ENGG2780 Statistics for Engineers

Ryan Chan

February 23, 2025

Abstract

This is a note for **ENGG2780 - Statistics for Engineers** for self-revision purpose ONLY. Some contents are taken from lecture notes and reference book.
Mistakes might be found. So please feel free to point out any mistakes.
Contents are adapted from the lecture notes of ENGG2780, prepared by Sinno Jialin Pan, as well as some online resources.
This course heavily relies on prior knowledge of probability (which you can refer to in the notes I wrote for ENGG2760). Therefore, before proceeding with this course, make sure you understand the foundation, as I will take them for granted.

Contents

Chapter 1

Bayesian Statistic

1.1 Statistic v.s. Probability

Statistics focuses on real-life applications where the underlying distribution is often unknown. To address this, we use **statistical inference** to analyze observed data and estimate the unknown distribution. Rather than finding the exact distribution, we approximate it using models such as parametric (e.g., normal, exponential) or non-parametric approaches. Once a suitable model is chosen, probability laws help us make predictions and draw conclusions, though these approximations involve assumptions and uncertainties.

Now, let's move on to our first topic in statistics:

1.2 Bayesian Statistics

1.2.1 Introduction

In the probability course, we learned Bayes' Rule ENGG2760: Theorem 3.2.1, which helps us calculate conditional probabilities and, at times, update our beliefs based on new evidence.

And it turns out that one of the statistical inferences we use is based on Bayes' rule, namely Bayesian statistical inference. In Bayesian statistical inference, we: (1) assign prior probabilities to parameters; (2) observe data; and (3) update probabilities via Bayes' rule:

$$\underbrace{f_{\Theta|X}(\theta|x)}_{\text{Posterior}} = \frac{\overbrace{f_\Theta(\theta)}^{\text{Prior}}\overbrace{f_{X|\Theta}(x|\theta)}^{\text{Observation}}}{f_X(x)}$$

Here we have both the posterior and prior probabilities of the parameters θ and observations x.

We have four variations of the Bayes' rule shown above.

Condition	Bayes' rule
Θ discrete, X discrete	$p_{\Theta\|X}(\theta\|x) = \frac{p_\Theta(\theta)p_{X\|\Theta}(x\|\theta)}{\sum_{\theta'} p_\Theta(\theta')p_{X\|\Theta}(x\|\theta')}$
Θ discrete, X continuous	$p_{\Theta\|X}(\theta\|x) = \frac{p_\Theta(\theta)f_{X\|\Theta}(x\|\theta)}{\sum_{\theta'} p_\Theta(\theta')f_{X\|\Theta}(x\|\theta')}$
Θ continuous, X discrete	$f_{\Theta\|X}(\theta\|x) = \frac{f_\Theta(\theta)p_{X\|\Theta}(x\|\theta)}{\int f_\Theta(\theta')p_{X\|\Theta}(x\|\theta')}$
Θ continuous, X continuous	$f_{\Theta\|X}(\theta\|x) = \frac{f_\Theta(\theta)f_{X\|\Theta}(x\|\theta)}{\int f_\Theta(\theta')f_{X\|\Theta}(x\|\theta')}$

We can use $Z(x)$ to denote the denominator for both discrete and continuous cases. It depends only on the observed data x.

Example (Probability Review). We flip a coin. How likely is it to get 2 heads in 3 coin flips if the probability of heads is p, where p could be 0.5, 0.7, and 1?

Also, use the Central Limit Theorem to estimate the probability of at least 200 heads in 300 coin flips.

Solution:

$$\mathbb{P}(H=2)=\binom{3}{2}p^2(1-p)$$

$p=0.5:\mathbb{P}(H=2)=\binom{3}{2}\times 0.5^2\times 0.5=0.375$

$p=0.7:\mathbb{P}(H=2)=\binom{3}{2}\times 0.7^2\times 0.3=0.441$

$p=1:\mathbb{P}(H=2)=\binom{3}{2}\times 1^2\times 0=0$

For the probability of at least 200 heads in 300 coin-flips,

$$\mathrm{H}\sim\text{Binomial}(300,p),\quad \mu=300p,\quad \sigma=\sqrt{300p(1-p)}$$

$p=0.5:\mu=150,\sigma=8.66$

$$\begin{aligned}\mathbb{P}(H\geq 200)&=\mathbb{P}(\frac{H-150}{8.66}\geq\frac{200-150}{8.66})\\&=\mathbb{P}(z\geq 5.77)\\&\approx 0\end{aligned}$$

$p=0.7:\mu=210,\sigma=7.94$

$$\begin{aligned}\mathbb{P}(H\geq 200)&=\mathbb{P}(\frac{H-210}{7.94}\geq\frac{200-210}{7.94})\\&=\mathbb{P}(z\geq -1.26)\\&=\Phi(1.26)\\&=0.896\end{aligned}$$

Above shows that we have a lower probability for $p=0.5$, which means $p=0.7$ is a better assumption. This is also quite intuitive, since with 200 heads in 300 coin flips, there is a certain probability that the coin is biased.

Again, we flip a coin three times and get two heads. You are told that there are three types of coins with different priors, but you don't know which coin you are flipping. It is obvious that the first coin flip will affect your belief (prior) about which coin you have. For example, if you see 100 heads out of 100 flips, you might strongly believe that both sides of the coin are heads. But to what extent does each flip influence your belief? This brings us to the problem of statistics.

Example. A coin can be one of three types:

1. A fair coin $\theta=1$ with one head and one tail – 90%
2. A coin $\theta=2$ with both sides as heads – 5%
3. A coin $\theta=3$ with both sides as tails – 5%

Now, you flip a head without knowing which coin you have. How should you update your belief (priors)?

Solution:

$$\mathbb{P}(\theta=1|H_1)=\frac{\mathbb{P}(H_1|\theta=1)\mathbb{P}(\theta=1)}{Z(H_1)}=\frac{0.5\times 0.9}{Z(H_1)}=\frac{0.45}{Z(H_1)}$$

$$\mathbb{P}(\theta=2|H_1)=\frac{\mathbb{P}(H_1|\theta=2)\mathbb{P}(\theta=2)}{Z(H_1)}=\frac{1\times 0.05}{Z(H_1)}=\frac{0.05}{Z(H_1)}$$

$$\mathbb{P}(\theta=3|H_1)=0$$

Then we have $\mathbb{P}(H_1) = Z(H_1) = 0.45 + 0.05 + 0 = 0.5$

$$\mathbb{P}(\theta = 1|H_1) = \frac{0.45}{Z(H_1)} = 0.9 \quad \mathbb{P}(\theta = 1|H_1) = \frac{0.05}{Z(H_1)} = 0.1 \quad \mathbb{P}(\theta = 1|H_1) = 0$$

From this, we can update our belief, which we can then use to further readjust our belief if the second flip also results in a head.

$$\mathbb{P}(\theta = 1|H_2H_1) = \frac{\mathbb{P}(H_2|\theta = 1, H_1)\mathbb{P}(\theta = 1|H_1)}{Z(H_2, H_1)} = \frac{0.5 \times 0.9}{Z(H_2, H_1)} = \frac{0.45}{Z(H_2, H_1)}$$

$$\mathbb{P}(\theta = 2|H_2H_1) = \frac{\mathbb{P}(H_2|\theta = 2, H_1)\mathbb{P}(\theta = 2|H_1)}{Z(H_2, H_1)} = \frac{1 \times 0.1}{Z(H_2, H_1)} = \frac{0.1}{Z(H_2, H_1)}$$

$$\mathbb{P}(\theta = 3|H_2H_1) = 0$$

Then we have $\mathbb{P}(H_2H_1) = Z(H_2H_1) = 0.45 + 0.01 + 0 = 0.55$

$$\mathbb{P}(\theta = 1|H_2H_1) = \frac{0.45}{Z(H_2H_1)} = 0.82 \quad \mathbb{P}(\theta = 1|H_2H_1) = \frac{0.1}{Z(H_2H_1)} = 0.18 \quad \mathbb{P}(\theta = 1|H_2H_1) = 0$$

1.2.2 Bayesian Statistical Inference

For Bayesian statistics, we have only one formula: Bayes's rule:

$$\underbrace{f_{\Theta|X}(\theta|x)}_{\text{posterior}} \propto \underbrace{f_{X|\Theta}(x|\theta)}_{\text{likelihood}} \underbrace{f_{\Theta}(\theta)}_{\text{prior}}$$

We have some prior knowledge, and after observing something, we can use the prior (assumption) and likelihood to update our belief, which gives us the posterior. This posterior can later serve as the prior for another observation, allowing us to continuously update our belief throughout the observation process.

Example. Romeo is waiting for Juliet on their first date. He wants to estimate how long he will have to wait for her. Given that Romeo has some prior dating experience, he already has some prior knowledge about how late girls tend to be.

Girl A - $X \sim \text{Uniform}(0, 0.3)$;

Girl B - $X \sim \text{Uniform}(0, 0.8)$;

Girl C - $X \sim \text{Uniform}(0, 0.6)$,

where the uniform random variable shows the range of lateness. For example, for girl A, she will be late between the dating time and the dating time plus 0.3 hours. Then, how could you use Bayesian statistics to estimate the waiting time for Romeo's new girlfriend?

