ENGG2780 Statistics for Engineers

Ryan Chan

May 12, 2025

Abstract

This is a note for **ENGG2780 - Statistics for Engineers**.
Contents are adapted from the lecture notes of ENGG2780, prepared by Sinno Jialin Pan and Andrej Bogdanov, as well as some online resources.
This note is intended solely as a study aid. While I have done my best to ensure the accuracy of the content, I do not take responsibility for any errors or inaccuracies that may be present. Please use the material thoughtfully and at your own discretion.
If you believe any part of this content infringes on copyright, feel free to contact me, and I will address it promptly.
Mistakes might be found. So please feel free to point out any mistakes.
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

1 Bayesian Statistic **2**
- 1.1 Statistic v.s. Probability . . . 2
- 1.2 Bayesian Statistics . . . 2
- 1.3 Conjugate Priors . . . 7
- 1.4 Applications of Bayesian Statistic . . . 11

2 Sampling Statistics **18**
- 2.1 Sample Statistics . . . 18
- 2.2 Point Estimation . . . 22

3 Confidence Intervals **26**
- 3.1 Definition . . . 26
- 3.2 Confidence Interval for Mean . . . 27
- 3.3 One Sided Confidence Intervals . . . 29
- 3.4 Student's t-distribution . . . 30
- 3.5 Summary . . . 35

4 Hypothesis Testing **36**
- 4.1 Terminology . . . 36
- 4.2 Likelihood Ratio . . . 37

5 Composite Hypothesis **40**
- 5.1 Overview . . . 40
- 5.2 Composite Hypothesis on Population Mean . . . 40
- 5.3 The p-value . . . 44

6 Comparing Population **47**

A Z TABLE **51**

B Student's t-Distribution **52**

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

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

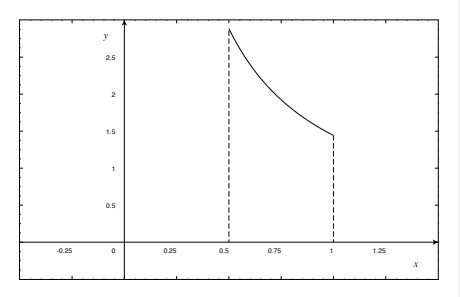
Here we have the prior $f_{\Theta}(\theta) = 1$ if $0 \le \theta \le 1$, and the likelihood $f_{X|\Theta}(\frac{1}{2}|\theta) = \frac{1}{\theta}$ if $\frac{1}{2} \le \theta \le 1$. Keep in mind that the prior comes from Romeo's model, where he has never dated a girl who is

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.

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

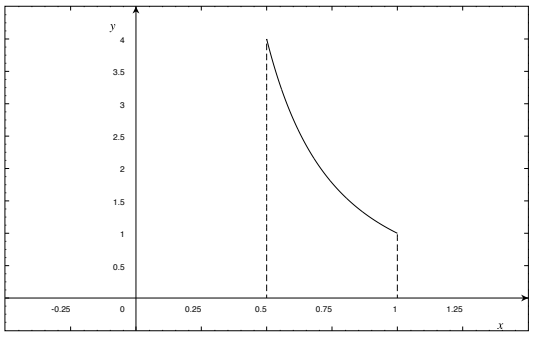
Here, since X_1 and X_2 are independent, we can discard X_1 in the calculation.

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

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:

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

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,

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

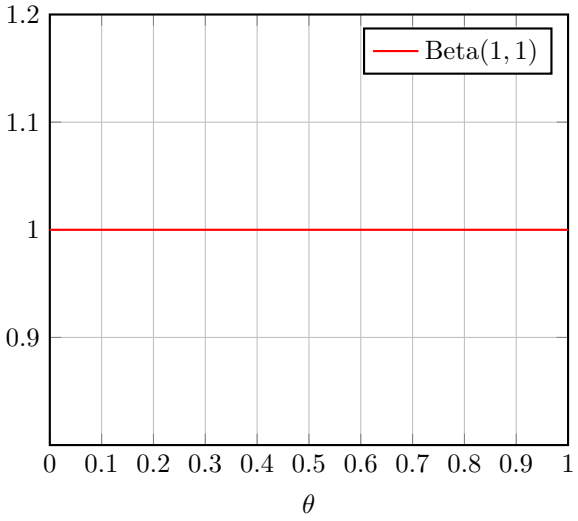
The reason why $B(\alpha, \beta)$ appears in the denominator of the PDF is that it serves as the normalization constant, ensuring that the integral equals 1 so that it is a valid PDF.

