation (normal mixture)

In the last example, we assume samples are i.i.d. from a normal distribution. How about a distribution with a complicated form? Assume $X_1, \dots, X_n$ from

$$p(x|\Theta) = \pi p(x|\mu_1, \sigma_1^2) + (1-\pi)p(x|\mu_2, \sigma_2^2), \tag{1.3}$$

where $\Theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi)$ and $p(x|\mu, \sigma^2)$ represents the density function of a normal distribution with mean μ and variance σ^2. We also call $\pi p(x|\mu_1, \sigma_1^2) + (1-\pi)p(x|\mu_2, \sigma_2^2)$ the normal mixture model.

How is each sample X_i drawn from $p(x|\Theta)$? We introduce a *latent* variable Z_i, which represents the group where X_i is from. When $Z_i = 1$, $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$; when $Z_i = 2$, $X_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$. It follows that

$$\begin{aligned} p(X_i|\Theta) &= \sum_{k=1}^{2} p(X_i, Z_i = k|\Theta) \\ &= \sum_{k=1}^{2} p(X_i|Z_i = k|\Theta) \cdot p(Z_i = k|\Theta) \\ &= \pi \cdot p(X_i|\mu_1, \sigma_1^2) + (1-\pi) \cdot p(X_i|\mu_2, \sigma_2^2). \end{aligned}$$

The calculation results in

$$X_i \sim p(x|\Theta)(1.3) \Leftrightarrow Z_i = \begin{cases} 1 & \textit{with probability } \pi \\ 2 & \textit{with probability } 1-\pi \end{cases} \qquad X_i \sim p(x|\mu_{Z_i}, \sigma_{Z_i}^2)$$

How to estimate unknown parameters Θ? We first investigate the likelihood function for all samples.

$$\begin{aligned} L(\mu_1, \mu_2, \sigma_1, \sigma_2, \pi|x_1, \dots, x_n) &= \prod_{i=1}^{n} p(x_i|\mu_1, \mu_2, \sigma_1, \sigma_2, \pi) \\ &= \prod_{i=1}^{n} \pi p(x|\mu_1, \sigma_1^2) + (1-\pi)p(x|\mu_2, \sigma_2^2) \end{aligned}$$

To derive the MLE of Θ, we need to search the values at which $L(\mu_1, \mu_2, \sigma_1, \sigma_2, \pi|x_1, \dots, x_n)$ attains maximum. Usually, the function L is too complex to be solved analytically. In practice, we can make use of numerical optimization methods (e.g. gradient descent) or statistical optimization algorithms (e.g. EM) or sampling algorithm (e.g. MCMC), all of which are computationally efficient and guaranteed by mathematical theories. Our course will focus on the last two statistical approaches.

2 Solution to Nonlinear Equations

2.1 Bisection

Problem: given a univariate and continuous function $g(x)$, we are interested in a value x_0 such that $g(x_0) = 0$. x_0 is called *zero point* of g.

Fact: If $f(a) \cdot f(b) < 0$ $(a < b)$, due to the continuity of function f, there exists at least one zero point between a and b.

Solution: based on the fact, we first initialize $a^{(0)}$ and $b^{(0)}$ s.t.(such that) $f(a^{(0)}) \cdot f(b^{(0)}) < 0$. $(*)$ Then we calculate the middle point $c^{(0)}$ of the interval $[a^{(0)}, b^{(0)}]$. There are three cases now: 1)If $f(c^{(0)})$ equals 0, we found one zero point and end the calculation procedure; 2) If $f(c^{(0)}) \cdot f(a^{(0)}) < 0$, then the zero point must fall in the interval $[a^{(0)}, c^{(0)}]$. In this case, we let $a^{(1)}$ be $a^{(0)}$ and let $b^{(1)}$ be $c^{(0)}$; 3)If $f(c^{(0)}) \cdot f(b^{(0)}) < 0$, we let $a^{(1)}$ be $c^{(0)}$ and let $b^{(1)}$ be $b^{(0)}$. In the last two cases, we obtained a new interval $[a^{(1)}, b^{(1)}] \subset [a^{(0)}, b^{(0)}]$, and then we repeat the $(*)$ procedure until $b^{(t)} - a^{(t)}$ is less than a specified tolerance.

```
Algorithm:
INPUT: continuous and univariate function $f$ and interval $[a, b]$ with $f(a)f(b) < 0$.
INITIALIZE: $a^{(0)} \leftarrow a$ and $b^{(0)} \leftarrow b$, and $t \leftarrow 0$.
Repeat
  calculate $c^{(t)} \leftarrow \frac{a^{(t)}+b^{(t)}}{2}$;
  If $f(c^{(t)}) \cdot f(a^{(t)}) < 0$, let $a^{(t+1)} \leftarrow a^{(t)}$ and $b^{(t+1)} \leftarrow c^{(t)}$;
    else if $f(c^{(t)}) \cdot f(b^{(t)}) < 0$, let $a^{(t+1)} \leftarrow c^{(t)}$ and $b^{(t+1)} \leftarrow b^{(t)}$;
    else break;
  $t \leftarrow t + 1$;
Until $|a^{(t)} - b^{(t)}| < \epsilon$.
OUTPUT: $a^{(t)}$, $b^{(t)}$ in the last iteration. $c^{(t)} \leftarrow \frac{a^{(t)}+b^{(t)}}{2}$ is the final answer.
```