Solution: Here we can set up the uniform random variable $\text{Uniform}(0, \Theta)$, where Θ depends on the girls. Then what we need to find is the θ for Juliet. We can then have

$$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \le x \le \theta; \\ 0, & \text{otherwise.} \end{cases}$$

In Romeo's model, θ is also a uniform random variable $\theta \sim \text{Uniform}(0, 1)$, where $X \sim \text{Uniform}(0, \Theta)$. It means that Romeo has a prior belief that all the girls would be late for at most 1 hour, and the likelihood of Juliet being late is described by X, which states that she could be θ hour late. Given that on their first date, Juliet arrived $\frac{1}{2}$ hours late, we have

$$f_{\Theta|X}(\theta|\frac{1}{2}) \propto f_{\Theta}(\theta) f_{X|\Theta}(\frac{1}{2}|\theta) = \frac{1}{\theta}$$

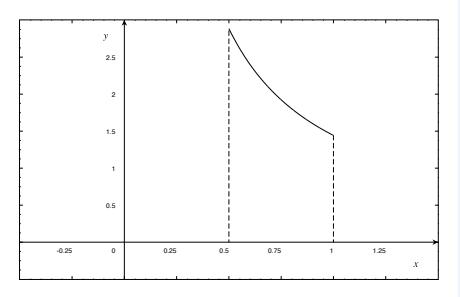
Here we have the prior $f_{\Theta}(\theta) = 1$ if $0 \le \theta \le 1$, and the likelihood $f_{X|\Theta}(\frac{1}{2}|\theta) = \frac{1}{\theta}$ if $\frac{1}{2} \le \theta \le 1$. Keep in mind that the prior comes from Romeo's model, where he has never dated a girl who is

late for more than 1 hour, and it may not be valid if $\theta > 1$, which shows the limitation of Bayesian statistics. Also, the observation (likelihood) shows the probability of Juliet arriving precisely at (or within a very small interval around) time plus 0.5. Therefore, we have $\theta \geq \frac{1}{2}$. Otherwise, if $\theta < \frac{1}{2}$, it is not possible for Juliet to arrive $\frac{1}{2}$ hour late, since it is not included in Romeo's belief.

For the integral to be equal to 1, we need to find the constant term. This can be found using calculus:

$$\int_{\frac{1}{2}}^{1} \frac{1}{\theta} d\theta = \ln\theta\Big|_{\frac{1}{2}}^{1} = \ln 2 \Longrightarrow f_{\Theta|X}(\theta|\frac{1}{2}) = \frac{1}{\theta \ln 2}$$

Here we have $\theta < \frac{1}{2} = 0$ because from the data, we know that $\theta \geq \frac{1}{2}$, which means the lateness parameter is at least $\frac{1}{2}$, so it is not possible for Juliet to arrive between the dating time and dating time plus 0.5. We also have $\theta > 1 = 0$ because from Romeo's prior knowledge, he knows that a girl would not be later than 1 hour.



On their second date, Juliet arrived $\frac{1}{4}$ hours late. We then need to readjust the prior based on the previous model to find the new posterior.

$$f_{\Theta|X_1,X_2}\left(\theta\middle|\frac{1}{2},\frac{1}{4}\right) \propto f_{\Theta|X_1}\left(\theta\middle|\frac{1}{2}\right) f_{X_2|\Theta,X_1}\left(\frac{1}{4}\middle|\theta,\frac{1}{2}\right)$$

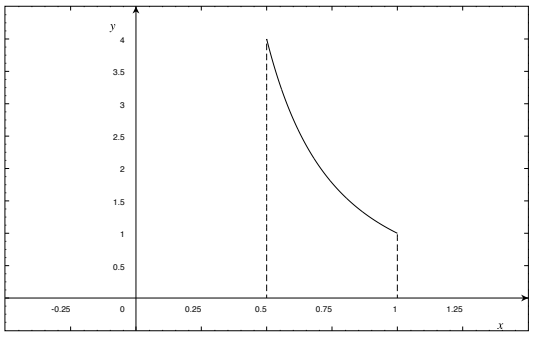
Here, since X_1 and X_2 are independent, we can discard X_1 in the calculation.

$$f_{\Theta|X_1,X_2}\left(\theta\middle|\frac{1}{2},\frac{1}{4}\right) \propto f_{\Theta|X_1}\left(\theta\middle|\frac{1}{2}\right) f_{X_2|\Theta}\left(\frac{1}{4}\middle|\theta\right) = \frac{1}{\theta \ln 2} \times \frac{1}{\theta} = \frac{1}{\theta^2 \ln(2)} \propto \frac{1}{\theta^2}$$

The same as above, we have $f_{X_2|\Theta}(\frac{1}{4}|\theta) = \frac{1}{\theta}$ for $\theta \geq \frac{1}{4}$ since it is not possible for the lateness to be less than $\frac{1}{4}$ hours. Also, given the prior as calculated in the first part, we have $f_{\Theta|X_1}(\theta|\frac{1}{2}) = \frac{1}{\theta \ln 2}$ if $\frac{1}{2} \leq \theta \leq 1$.

For the integral to be equal to 1, we need to find the constant term. This can be found using calculus:

$$\int_{\frac{1}{2}}^{1} \frac{1}{\theta^2} d\theta = 1 \Longrightarrow f_{\Theta|X_1,X_2}\left(\theta\middle|\frac{1}{2},\frac{1}{4}\right) = \frac{1}{\theta^2}$$



Remark (Bayes' rule variant).

$$\mathbb{P}(\theta|x_1,x_2) = \frac{\mathbb{P}(x_2|\theta,x_1)\mathbb{P}(\theta|x_1)}{\mathbb{P}(x_2|x_1)}$$

Proof.

$$\begin{aligned}
f_{\Theta|X_1,X_2}\left(\theta\middle|\frac{1}{2},\frac{1}{4}\right) &= \frac{f_{\Theta,X_1,X_2}\left(\theta,\frac{1}{2},\frac{1}{4}\right)}{f_{X_1,X_2}\left(\frac{1}{2},\frac{1}{4}\right)} \\
&= \frac{f_{X_2|\Theta,X_1}\left(\frac{1}{4}\middle|\theta,\frac{1}{2}\right) f_{\Theta,X_1}\left(\theta,\frac{1}{2}\right)}{f_{X_1,X_2}\left(\frac{1}{2},\frac{1}{4}\right)} \\
&= \frac{f_{X_2|\Theta,X_1}\left(\frac{1}{4}\middle|\theta,\frac{1}{2}\right) f_{\Theta|X_1}\left(\theta\middle|\frac{1}{2}\right) f_{X_1}\left(\frac{1}{2}\right)}{f_{X_1,X_2}\left(\frac{1}{2},\frac{1}{4}\right)} \\
&= \frac{f_{X_2|\Theta,X_1}\left(\frac{1}{4}\middle|\theta,\frac{1}{2}\right) f_{\Theta|X_1}\left(\theta\middle|\frac{1}{2}\right)}{f_{X_2|X_1}\left(\frac{1}{4}\middle|\frac{1}{2}\right)}
\end{aligned}$$

Thus,

$$f_{\Theta|X_1,X_2}\left(\theta\middle|\frac{1}{2},\frac{1}{4}\right) \propto f_{X_2|\Theta,X_1}\left(\frac{1}{4}\middle|\theta,\frac{1}{2}\right) f_{\Theta|X_1}\left(\theta\middle|\frac{1}{2}\right)$$

■

Now it's a bit tedious since we need to perform calculations and adjust our prior each time we obtain new data or observations. However, we also have Bayes's rule for multiple random variables, which simplifies the process.

$$\begin{aligned}
f_{\Theta|X_1,\cdots,X_n}(\theta|x_1,\cdots,x_n) &= \frac{f_{X_1,\cdots,X_n|\Theta}(x_1,\cdots,x_n|\theta) f_\Theta(\theta)}{Z(x_1,\cdots,x_n)} \\
&\propto f_{X_1,\cdots,X_n|\Theta}(x_1,\cdots,x_n|\theta) f_\Theta(\theta) \\
&= \underbrace{f_{X_1|\Theta}(x_1|\theta)\cdots f_{X_n|\Theta(x_n|\theta)}}_{\text{product of likelihood}} \underbrace{f_\Theta(\theta)}_{\text{prior}}
\end{aligned}$$

if $X_1,\cdots,X_n$ are independent given Θ.

Example (Cont'd). Given that Juliet is late by $\frac{1}{4}$ hours on their third date, how do we find the posterior?

Solution:

$$f_{\Theta|X_1,X_2,X_3}\left(\theta\middle|\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) \propto f_{X_1|\Theta}\left(\frac{1}{2}\middle|\theta\right) f_{X_2|\Theta}\left(\frac{1}{4}\middle|\theta\right) f_{X_3|\Theta}\left(\frac{1}{4}\middle|\theta\right) f_\Theta(\theta) = \frac{1}{\theta^3}$$

For $f_{X_1|\Theta}, f_{X_2|\Theta}, f_{X_3|\Theta}$, they are all equal to $\frac{1}{\theta}$ for $\theta \geq \frac{1}{2}$ and $\theta \geq \frac{1}{4}$ for the same reason shown before. We also have $f_\Theta(\theta) = 1$ if $0 \leq \theta \leq 1$. Taking the intersection, we obtain $\frac{1}{\theta^3}$ for $\frac{1}{2} \leq \theta \leq 1$. For the integral to be equal to 1, we need to determine the constant term, which can be found using calculus.

$$\int_{\frac{1}{2}}^{1} \frac{1}{\theta^2} d\theta = \frac{3}{2} \Longrightarrow f_{\Theta|X_1,X_2,X_3}\left(\theta\middle|\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) = \frac{2}{3\theta^3}$$

Example (Biased Coin). A coin of unknown bias flips HTT. What is the bias?

Solution: Let $X \sim \text{Bernoulli}(\Theta)$, where $\Theta = \mathbb{P}(X = H)$. We have a prior $\Theta \sim \text{Uniform}(0, 1)$. To find the posterior (bias), we have:

$$\begin{aligned} f_{\Theta|X_1,X_2,X_3}(\theta|H,H,T) &\propto p_{X_1|\Theta}(H|\theta)p_{X_2|\Theta}(T|\theta)p_{X_3|\Theta}(T|\theta)f_\Theta(\theta) \\ &= \theta(1-\theta)(1-\theta) \times 1 \\ &= \theta(1-\theta)^2 \\ \implies f_{\Theta|X_1,X_2,X_3}(\theta|H,H,T) &= \frac{\theta(1-\theta)^2}{\int_0^1 \theta(1-\theta)^2 d\theta} = 12\theta(1-\theta)^2 \end{aligned}$$

To find the posterior, we often need to find the denominator $Z(x)$, which requires some calculus techniques and can sometimes be difficult to solve. However, there are some techniques that come in handy.