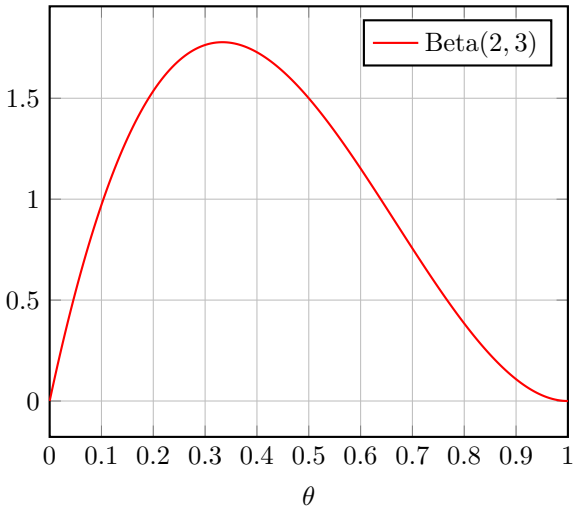
The Beta random variable is widely used to model the prior distribution of a random variable which range is $[0, 1]$, where α and β are hyperparameter.

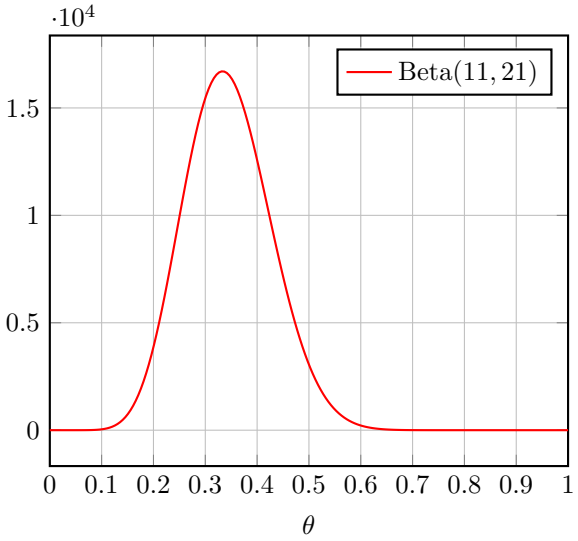
Recalling the coin flip example above, with the prior Θ and observation X remaining unchanged, we can use the Beta distribution to perform the calculation. We have $\Theta \sim \text{Uniform}(0, 1) = \text{Beta}(1, 1)$, and for $h = 1, t = 2$, we have:

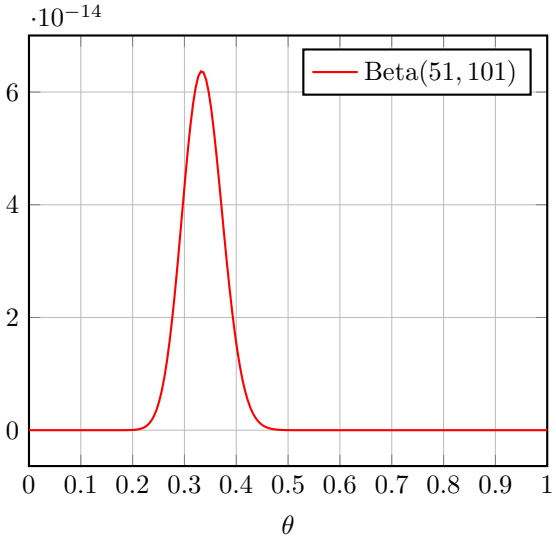
$$f_{\Theta|X_1,X_2,X_3}(\theta|H,T,T) = \frac{1}{\text{Beta}(h+1,t+1)}\theta^{2-1}(1-\theta)^{3-1} = 12\theta(1-\theta)^2$$

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 $(\theta|h \text{ heads}, t \text{ tails}) \sim \text{Beta}(h + 1, t + 1)$.