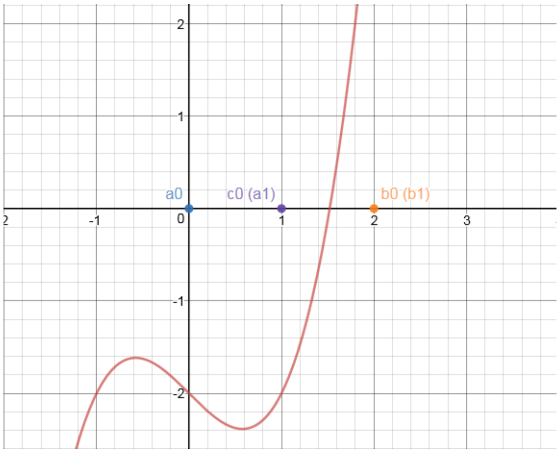

Figure 1: Figure demonstration for the bisection approach to obtain the zero point of $f(x) = x^3 - x - 2$.

Example: find the shortest confidence interval.

Problem: Y is a random variable with known probability density function $f(y)$. Given α_0 (e.g. $\alpha_0 = 0.95$),we want to find the shortest interval $[a, b]$ s.t. $P(a \leq Y \leq b) = \alpha_0$ or equivalently $\int_a^b f(y)dy = \alpha_0$.

Proposition: assume f is unimodal. If one interval $[a^*, b^*]$ satisfies $\int_{a^*}^{b^*} f(y)dy = \alpha_0$ and $f(a^*) = f(b^*)$, then for any other interval $[a, b]$ with $\int_a^b f(y)dy = \alpha_0$, we have $b^* - a^* < b - a$. (The proof will be discussed in the tutorial.)

Solution: based on the proposition, the shortest interval $[a, b]$ must satisfy $f(a) = f(b)$. We let λ be the value of $f(a)$ and $f(b)$. For any fixed λ, we use the bisection method to find a_λ and b_λ, respectively, s.t. $f(a_\lambda) \approx \lambda$, $f(b_\lambda) \approx \lambda$ and $a_\lambda < b_\lambda$. Then we numerically calculate the value of $\int_{a_\lambda}^{b_\lambda} f(y)dy$, $\alpha(\lambda)$. Finally we apply the bisection method to finding the zero point of $\alpha(\lambda) - \alpha_0$.

```
**Algorithm**:
INPUT: continuous, univariate and unimodal function $f$, $\alpha_0$(e.g. 0.95), tolerance $\epsilon$(e.g. e-5),
   a small value $\lambda_{lw}$ (near zero), and a large value $\lambda_{up}$ (near the maximum of $f$).
INITIALIZE: $\lambda_{lw}^{(0)} \leftarrow \lambda_{lw}$ and $\lambda_{up}^{(0)} \leftarrow \lambda_{up}$, and $t \leftarrow 0$.
**Repeat**
  calculate $\lambda_{mid}^{(t)} \leftarrow \frac{\lambda_{lw}^{(t)} + \lambda_{up}^{(t)}}{2}$ ;
  Use the bisection method to find one zero point of $f(x) - \lambda_{mid}^{(t)} = 0$, $\tilde{a}$ ;
  Use the bisection method again to find another zero point of $f(x) - \lambda_{mid}^{(t)} = 0$, $\tilde{b}$ ;
  (Without loss of generality, we assume $\tilde{a} < \tilde{b}$.)
  Numerically integrate $f(x)$ from $\tilde{a}$ to $\tilde{b}$, the result is denoted by $\alpha(\lambda_{mid}^{(t)})$;
  If $\alpha(\lambda_{mid}^{(t)}) < \alpha_0$, let $\lambda_{lw}^{(t+1)} \leftarrow \lambda_{lw}^{(t)}$ and $\lambda_{up}^{(t+1)} \leftarrow \lambda_{mid}^{(t)}$;
    else if $\alpha(\lambda_{mid}^{(t)}) > \alpha_0$, let $\lambda_{lw}^{(t+1)} \leftarrow \lambda_{mid}^{(t)}$ and $\lambda_{up}^{(t+1)} \leftarrow \lambda_{up}^{(t)}$;
    else break;
  $t \leftarrow t + 1$;
**Until** $|\lambda_{up}^{(t)} - \lambda_{lw}^{(t)}| < \frac{\epsilon}{2}$.
OUTPUT: $[\tilde{a}, \tilde{b}]$ is the shortest confidence interval with confidence level$\alpha_0$.
```

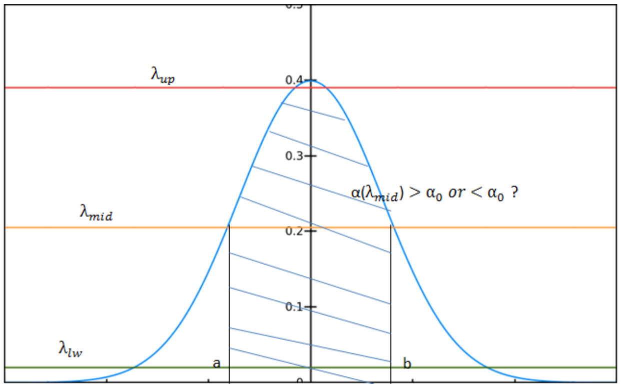

Figure 2: Figure demonstration for the nested bisection approach to find the shortest confidence interval.