1.3 Conjugate Priors

Definition 1.3.1 (Conjugate Priors). The posterior distribution $f_{\Theta|X}(\theta|x)$ is in the same probability distribution family as the prior distribution $f_\Theta(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $f_{X|\Theta}(x|\theta)$.

There are four types of conjugate priors to consider.

1.3.1 Conjugate Prior for Bernoulli

Definition 1.3.2. Suppose $X_1, \cdots, X_n$ form a random sample from Bernoulli distribution with an unknown parameter θ $(0 < \theta < 1)$. If the prior distribution $f_\Theta(\theta)$ is the Beta distribution $\text{Beta}(\alpha, \beta)$ $(\alpha, \beta > 0)$, then the posterior distribution $f_{\Theta|X}(\theta|x)$ given $\{X_i = x_i\}_i = 1^n$ is the Beta distribution $\text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$.

Here we introduce the Beta random variable. It has the PDF as follows:

$$f_\Theta(\theta) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} & \text{for } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases},$$

where

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx = (\alpha-1)! \text{ (for positive integer} \alpha)$$

or equivalently,

$$B(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

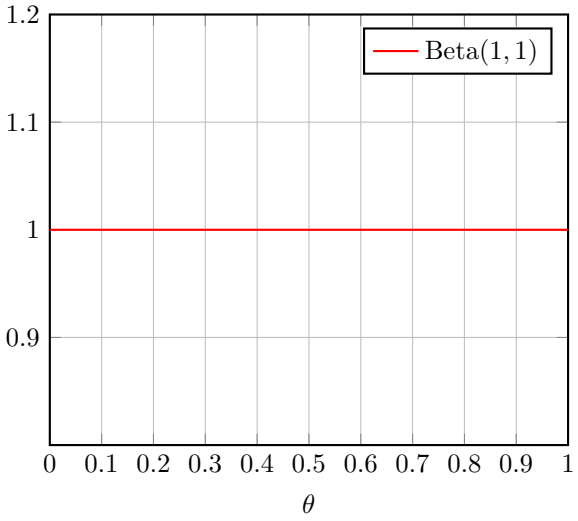
The reason why $B(\alpha, \beta)$ appears in the denominator of the PDF is that it serves as the normalization constant, ensuring that the integral equals 1 so that it is a valid PDF.

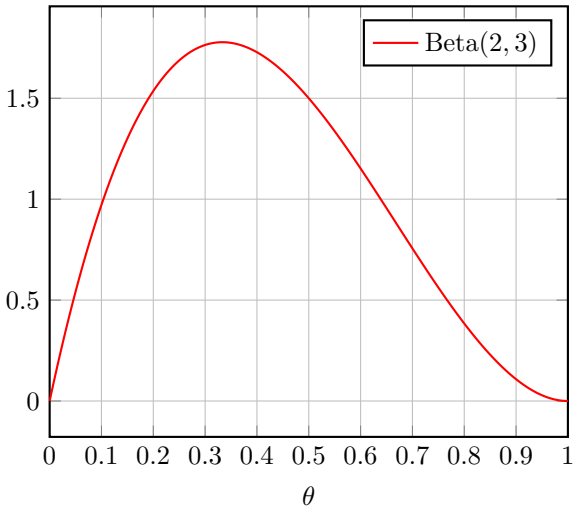
The Beta random variable is widely used to model the prior distribution of a random variable which range is $[0, 1]$, where α and β are hyperparameter.

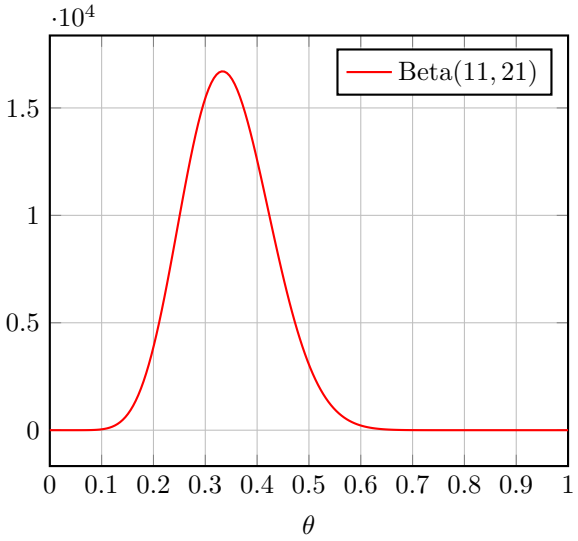
Recalling the coin flip example above, with the prior Θ and observation X remaining unchanged, we can use the Beta distribution to perform the calculation. We have $\Theta \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1)$, and for $h = 1, t = 2$, we have:

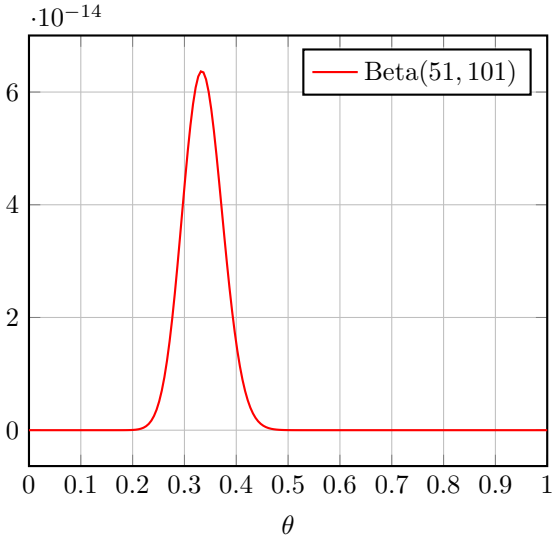
$$f_{\Theta|X_1,X_2,X_3}(\theta|H,H,T) = \frac{1}{\text{Beta}(h+1,t+1)}\theta^{2-1}(1-\theta)^{3-1} = 12\theta(1-\theta)^2$$

In general, for a coin of unknown bias flips n times and gets h heads and $(n-h)$ tails (or t tails), we can have prior of $\Theta \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1)$, and $(\theta|h \text{ heads}, t \text{ tails}) \sim \text{Beta}(h + 1, t + 1)$.







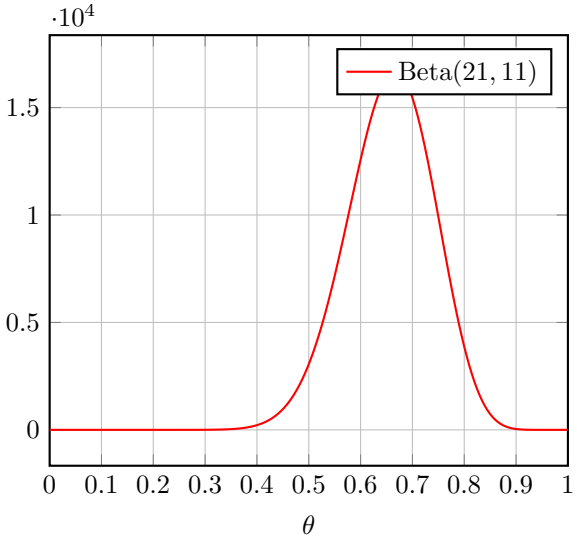


The above shows that we can perform estimation based on the number of experiments, which will result in a different PDF. With more data in hand, the accuracy of the data is higher. However, they share a common feature: the value at which the PDF or PMF reaches its maximum is

$$\text{mode}[\theta] = \frac{\alpha - 1}{\alpha - 1 + \beta - 1} \quad \text{when } \alpha, \beta > 1$$

Also, we can treat the different parameters as a change in belief. For example, if Beta(2, 3) is our prior, and we readjust our belief based on observations, we then obtain Beta(21, 11). This shows that the area below the original mode $\frac{1}{3}$ decreases, making it less probable.

The last thing to note is that hyperparameter, in the coin flip case, h, t, don't matter if we observe a large number of data samples, meaning the posterior mainly depends on the observed data. However, if the prior contains a large dataset or the size of the observed data is small, then the prior plays an important role in the posterior.



1.3.2 Conjugate Prior for Poisson

Definition 1.3.3. Suppose $X_1, \cdots, X_n$ form a random sample from Poisson distribution with an unknown mean $\Theta > 0$. If the prior distribution $f_\Theta(\theta)$ is the Gamma distribution Gamma(α, β) $(\alpha, \beta > 0)$, then the posterior distribution $f_{\Theta|X}(\theta|x)$ given $\{X_i = x_i\}_{i=1}^n$ is the Gamma distribution Gamma$(\alpha + \sum_{i=1}^n x_i, \beta + n)$.

Here we introduce another random variable that is often used as prior, Gamma random variable. It has the PDF as follows:

$$f_\Theta(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} & \text{for } \theta > 0 \\ 0 & \text{for } \theta \leq 0 \end{cases},$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx = (\alpha - 1)! \text{ (for positive integer } \alpha).$$

Again, we have the Gamma random variable as the denominator because the integral needs to be equal to 1.

Example. At an Apple Store, the number of iPhones sold per day is modeled as a Poisson distribution with unknown mean Θ. Suppose the prior distribution of Θ is Gamma(3, 2). Let X be the number of iPhones sold in a specific day. If $X = 3$ is observed, what is the updated distribution of θ?