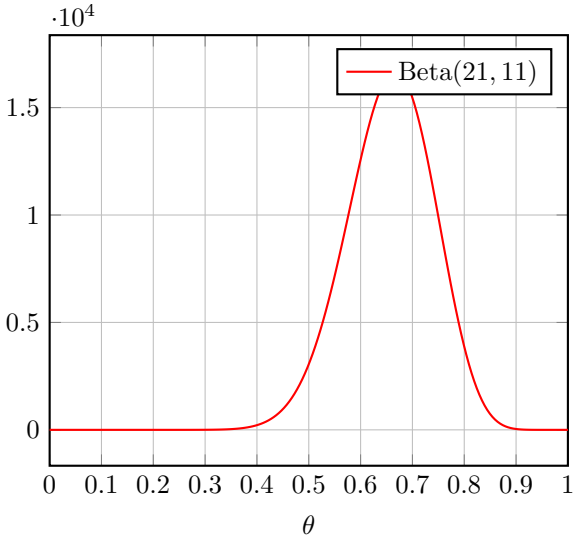


The above shows that we can perform estimation based on the number of experiments, which will result in a different PDF. With more data in hand, the accuracy of the data is higher. However, they share a common feature: the value at which the PDF or PMF reaches its maximum is

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

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 $X \sim$ 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

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

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

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

Chapter 3

Confidence Intervals

In this chapter, we continue the discussion of classical statistics.

In the previous chapter, we discussed the estimation of the value of estimators. However, what we did not discuss is how much the value deviates from the actual one. In other words, how likely is it that the estimated value we have will be the actual one? Since, in classical statistics, the parameters are unknown and not deterministic, we use confidence intervals to determine this probability.

3.1 Definition

In the previous discussion, we stated that θ is an unknown parameter. We then use the estimator $\hat{\Theta}$ to estimate the value of θ, where $\hat{\Theta}$ is a random variable with sample size n. Given different sample sets, the estimate of θ varies. We have discussed unbiasedness, asymptotic unbiasedness, and consistency, which are properties of the estimator rather than a specific estimate.

Thus, besides obtaining a single numerical estimate $\hat{\theta}_n$ of θ based on a specific set of n observed samples, we also want to construct a so-called confidence interval, which not only provides a point estimate but also estimates an interval of values that we are confident contains the unknown θ.

A confidence interval is an interval that contains θ with a certain high probability. For example, we could say that there is a 90% probability that θ lies within the interval.

Based on the point estimate $\hat{\theta}_n$, we construct an interval $[\hat{\theta}_n^-, \hat{\theta}_n^+]$, where $\hat{\theta}_n^- < \hat{\theta}_n^+$, such that we are confident that θ falls within the interval:

$$\mathbb{P}(\hat{\theta}_n^- \leq \theta \leq \hat{\theta}_n^+) \geq \underbrace{1-\alpha}_{\text{Confidence Level}}$$

Here, $[\hat{\theta}_n^-, \hat{\theta}_n^+]$ is called the $(1-\alpha)$ confidence interval, where $\hat{\theta}_n^-$ is the lower confidence limit, and $\hat{\theta}_n^+$ is the upper confidence limit.

> **Remark.** Note that in the above, θ is a true parameter rather than a random variable. To find the probability, one must use the random variable Θ; otherwise, we are unable to determine the likelihood.

We define the width of the confidence interval as $\hat{\theta}_n^+ - \hat{\theta}_n^-$, and α as the confidence parameter. For example, if the confidence parameter is $\alpha = 5\%$, then the confidence level is 95%. One needs to ensure that the confidence parameter is low while maintaining a high confidence level.

For $\hat{\theta}_n^-$ and $\hat{\theta}_n^+$, we can theoretically choose any values. For example, setting them to $-\infty$ and $+\infty$ would give a 100% confidence level, but such an interval is uninformative. Additionally, in most cases, $\hat{\theta}_n^-$ and $\hat{\theta}_n^+$ are symmetric in magnitude since we aim to find the narrowest confidence interval.

Naturally, a question arises: what is the best confidence interval we can choose?