Solution: Here we have

$$X \sim \text{Poisson}(\Theta) = \begin{cases} \frac{e^{-\theta}\theta^x}{x!} & \text{for } x = 0, 1, 2 \dots \\ 0 & \text{otherwise} \end{cases};$$

$$\Theta \sim \text{Gamma}(\alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} & \text{for } \theta > 0 \\ 0 & \text{for } \theta \leq 0 \end{cases}.$$

Since we have observed $X = 3$,

$$f_{\Theta|X}(\theta|3) \propto f_\Theta(\theta) f_{X|\Theta}(3|\theta)$$

where

$$f_\Theta(\theta) = \text{Gamma}(3, 2) = \frac{2^3}{2!}\theta^{3-1}e^{-2\theta}, f_{X|\Theta}(3|\theta) = \text{Poisson}(\theta) = \frac{e^{-\theta}\theta^3}{3!}.$$

Then we have

$$f_{\Theta|X}(\theta|3) \propto f_\Theta(\theta) f_{X|\Theta}(3|\theta) = \frac{2^2}{3!}\theta^5 e^{-3\theta} \propto \theta^5 e^{-3\theta}$$

$$f_{\Theta|X}(\theta|3) = \frac{\theta^{6-1}e^{-3\theta}}{Z}, \quad Z = \int_0^\infty \theta^{6-1}e^{-3\theta}d\theta = \frac{\Gamma(6)}{3^6}$$

Finally, we have the posterior

$$f_{\Theta|X}(\theta|3) = \text{Gamma}(6, 3).$$

Above is the same as taking $\alpha = 3, \beta = 2, n = 1$ and $x = 3$, then we have $\alpha + x = 6$, $\beta + n = 3$. This directly gives us Gamma(6, 3).

1.3.3 Conjugate Prior for Exponential

Definition 1.3.4. Suppose $X_1, \cdots, X_n$ form a random sample from Exponential distribution with an unknown parameter $\theta > 0$. If the prior distribution $f_\Theta(\theta)$ is the Gamma distribution Gamma(α, β) $(\alpha, \beta > 0)$, then the posterior distribution $f_{\Theta|X}(\theta|x)$ given $\{X_i = x_i\}_{i=1}^n$ is the Gamma distribution Gamma$(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

In the case of Exponential prior, we have $\alpha =$ no. of trials + 1, $\beta =$ sum of data + prior.

Example. If the number of iPhones sold per hour follows a Poisson distribution with unknown mean Θ, then the time between two successive iPhones sold follow an exponential distribution with parameter Θ. Suppose the prior distribution of Θ is Gamma(1, 2). Let X be the time interval (in hour) between successive iPhones sold.

Assume that we have $X_1 = 1.5, X_2 = 2, X_3 = 2.5$.

Solution: Here we have

$$X \sim \text{Exponential}(\Theta) = \begin{cases} \theta e^{-\theta x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases};$$

$$\Theta \sim \text{Gamma}(1, 2).$$

Since we have observed X_1, X_2, X_3,

$$f_{\Theta|X_1,X_2,X_3}(\theta|1.5, 2, 2.5) \propto f_\Theta(\theta) f_{X_1,X_2,X_3|\Theta}(1.5, 2, 2.5|\theta)$$

where

$$f_\Theta(\theta) = \text{Gamma}(1, 2) = \frac{2^1}{1!}\theta^{1-1}e^{-2\theta}, f_{X_1,X_2,X_3|\Theta}(1.5, 2, 2.5|\theta) = (\theta e^{-1.5\theta})(\theta e^{-2\theta})(\theta e^{-2.5\theta}).$$

Then we have

$$f_{\Theta|X_1,X_2,X_3}(\theta|1.5, 2, 2.5) \propto f_\Theta(\theta) f_{X_1,X_2,X_3|\Theta}(1.5, 2, 2.5|\theta) = 2\theta^3 e^{-(2+6)\theta} \propto \theta^3 e^{-(2+6)\theta}$$

$$f_{\Theta|X}(\theta|3) = \frac{\theta^3 e^{-(2+6)\theta}}{Z}, \quad Z = \int_0^\infty \theta^3 e^{-(2+6)\theta} d\theta = \frac{\Gamma(4)}{8^4}$$

Finally, we have the posterior

$$f_{\Theta|X}(\theta|3) = \text{Gamma}(4, 8).$$

Above is the same as taking $\alpha = 1, \beta = 2, n = 3, x_1 = 1.5, x_2 = 2$ and $x_3 = 2.5$, then we have $\alpha + n =$ no. of trials $+ 1 = 3 + 1 = 4$, $\beta + n =$ sum of trials + prior $= 6 + 2 = 8$. This directly gives us Gamma(4, 8).

1.3.4 Conjugate Prior for Normal Distribution

Definition 1.3.5. Suppose $X_1, \cdots, X_n$ form a random sample from a normal distribution with an unknown mean μ and a known variance $\sigma^2 > 0$. If the prior distribution $f_\Theta(\mu)$ is the normal distribution $\mathcal{N}(\mu, \sigma_0^2)$, then the posterior distribution $f_{\Theta|X}(\mu|x)$ given $\{X_i = x_i\}_{i=1}^n$ is the normal distribution $\mathcal{N}(\mu', \sigma'^2)$, where

$$\mu' = \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{i=1}^n x_i}{\sigma^2 + n\sigma_0^2} \qquad \sigma'^2 = \frac{\sigma^2\sigma_o^2}{\sigma^2 + n\sigma_o^2}$$

Definition 1.3.6 (A more general case). Suppose $X_1, \cdots, X_n$ form a random sample from a normal distribution with a common unknown mean θ and the known variance $\sigma_i^2 > 0$. If the prior distribution $f_\Theta(\theta)$ is the normal distribution $\mathcal{N}(\mu_0, \sigma_0^2)$, then the posterior distribution $f_{\Theta|X}(\theta|x)$ given that $\{X_i = x_i\}_{i=1}^n$ is the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where

$$\frac{\mu}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{x_1}{\sigma_1^2} + \cdots + \frac{x_n}{\sigma_n^2} \qquad \frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2} + \cdots + \frac{1}{\sigma_n^2}$$

Here we need to consider a special case when both σ_0^2 and σ^2 are equal to 1, then we have

$$\mu' = \frac{\mu_0 + \sum_{i=1}^n x_i}{1 + n} \qquad \sigma'^2 = \frac{1}{1 + n}$$

Example. An $\mathcal{N}(\Theta, 1)$ random variable takes value 3.97. Θ follows a standard normal. What is the posterior of Θ?

Solution: Here we have the PDF of $\mathcal{N}(\mu, \sigma^2)$

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

Given that the prior $= \Theta \sim \mathcal{N}(0,1)$, posterior $= f_{\Theta|X}(\theta|x) \propto f_\Theta(\theta) f_{X|\Theta}(x|\theta)$, we have

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} \qquad f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$$

$$\begin{aligned} f_{\Theta|X}(\theta|x) \propto f_\Theta(\theta) f_{X|\Theta}(x|\theta) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \\ &\propto e^{-\frac{1}{2}\theta^2} \times e^{-\frac{1}{2}(x-\theta)^2} \\ &= e^{-\frac{1}{2}\theta^2 - \frac{1}{2}(x-\theta)^2} \\ &= e^{-(\sqrt{2}\theta - \frac{1}{\sqrt{2}}x)^2} \underbrace{e^{-\frac{x^2}{4}}}_{\text{constant term}} \\ &\propto e^{-(\sqrt{2}\theta - \frac{1}{\sqrt{2}}x)^2} \\ &= e^{-\frac{1}{2}\frac{(\theta - \frac{x}{2})^2}{(\frac{1}{\sqrt{2}})^2}} \end{aligned}$$

Then we have

$$\mu = \frac{x}{2} = \frac{3.97}{2} = 1.985 \qquad \sigma^2 = (\frac{1}{\sqrt{2}})^2 = \frac{1}{2}$$

Finally, we have the posterior

$$f_{\Theta|X}(\theta|3) = \mathcal{N}(1.985, \frac{1}{2})$$

Above is the same as taking $\mu_0 = 0, x_1 = 3.97, \sigma_0 = 1$ and $\sigma_1 = 1$, then we have

$$\frac{1}{\sigma^2} = \frac{1}{1} + \frac{1}{1} \Longrightarrow \sigma = \frac{1}{\sqrt{2}} \qquad \frac{\mu}{\frac{1}{2}} = \frac{0}{1} + \frac{3.97}{1} \Longrightarrow \mu = 1.985,$$

which directly gives us $\mathcal{N}(1.985, \frac{1}{2})$.

When $\sigma_0 = \sigma_1 = \cdots = 1$, we can find σ and μ by:

$$\sigma = \frac{1}{\sqrt{n+1}}, \quad \mu = \frac{x_0 + x_1 + \cdots + x_n}{n+1}$$

Example. Three independent $\mathcal{N}(\Theta, 1)$ random variables take values 3.97, 4.09, 3.11. What is Θ?

Solution: Here we assume the priors are $\Theta \sim \mathcal{N}(0,1)$, and from observation we have $x_1 = 3.97, x_2 = 4.09, x_3 = 3.11$.

Then, for the posterior, we have

$$f_{\Theta|X_1,X_2,X_3}(\theta|x_1,x_2,x_3) \sim \mathcal{N}\left(\frac{0 + 3.97 + 4.09 + 3.11}{1+3}, \left(\frac{1}{\sqrt{1+3}}\right)^2\right) \approx \mathcal{N}(2.79, \frac{1}{4})$$

1.4 Applications of Bayesian Statistic

In this section, we will study the use of Bayesian Statistics.

To begin with, think about the coin flips event. Assume that you have observed some data, i.e., the first 10 coin flips give the sequence `H T T H T T H T T T`. You now have the model; then, what can it

be used for? It turns out that we can use it to make predictions, which tell the probability of the next flip being a head. We can also use it to do estimation, such as determining the probability of heads for this coin. Additionally, we can perform something called hypothesis testing, which helps us find the best guess for the estimation.

1.4.1 Prediction

Let's revisit the previous dating scenario.

Example. On her first date, Juliet arrives $\frac{1}{2}$ hour late. How likely is she to arrive more than $\frac{3}{4}$ hour late next time?

Solution: Let $X_1, X_2 \sim \text{Uniform}(0, \Theta)$, where $\Theta = \text{Uniform}(0, 1)$. From the posterior that we calculated before, we have

$$f_{\Theta|X}(\theta|\frac{1}{2}) = \begin{cases} \dfrac{1}{\theta \ln 2} & \text{if } \dfrac{1}{2} \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We can then use this posterior to make predictions.

$$\mathbb{P}(X_2 \ge \frac{3}{4}|X_1 = \frac{1}{2}) = \int_{-\infty}^{+\infty} \underbrace{\mathbb{P}\left(X_2 \ge \frac{3}{4}|X_1 = \frac{1}{2}, \Theta = \theta\right) \mathbb{P}\left(\theta|X_1 = \frac{1}{2}\right)}_{\text{Total Probability Theorems}} d\theta$$

$$(*) = \int_{\frac{1}{2}}^{1} \mathbb{P}\left(X_2 \ge \frac{3}{4}|X_1 = \frac{1}{2}, \Theta = \theta\right) f_{\Theta|X}(\theta|\frac{1}{2}) d\theta$$

$$(**) = \int_{\frac{3}{4}}^{1} \mathbb{P}\left(X_2 \ge \frac{3}{4}|\Theta = \theta\right) f_{\Theta|X}(\theta|\frac{1}{2}) d\theta$$

$$(***) = \int_{\frac{3}{4}}^{1} (\theta - \frac{3}{4}) \frac{1}{\theta} \frac{1}{\theta \ln 2} d\theta$$

$$= \int_{\frac{3}{4}}^{1} \frac{1}{\theta \ln(2)} d\theta - \int_{\frac{3}{4}}^{1} \frac{3}{4\theta^2 \ln(2)} d\theta$$

$$= \frac{\ln \frac{4}{3} - \frac{1}{4}}{\ln 2}$$

$$= 0.054$$

In (*), we change the lower boundary from $-\infty$ to $\frac{1}{2}$ and the upper boundary from $+\infty$ to 1 because $f_{\Theta|X}(\theta|\frac{1}{2})$ would be 0 outside $[\frac{1}{2}, 1]$. Then, in (**), we again update the lower boundary to $\frac{3}{4}$ because for $\frac{1}{2} \le \theta \le \frac{3}{4}$, $\mathbb{P}(X_2 \ge \frac{3}{4}|\Theta = \theta)$ would be equal to 0. In (***), we can directly find the left-hand side by $(\theta - \frac{3}{4})\frac{1}{\theta}$ because $X_2 \sim \text{Uniform}(0, \theta)$. The PDF can be directly computed by finding the area.

Remark. One may start with

$$\int_{-\infty}^{+\infty} \mathbb{P}\left(X_2 \ge \frac{3}{4}, \Theta = \theta|X_1 = \frac{1}{2}\right) d\theta$$

where

$$\begin{aligned}\mathbb{P}\left(X_2 \geq \frac{3}{4}, \Theta = \theta | X_1 = \frac{1}{2}\right) &= \frac{\mathbb{P}\left(X_2 \geq \frac{3}{4}, \Theta = \theta, X_1 = \frac{1}{2}\right)}{\mathbb{P}\left(X_1 = \frac{1}{2}\right)} \\ &= \frac{\mathbb{P}\left(X_2 \geq \frac{3}{4} | X_1 = \frac{1}{2}, \Theta = \theta\right) \mathbb{P}\left(X_1 = \frac{1}{2}, \Theta = \theta\right)}{\mathbb{P}\left(X_1 = \frac{1}{2}\right)} \\ &= \mathbb{P}\left(X_2 \geq \frac{3}{4} | X_1 = \frac{1}{2}, \Theta = \theta\right) \mathbb{P}\left(\theta | X_1 = \frac{1}{2}\right)\end{aligned}$$

If we have past data and a prior distribution, we can often make predictions.

Example. Assume that we have observed n heads in coin flips. What is the probability that the next coin flip will also be a head?

Solution: For coin flips, we can use $X \sim \text{Bernoulli}(\Theta)$, where $\Theta = \mathbb{P}(X = H)$. So for the prior, we have $\Theta \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1)$. Since the prior follows a beta distribution, the posterior also follows a beta distribution. Therefore, the posterior is given by:

$$\Theta | n \text{ Heads} \sim \text{Beta}(n + 1, 1)$$

$$f_{\Theta|X_1,\cdots,X_n}(\theta|nH) = \frac{(n+1)!}{n!1!}\theta^n = (n+1)\theta^n$$

We then use this posterior to update our belief, making it the prior for predicting whether the next coin flip will be heads.

$$\begin{aligned}\mathbb{P}(H^*|nH) &= \int_0^1 \mathbb{P}(H^*|\theta) f_{\Theta|X_1,\cdots,X_n}(\theta|nH) d\Theta \\ &= \int_0^1 \theta(n+1)\theta^n d\theta \\ &= \frac{n+1}{n+2}\end{aligned}$$

For example, if we have previously flipped $n = 100$ heads, the probability of the next coin flip being heads is $\frac{101}{102}$.

To summary, in Bayesian prediction, for observation $X = x$ (past data), if X is continuous, to predict $x^* \in [a, b]$

$$\mathbb{P}(x^* \in [a, b] | X = x) = \int_{-\infty}^{+\infty} \mathbb{P}(x^* \in [a, b] | \theta) \underbrace{f_{\Theta|X}(\theta|x)}_{\text{prior}} d\theta$$

where

$$\mathbb{P}(x^* \in [a, b] | \theta) = \int_a^b f_{X|\Theta}(x^*|\theta) dx^*.$$

If X is discrete, then to predict x^*

$$\mathbb{P}(x^*|X = x) = \int_{-\infty}^{+\infty} \mathbb{P}(x^*|\theta) f_{\Theta|X}(\theta|x) d\theta$$

1.4.2 Point Estimation

The question then arises: how do we turn the conditional PDF or PMF $f_{\Theta|X}(\theta|x)$ estimate into a single number? Or, to put it simply, how do we find the θ that is the best estimate of the parameter from the posterior? It turns out we have two methods, namely the Maximum a Posterior (MAP) estimator and the Conditional Expectation (CE) estimator.

For MAP, we find the most likely value:

$$\theta_{\text{MAP}} = \arg\max_{\theta} f_{\Theta|X}(\theta|x).$$

For CE, we find the average among all possible θ, and the expectation $\mu = \mathbb{E}[\Theta]$ will minimize the mean square error $\mathbb{E}[(\Theta - \theta)^2]$:

$$\mathbb{E}[\Theta|X = x].$$

To illustrate, let's return to the dating problem again.

Example. In Romeo's model, on their first date, Juliet arrived $\frac{1}{2}$ hour late. What would be his estimate for the probability of Juliet being late?

Solution:

MAP (optimistic method)

$$\text{Posterior } f_{\Theta|X}(\theta|\frac{1}{2}) = \frac{1}{\theta \ln 2} \quad \text{when } \frac{1}{2} \le \theta \le 1 \Longrightarrow \arg\max_{\theta} \frac{1}{\theta \ln 2} = \arg\max_{\theta} \frac{1}{\theta}$$

which gives

$$\theta_{\text{MAP}} = \frac{1}{2} \qquad \text{refers to the graph}$$

CE (conservative method)

$$\mathbb{E}[\Theta|X_1 = \frac{1}{2}] = \int_{\frac{1}{2}}^{1} \theta \frac{1}{\theta \ln 2} d\theta = \frac{1}{2 \ln 2} \approx 0.72$$

Remark. Note that prediction refers to forecasting the future value, while estimation involves calculating the likely value of a parameter based on samples.

Here we have two special cases:

1. Point estimation for a Beta random variable.

Given that the prior is $\Theta \sim \text{Beta}(1, 1)$, and the posterior is $\Theta|h \text{ Heads}, t \text{ Tails} \sim \text{Beta}(1+h, 1+t)$, where $\alpha = h + 1, \beta = t + 1$, we have:

$$\text{mode}[\text{Beta}(\alpha, \beta)] : \theta = \frac{\alpha - 1}{\alpha - 1 + \beta - 1} \text{ when } \alpha, \beta > 1.$$

$$\theta_{MAP} = \frac{\alpha - 1}{\alpha - 1 + \beta - 1} = \frac{h}{h + t}$$

$$\text{CE} = \frac{\alpha}{\alpha + \beta}$$

As the number of data points increases, the difference between MAP and CE will become smaller, and we will obtain a closer value.

2. Point estimation for Normal random variable.

Given that the prior is $\Theta \sim \mathcal{N}(\mu_0, 1)$, and the posterior is $\Theta|X_1, \cdots, X_n \sim \mathcal{N}(\frac{\mu_0 + x_1 + \cdots + x_n}{n+1}, \frac{1}{n+1})$, we have

$$\text{mode}[\mathcal{N}(\mu, \sigma^2)] : \theta = \mu$$

$$\theta_{MAP} = \frac{\mu_0 + x_1 + \cdots + x_n}{n + 1}$$

$$\text{CE} = \mathbb{E}[\mathcal{N}(\mu, \sigma^2)] = x$$

1.4.3 Hypothesis Testing

Suppose that in a hypothesis testing problem, Θ takes m values $\theta_1, \cdots, \theta_m$. Recall that in hypothesis testing, we want to find the best guess for the decision or classification, i.e., checking how likely the estimated parameter is to be the actual one given the observed data. Then, how do we choose the one for which $f_{\Theta|X}(\theta_i|x)$ is the largest (best guess), so that we have the optimal hypothesis θ?

Example (Estimation).

Now, you receive an email. It could be spam or legitimate, with $\Theta = 1$ indicating spam with a 20% chance, and $\Theta = 0$ indicating legit with an 80% chance. Suppose there are two patterns, X_1 and X_2, which are independent given a specific email, to classify whether the email is spam or legit.

Θ	$\mathbb{P}(X_1 = 1\|\theta)$	$\mathbb{P}(X_2 = 1\|\theta)$
$\Theta = 0$ legit	0.03	0.0001
$\Theta = 1$ spam	0.1	0.01

Then, in a specific email x, observe that $X_1 = 1$ and $X_2 = 0$. Is it spam or legitimate?

Solution:

$$\mathbb{P}(\Theta = 1|X_1 = 1, X_2 = 0) \propto \mathbb{P}(X_1 = 1, X_2 = 0|\Theta = 1)\mathbb{P}(\Theta = 1) = 0.1 \times 0.99 \times 0.2 \approx 0.0198$$

$$\mathbb{P}(\Theta = 0|X_1 = 1, X_2 = 0) \propto \mathbb{P}(X_1 = 1, X_2 = 0|\Theta = 0)\mathbb{P}(\Theta = 0) = 0.03 \times 0.9900 \times 0.8 \approx 0.0240$$

Thus, MAP $\Theta = 0$, shows that the email is legitimate.

Example (Hypothesis testing).

We have two coins, A and B. Coin A has a $\frac{2}{3}$ probability of landing heads, and coin B has a $\frac{2}{3}$ probability of landing tails. You flip a random coin and observe the sequence `H H T`. Which coin did you flip? What is the probability that you are wrong based on MAP, given the outcome is `H H T`?

Solution:

Since we have equally likely prior $\mathbb{P}(\Theta = A) = \mathbb{P}(\Theta = B) = 50\%$,

$$\mathbb{P}(\Theta = A|HHT) \propto \mathbb{P}(HHT|\Theta = A)\mathbb{P}(\Theta = A) = \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{2} = \frac{2}{27}$$
$$\mathbb{P}(\Theta = B|HHT) \propto \mathbb{P}(HHT|\Theta = B)\mathbb{P}(\Theta = B) = \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{27}$$

Thus, MAP $\Theta = A$.

$$\begin{aligned}\text{error} &= \mathbb{P}(B|HHT) \\ &= \frac{\mathbb{P}(HHT|\Theta = B)\mathbb{P}(\Theta = B)}{\mathbb{P}(HHT)} \\ &= \frac{\frac{1}{27}}{\frac{1}{27} + \frac{2}{27}} = \frac{1}{3}\end{aligned}$$

This shows that the event would be wrong at $\frac{1}{3}$ of the time.

We find the probability that, even if the calculation is correct, it is still possible for us to make a wrong guess from time to time. But then, what is the probability of being wrong on average?

Example. What is the probability that you are wrong on average based on the MAP estimate given the outcome of 3 flips?

Solution:

$$\begin{aligned}\mathbb{P}(\theta_{\text{MAP}} \neq \theta) &= \mathbb{P}(\theta_{\text{MAP}} = B, \theta = A) + \mathbb{P}(\theta_{\text{MAP}} = A, \theta = B) \\ &= \mathbb{P}(\theta_{\text{MAP}} = B | \theta = A)\mathbb{P}(\theta = A) + \mathbb{P}(\theta_{\text{MAP}} = A | \theta = B)\mathbb{P}(\theta = B)\end{aligned}$$

We can find the probability of the outcome given the coin type, which we used to find θ_{MAP}.

For example,

$$p_{3\text{H}|\theta=A} = \binom{3}{3}(\frac{2}{3})^3(1-\frac{2}{3})^0 = \frac{8}{27}; \qquad p_{2\text{H1T}|\theta=A} = \binom{3}{2}(\frac{2}{3})^2(1-\frac{2}{3})^1 = \frac{12}{27}$$

Then we have

Outcome	3H	2H1T	1H2T	3T
θ_{MAP}	A	A	B	B
$p_{\text{outcome}\|\theta=A}$	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$
$p_{\text{outcome}\|\theta=B}$	$\frac{1}{27}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{8}{27}$

Now we can find the probability of being wrong on average.

$$\begin{aligned}\mathbb{P}(\theta_{\text{MAP}} \neq \theta) &= \mathbb{P}(\theta_{\text{MAP}} = B | \theta = A)\mathbb{P}(\theta = A) + \mathbb{P}(\theta_{\text{MAP}} = A | \theta = B)\mathbb{P}(\theta = B) \\ &= (\mathbb{P}(1\text{H2T}|\theta = A) + \mathbb{P}(3\text{T}|\theta = A))\mathbb{P}(\theta = A) \\ &\quad + (\mathbb{P}(2\text{H1T}|\theta = B) + \mathbb{P}(3\text{H}|\theta = B))\mathbb{P}(\theta = B) \\ &= \left(\frac{6}{27} + \frac{1}{27}\right) \times \frac{1}{2} + \left(\frac{6}{27} + \frac{1}{27}\right) \times \frac{1}{2} \\ &= \frac{7}{27}\end{aligned}$$

For binary hypothesis testing error, we have $\theta = 0$ (negative) or $\theta = 1$ (positive), which represent the true state. Similarly, we have $\hat{\theta} = 0$ (negative) or $\hat{\theta} = 1$ (positive), which represent the estimated state. Then, $\mathbb{P}(\hat{\theta} = 1, \theta = 0)$ represents a false positive, and $\mathbb{P}(\hat{\theta} = 0, \theta = 1)$ represents a false negative. For the calculation, we can then simply use

$$\begin{aligned}\mathbb{P}(\hat{\theta} \neq \theta) &= \mathbb{P}(\hat{\theta} = 1, \theta = 0) + \mathbb{P}(\hat{\theta} = 0, \theta = 1) \\ &= \mathbb{P}(\hat{\theta} = 1 | \theta = 0)\mathbb{P}(\theta = 0) + \mathbb{P}(\hat{\theta} = 0 | \theta = 1)\mathbb{P}(\theta = 1)\end{aligned}$$

Example. A car-jack detector X outputs $\mathcal{N}(0, 1)$ if there is no intruder and $\mathcal{N}(1, 1)$ if there is one. When should the alarm activate? What is the error?

Solution:

Prior: $\mathbb{P}(\theta = 1) = p = 10\%$ (assume $p = 10\%$, and $\theta = 0$ for no intruder case).

Then for posterior, we have

$$f_{\Theta|X}(0|x^*) \propto \mathbb{P}_\Theta(0) f_{X|\Theta}(x^*|0) \propto (1-p)e^{-\frac{x^{*2}}{2}}$$

$$f_{\Theta|X}(1|x^*) \propto \mathbb{P}_\Theta(1) f_{X|\Theta}(x^*|1) \propto pe^{-\frac{(x^*-1)^2}{2}}$$

$$\frac{f_{\Theta|X}(1|x^*)}{f_{\Theta|X}(0|x^*)} = \frac{pe^{-\frac{(x^*-1)^2}{2}}}{(1-p)e^{-\frac{x^{*2}}{2}}} = \frac{p}{1-p}e^{x^*-\frac{1}{2}}$$

If the value is greater than 1, there will be an intruder. Otherwise, there will be no intruder. To check if the value is greater than 1, we can use a logarithmic trick.

$$\frac{p}{1-p}e^{x^*-\frac{1}{2}} > 1 \iff x^* > \frac{1}{2} + \ln\frac{1-p}{p} \approx 2.7$$

Therefore, when the signal strength is greater than 2.7, the alarm will be triggered.

$$\begin{aligned}
\text{error} &= \mathbb{P}(\hat{\theta} \neq 0) \\
&= \mathbb{P}(\theta = 0, x > 2.7) + \mathbb{P}(\theta = 1, x \leq 2.7) \\
&= \mathbb{P}(x > 2.7 | \theta = 0)\mathbb{P}(\theta = 0) + \mathbb{P}(x \leq 2.7 | \theta = 1)\mathbb{P}(\theta = 1) \\
&= \mathbb{P}(\mathcal{N}(0,1) > 2.7)\mathbb{P}(\theta = 0) + \mathbb{P}(\mathcal{N}(1,1) \leq 2.7)\mathbb{P}(\theta = 1) \\
&\approx 9.86\%
\end{aligned}$$

Chapter 2

Sampling statistics

Starting from this chapter, we will transition from Bayesian statistics to classical statistics. In Bayesian statistics, parameters are treated as random variables with prior distributions, rather than fixed but unknown values. In classical statistics, however, parameters are treated as deterministic (fixed) quantities that are simply unknown. Therefore, we use sampling distributions to estimate parameters.

2.1 Sample Statistics

A random sample of size n is a joint outcome of n independent random variables $X_1, \cdots, X_n$, each with the same PDF or PMF.

> **Remark.** By saying same PDF or PMF, we mean that
>
> $$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = \mu; \qquad \mathrm{Var}[X_1] = \cdots = \mathrm{Var}[X_n] = \sigma^2$$

The process of generating a specific random sample is called sampling. Note that repetition is allowed when taking samples.

2.1.1 Sampling Distributions

Given a random sample of n independent random variables $X_1, \cdots, X_n$ with the same PDF or PMF, the numerical descriptive measures of the sample are called statistics.

Sample mean: $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$;

Sample proportion: $\hat{p} = \dfrac{X_1 + \cdots + X_n}{n}$, where X_i are Bernoulli random variables;

Sample sum: $X = X_1 + \cdots + X_n$;

Sample variance: $s^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$.

Here, all the sample statistics are random variables, which are assumed to occur with repetitions. The probability distributions for statistics are called sampling distributions.

2.1.2 Sample Mean

> **Example.** Consider a fair coin X ($X = 1$ for heads, $X = 0$ for tails). Flip the coin twice, and we obtain X_1, X_2. Then, what is the PMF of $\overline{X}$?
>
> **Solution:** For the joint PMF of X_1, X_2, we have

Joint PMF	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	$\frac{1}{4}$	$\frac{1}{4}$
$X_2 = 1$	$\frac{1}{4}$	$\frac{1}{4}$

Then we have

x	0	1	2
$\mathbb{P}(X_1 + X_2 = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

x	0	$\frac{1}{2}$	1
$\mathbb{P}(\overline{X} = \frac{X_1 + X_2}{2} = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Thus, when we flip the coin n times, we have

$$n\overline{X} \sim \text{Binomial}(n, \frac{1}{2}),$$

where $\overline{X}$ is always a random variable.

In this example, we assume that $X \sim \text{Bernoulli}(p)$ with $p = \mathbb{P}(X = 1) = \frac{1}{2}$. However, in statistics, we do not know p. So how can we describe the distribution? In statistics, we can derive the sampling distribution of sample mean using the laws of probability.

Consider a class that has just finished an exam, and the grades have been released. Since you are a student, you are not supposed to know all the grades or data. So, how can you find out the average exam grade? The most naive approach is to ask your classmates for their grades. For example, you ask three of them, and their grades are 39, 30, and 43, respectively. Then, you can calculate a sample average, which is simply

$$\overline{x} = \frac{39 + 30 + 43}{3} \approx 37.33.$$

However, you cannot ensure that this is 100% accurate, as you might randomly ask three classmates who all happen to have low grades, such as 6, 7, and 5, resulting in a sample average of $\overline{x} = 6$. So how do we measure accuracy? Again, we use the laws of probability to do so.

The sample mean $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ is an estimator of the actual mean:

$$\mu = \mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n],$$

where X_i is a random variable. Also, from the Weak Law of Large Number, we have

$$\mathbb{P}(|\overline{X} - \mu| \geq \epsilon) \leq \delta,$$

The law of probability states that the probability of the sample mean being lower than the actual mean is small and is upper bounded by δ. This leads to an important property of the sample mean: it is **consistent**. In other words, for every positive ϵ and δ, there exists a sufficiently large sample size n such that the probability that $\overline{X}$ differs from the actual mean by more than ϵ is less than δ.

There is another important property of the sample mean: it is an **unbiased** estimator. This means that for every n, $\mathbb{E}[\overline{X}] = \mu$. This is an intuitive concept. Since each X_i is a random variable sampled from the population, the expected value of the sample mean is simply the mean of the actual population.

Proof.

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}\mathbb{E}[X_1 + \cdots + X_n] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \frac{1}{n} \times n\mu = \mu$$

■

Then, based on the Central Limit Theorem, we can find the sampling distribution of the sample mean.

Since we have

$$\mathbb{E}[\overline{X}] = \mu; \qquad \mathrm{Var}[\overline{X}] = \mathrm{Var}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{\sigma^2}{n},$$

for every t,

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{X}{n} \le \frac{\mathbb{E}[X]}{n} + \frac{t\sqrt{\mathrm{Var}[X]}}{n}\right) = \Phi(t);$$

$$\lim_{n\to\infty}\mathbb{P}\left(\overline{X} \le \mu + t\frac{\sigma}{\sqrt{n}}\right) = \Phi(t),$$

where

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim \mathcal{N}(0,1); \qquad \overline{X} \sim \mathcal{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

Note that $\overline{X}$ follows a normal distribution for sufficiently large n. This leads to the question of how to choose the ideal n.

Example. In a population of 1000, 200 people have disease X. For a sample of size 16, what is the probability that the sample mean is in the range of 10% to 30%? Also, consider that 100 people have disease Y out of 1000. For the same sample size, what is $\mathbb{P}(0.05 \le \overline{Y} \le 0.15)$?

Solution:

Disease X: From data we have

$$X_i \sim \text{Bernoulli}\left(p = \frac{200}{1000} = 0.2\right), X_i = 1: \text{ having disease } X$$

$$\overline{X} = \frac{X_1 + \cdots + X_{16}}{16} \Longrightarrow 16\overline{X} \sim \text{Binomial}(16, 0.2)$$

$$\mathbb{P}(0.1 \le \overline{X} \le 0.3) = \mathbb{P}(1.6 \le 16\overline{X} \le 4.8) = \mathbb{P}(2 \le \text{Binomial}(16, 0.2) \le 4) \approx 0.657$$

By using Central Limit Theorem,

$$X_i \sim \text{Bernoulli}(0.2),\ \mu(\overline{X}) = \mu_{X_i} = p = 0.2,\ \sigma(\overline{X}) = \frac{\sigma_{X_i}}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{\sqrt{0.2\times 0.8}}{\sqrt{16}} = 0.1$$

$$\mathbb{P}(0.1 \le \overline{X} \le 0.3) \approx \mathbb{P}\left(\frac{0.1-0.2}{0.1} \le \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}} \le \frac{0.3-0.2}{0.1}\right) = \mathbb{P}(-1 \le Z \le 1) = 0.683$$

Here the difference is within 2.6%.

Disease Y:

$$Y_i \sim \text{Bernoulli}\left(p = \frac{100}{1000} = 0.1\right), Y_i = 1: \text{ having disease } Y, 16\overline{Y} \sim \text{Binomial}(16, 0.1)$$

$$\mathbb{P}(0.05 \le \overline{Y} \le 0.15) = \mathbb{P}(0.8 \le 16\overline{Y} \le 2.4) = \mathbb{P}(1 \le \text{Binomial}(16, 0.2) \le 2) \approx 0.604$$

By using Central Limit Theorem,

$$Y_i \sim \text{Bernoulli}(0.1),\ \mu(\overline{Y}) = 0.1,\ \sigma(\overline{Y}) = \frac{\sigma_{Y_i}}{\sqrt{n}} = \frac{\sqrt{0.1\times 0.9}}{\sqrt{16}} = 0.075$$

$$\mathbb{P}(0.05 \le \overline{Y} \le 0.15) \approx \mathbb{P}\left(\frac{0.05-0.1}{0.075} \le \frac{\overline{Y}-\mu_{\overline{Y}}}{\sigma_{\overline{Y}}} \le \frac{0.15-0.1}{0.075}\right) = \mathbb{P}(-0.666 \le Z \le 0.666) = 0.495$$

Here the difference is within 11%.

Therefore, if the population data is normal, then the sampling distribution of $\overline{X}$ is also normal, regardless of the sample size. For $n \ge 30$, the Central Limit Theorem (CLT) usually applies. However, it depends on the data and the desired precision.

Remark. Again, note that in statistics, we normally don't have the actual data. We are more likely asked to find a function or model to describe the distribution. The data being used are just for demonstration purposes.

2.1.3 Sample Variance

Above, we talked about the unbiased estimator, the sample mean. However, in terms of sample variance, it is a biased estimator due to the biased expectation.

Consider again the exam grade example that was used for illustration earlier. We have a sample mean $\overline{x} = 37.33$, and then we can find the sample variance.

$$s^2 = \frac{(39-37.33)^2 + (30-37.33)^2 + (43-37.33)^2}{3} \approx 29.56.$$

However, as mentioned above, once the sample we take is different, it leads to a different sample variance. In the case of sample variance, the average sample variance, or the expected value of the sample variance, is often smaller than the actual population variance.

For example, we now have data on some $X \sim \text{Bernoulli}(p), p = \frac{1}{2}$. To find σ^2, we can start with the variance for a Bernoulli random variable, in which $\text{Var}[X] = p(1-p)$. Then, we have the actual variance $\sigma^2 = \frac{1}{4}$. When we take two samples, we find that the PMF of $s^2 = \frac{1}{2}((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2)$.

Joint PMF	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	$\frac{1}{4}$	$\frac{1}{4}$
$X_2 = 1$	$\frac{1}{4}$	$\frac{1}{4}$

If $X_1 = X_2$, then $\overline{X} = X_1 = X_2, s^2 = 0$; If $X_1 \neq X_2$, then $\overline{X} = \frac{1}{2}, s^2 = \frac{1}{4}$. This gives

s^2	0	$\frac{1}{4}$
$\mathbb{P}(S^2 = s^2)$	$\frac{1}{2}$	$\frac{1}{2}$

Then we have

$$\mathbb{E}[S^2] = 0 \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{1}{8} = \frac{1}{2}\sigma^2,$$

which is smaller than the actual variance.

In the general case, a random sample of size n consists of independent random variables $X_1, \cdots, X_n$ with the same PDF or PMF.

$$\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2,$$

which shows that we tend to underestimate. However, for a sufficiently large $n \to \infty$, $\frac{n-1}{n} \to 1$.

We can correct the sample variance using the formula above by using $\frac{n-1}{n}$, such that

$$\mathbb{E}[\frac{n}{n-1}S^2] = \sigma^2 \qquad \left(\frac{n}{n-1}S^2 = \frac{n}{n-1}\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}\right)$$

Note that the factor is not significant when n is large, but it is important when n is small.

Proof.

$$
\begin{aligned}
s^2 &= \frac{1}{n}\left((X_1-\overline{X})^2+\cdots+(X_n-\overline{X})^2\right)\\
&= \frac{1}{n}\left(\left(X_1-\frac{X_1+\cdots+X_n}{n}\right)^2+\cdots+\left(X_n-\frac{X_1+\cdots+X_n}{n}\right)^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^n X_i^2+\frac{n(\sum_{i=1}^n X_i)^2}{n^2}-2\left(\sum_{i=1}^n X_i\right)\frac{\sum_{i=1}^n X_i}{n}\right)\\
&= \frac{\sum_{i=1}^n X_i^2}{n}-\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2\\
&= \frac{\sum_{i=1}^n X_i^2}{n}-\overline{X}^2
\end{aligned}
$$

$$\mathbb{E}[s^2]=\mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{n}-\overline{X}^2\right]=\frac{\sum_{i=1}^n \mathbb{E}[X_i^2]}{n}-\mathbb{E}[\overline{X}^2]$$

$$\mathrm{Var}[X_i]=\mathbb{E}[X_i^2]-\mathbb{E}[X_i]^2;\quad \mathbb{E}[X_i^2]=\sigma^2+\mu^2$$

$$\mathrm{Var}[\overline{X}]=\mathbb{E}[\overline{X}^2]-\mathbb{E}[\overline{X}]^2;\quad \mathbb{E}[\overline{X}^2]=\frac{\sigma^2}{n}+\mu^2$$

By substitution, we have

$$\mathbb{E}[s^2]=\frac{\sum_{i=1}^n \mathbb{E}[X_i^2]}{n}-\mathbb{E}[\overline{X}^2]=\sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2=\frac{n-1}{n}\sigma^2$$

■

2.2 Point Estimation

Previously, in Bayesian statistics, we used MAP for point estimation. In classical statistics, there is also a method for point estimation, called Maximum Likelihood Estimation (MLE).

Recall that in classical statistics, the parameter θ is a deterministic quantity that happens to be unknown, and we try to estimate this parameter. Therefore, we develop an estimator $\hat{\theta}$ based on the observations.

2.2.1 Estimators

Suppose that $X_1,\cdots,X_n$ are independent samples with the same PDF/PMF parameterized by θ. Then we can define the following random variables:

$$\text{Estimator: } \widehat{\Theta}_n = g(X_1,\cdots,X_n);$$
$$\text{Estimate: } \hat{\theta}_n = g(X_1=x_1,\cdots,X_n=x_n),$$

where Θ is the random variable that estimates θ, for example, the sample mean.

Then we have:

Unbiased: $\mathbb{E}[\widehat{\Theta}_n]=\theta$

Asymptotically unbiased: $\lim_{n\to\infty}\mathbb{E}[\widehat{\Theta}_n]=\theta$

Consistent: $\widehat{\Theta}_n$ converges to θ in probability

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\mathbb{P}(|\widehat{\Theta}_n-\theta|\geq\varepsilon)=0$$

For an asymptotically unbiased estimator, when n is large enough, i.e., with a sufficiently large sample size, we can approximate the estimator to the actual value. Therefore, we can also use the weak law of large numbers, which states that with a sufficiently large sample size, $\mathbb{P}(\text{sample error } > 0)$ becomes small, meaning $\widehat{\Theta}_n$ is a good estimator.

2.2.2 Maximum Likelihood Estimation

Suppose that $X_1, \cdots, X_n$ are independent samples with the same PDF $f_X(X|\theta)$ (or PMF $\mathbb{P}_X(X|\theta)$). Then, for the maximum likelihood estimate of θ, we have

$$\hat{\theta}_n = \arg\max_{\theta} f_X(x_1, \cdots, x_n|\theta).$$

Through the observation process, we estimate θ using different values. The maximum likelihood estimate is the value of θ that maximizes the likelihood function, representing the parameter value most likely to have produced the observed data:

$$f_X(x|\hat{\theta}) = \max_{\theta} f_X(x|\theta)$$

Example. What is the MLE for θ from Uniform$(0, \theta)$ samples?

Solution: As we observe x_1, x_2, x_3 independently from Uniform$(0, \theta)$, we have:

$$f_X(x_1, x_2, x_3|\theta) = f_X(x_1|\theta) f_X(x_2|\theta) f_X(x_3|\theta) = \frac{1}{\theta^3}(\text{ if } \theta \geq x_1, x_2, x_3 > 0)$$

Here, $\frac{1}{\theta^3}$ is a decreasing function when $\theta > 0$. To maximize the probability, we want to minimize θ. However, the constraint is that $\theta \geq \max\{x_1, x_2, x_3\} > 0$. Therefore, we choose $\theta = \max\{x_1, x_2, x_3\}$, where $\frac{1}{\theta^3}$ reaches its maximum.

$$\theta_{\text{MLE}} = \max\{x_1, x_2, x_3\}$$

Remark. Notice that here θ is treated as an unknown value.

Example. Now we try to find the MLE for Bernoulli(θ). Suppose we observe k heads and $n - k$ tails. What is θ_{MLE}?

Solution:

$$\begin{aligned}
\theta_{\text{MLE}} &= \arg\max_{\theta} f_X(x_1, \cdots, x_n|\theta) \\
&= \arg\max_{\theta} \theta^k (1-\theta)^{n-k} \\
&= \arg\max_{\theta} \text{Beta}(k+1, n-k+1)
\end{aligned}$$

Since

$$\text{Beta}(k+1, n-k+1) = \begin{cases} \dfrac{1}{B(k+1, n-k+1)} \theta^k (1-\theta)^{n-k} & \text{if } 0 < \theta < 1; \\ 0 & \text{otherwise} \end{cases}$$

and we have

$$\text{mode}(\text{Beta}(\alpha, \beta)) = \frac{\alpha - 1}{\alpha - 1 + \beta - 1}.$$

Thus,

$$\theta_{\text{MLE}} = \frac{k}{n}.$$

2.2.3 Systematic Approach to the MLE

We can have a general approach to find MLE. As before, we have MLE: $\hat{\theta} = \arg\max_\theta f_X(x_1, \cdots, x_n|\theta)$. If θ has discrete values, we then compute $f_X(x_1, \cdots, x_n|\theta)$ for each possible value and choose the one that maximizes the likelihood. If θ has continuous values, then we can rely on the properties of $f_X(x_1, \cdots, x_n|\theta)$ to find θ_{MLE}. However, for complicated cases, we need to use another approach.

Since $f_X(x_1, \cdots, x_n|\theta)$ is a function of θ, we can find the θ that maximizes the function by using derivatives if $f_X(x_1, \cdots, x_n|\theta)$ is differentiable with respect to θ (we also consider the boundary cases).

$$\frac{\partial f_X(x_1, \cdots, x_n|\theta)}{\partial \theta} = 0$$

If such an equation can be solved, then we get a closed-form (analytical) solution for θ_{MLE}.

Moreover, if there are more than one parameter to estimate, we can solve the equations jointly.

$$\{\hat{\theta}_1, \cdots, \hat{\theta}_m\} = \arg \max_{\{\hat{\theta}_1, \cdots, \hat{\theta}_m\}} f_X(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m)$$

$$\begin{cases} \dfrac{\partial f_X(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m)}{\partial \theta_1} = 0 \\ \cdots \\ \dfrac{\partial f_X(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m)}{\partial \theta_m} = 0 \end{cases}$$

However, it can become complicated when n is large, as $X_1, \cdots, X_n$ are independent.

$$f_X(x_1, \cdots, x_n | \theta) = \prod_{i=1}^{n} f_X(x_i | \theta) \Longrightarrow \frac{\partial f_X(x_1, \cdots, x_n | \theta)}{\partial \theta} = \frac{\partial \prod_{i=1}^{n} f_X(x_i | \theta)}{\partial \theta}$$

Therefore, we introduce the log-likelihood. For maximum likelihood, we have:

$$\hat{\theta} = \arg \max_{\theta} f_X(x_1, \cdots, x_n | \theta).$$

For maximum log-likelihood, we have

$$\hat{\theta} = \arg \max_{\theta} \ln \left(f_X(x_1, \cdots, x_n | \theta) \right).$$

This is because the $\ln(\cdot)$ function converts the product into sum. Then we have

$$\ln \left(f_X(x_1, \cdots, x_n | \theta) \right) = \ln \left(\prod_{i=1}^{n} f_X(x_i | \theta) \right) = \sum_{i=1}^{n} \ln \left(f_X(x_i | \theta) \right)$$

Also, the $\ln(\cdot)$ function is a strictly increasing function. If $\hat{\theta}$ maximizes $\ln \left(f_X(x_1, \cdots, x_n | \theta) \right)$, it also maximizes $f_X(x_1, \cdots, x_n | \theta)$.

Example. A $\mathcal{N}(\mu, \sigma^2)$ random variable takes the values 2.9 and 3.3. What is the MLE for μ and σ^2?

Solution: Denote $v = \sigma^2$. For likelihood, we have

$$f_X(2.9, 3.3 | \mu, v) = \frac{1}{\sqrt{2\pi v}} e^{\left(-\frac{(2.9-\mu)^2}{2v} \right)} \frac{1}{\sqrt{2\pi v}} e^{\left(-\frac{(3.3-\mu)^2}{2v} \right)} = \frac{1}{2\pi v} e^{\left(-\frac{(2.9-\mu)^2}{2v} \right)} e^{\left(-\frac{(3.3-\mu)^2}{2v} \right)}$$

For log-likelihood, we have

$$\begin{aligned} \ln f_X(2.9, 3.3 | \mu, v) &= \ln e^{\left(-\frac{(2.9-\mu)^2}{2v} \right)} + \ln e^{\left(-\frac{(3.3-\mu)^2}{2v} \right)} - \ln 2\pi v \\ &= -\frac{(2.9-\mu)^2 + (3.3-\mu)^2}{2v} - \ln 2\pi - \ln v \end{aligned}$$

Then we differentiate the log-likelihood.

$$\begin{aligned} \frac{\partial \ln f_X(2.9, 3.3 | \mu, v)}{\partial \mu} &= 0 \\ \frac{2.9 - \mu + 3.3 - \mu}{v} &= 0 \\ \hat{\mu} = \frac{2.9 + 3.3}{2} &= 3.1 \end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln f_X(2.9, 3.3|\mu, v)}{\partial v} &= 0 \\
\frac{(2.9-\mu)^2 + (3.3-\mu)^2}{2v^2} - \frac{1}{v} &= 0 \\
\frac{(2.9-\mu)^2 + (3.3-\mu)^2 - 2v}{2v^2} &= 0 \\
v &= \frac{0.04 + 0.04}{2} = 0.04
\end{aligned}$$

In general, for a random sample of size n, $X_1, \cdots, X_n$ drawn from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the maximum likelihood estimations for μ and σ^2 are:

$$\begin{cases} \hat{\mu} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \end{cases},$$

where the sample mean is an unbiased estimator $\mathbb{E}[\hat{\mu}] = \mu$, and the sample variance is a biased estimator $\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$.

Notice that in practice, we use the corrected unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Appendix A

Z TABLE

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990