3.2 Confidence Interval for Mean

Suppose we have independent samples $X_1, \cdots, X_n$ with the same PDF or PMF, i.e., they share the same mean μ and variance σ^2. If these samples are normally distributed, i.e., $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

also follows a normal distribution:

$$\overline{X} \sim \mathcal{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

If $X_1, \cdots, X_n$ are not normally distributed but n is large ($n \geq 30$), then by the central limit theorem, the sample mean $\overline{X}$ can still be approximated by the same normal distribution.

As shown before, if $X_1, \cdots, X_n$ are normally distributed as $\mathcal{N}(\mu, \sigma^2)$ or if n is large, then

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

Suppose σ^2 is known. Then, a $(1-\alpha)$-confidence interval for the mean μ is given by:

$$\overline{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \Longleftrightarrow \left[\overline{x} - z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right), x + z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

where $\overline{x}$ is the sample mean estimate based on observed samples, and $z_{\frac{\alpha}{2}}$ is called the z-value or z-score, which satisfies the property that the area to its right under the standard normal curve is $\frac{\alpha}{2}$.

However, since $\overline{x}$ here is a single estimate with a fixed value rather than a random variable, we cannot calculate its probability.

We can derive the probability function for finding such an interval. Since what we are trying to find is the interval such that the actual mean μ will fall into an interval around the sample mean $\overline{X}$, we can define ε as the margin of error, controlling the width of the interval, where $\varepsilon = z_{\frac{\alpha}{2}}$. Then we have:

$$\begin{aligned}
\mathbb{P}\left(\overline{X} - \varepsilon \leq \mu \leq \overline{X} + \varepsilon\right) &= 1 - \alpha \\
\mathbb{P}\left(-\varepsilon \leq \overline{X} - \mu \leq \varepsilon\right) &= 1 - \alpha \\
\mathbb{P}\left(-\varepsilon \leq \frac{\sigma}{\sqrt{n}}\mathcal{N}(0,1) \leq \varepsilon\right) &= 1 - \alpha \\
\mathbb{P}\left(-\frac{\varepsilon\sqrt{n}}{\sigma} \leq \mathcal{N}(0,1) \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right) &= 1 - \alpha \\
\mathbb{P}\left(-z_{\frac{\alpha}{2}}\frac{\sqrt{n}}{\sigma} \leq \mathcal{N}(0,1) \leq z_{\frac{\alpha}{2}}\frac{\sqrt{n}}{\sigma}\right) &= 1 - \alpha
\end{aligned}$$

Alternatively, we can also express this as:

$$\begin{aligned}
\mathbb{P}\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) &= 1 - \alpha \\
\mathbb{P}\left(-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\
\mathbb{P}\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha
\end{aligned}$$

Note that $\overline{X}$ is a random variable, and this function describes a random interval centered at $\overline{X}$ that has a $(1-\alpha)$ probability of containing the population mean μ before a sample is drawn. Once the specific sample is observed, the sample mean $\overline{x}$ becomes fixed, and the interval becomes:

$$\left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

If we repeatedly sampled and computed intervals this way, $100(1-\alpha)\%$ of those intervals would contain μ.

We have talked about the case that σ^2 is known. Now, suppose that σ^2 is unknown, but n is large ($n \geq 30$), then a confidence interval for the mean can also be found by:

$$\overline{x} \pm z_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right) \Longleftrightarrow \left[\overline{x} - z_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right), \overline{x} + z_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right)\right]$$

where s^2 is an unbiased sample standard deviation estimate based on observed samples:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Example. Given a 95% confidence interval for the mean from 30 $\mathcal{N}\left(\mu, \left(\frac{1}{2}\right)^2\right)$ samples.

Solution: As $1-\alpha = 95\%$, thus $\frac{\alpha}{2} = 2.5\% = 0.025$.

Since $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ is a normal random variable, we have

$$\overline{X} \sim \text{Normal}\left(\mu, \left(\frac{1}{2\sqrt{30}}\right)^2\right)$$

Given $\sigma = \frac{1}{2}$, we have

$$\mathbb{P}\left(\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$
$$\mathbb{P}\left(\overline{X} - z_{0.025}\frac{\frac{1}{2}}{\sqrt{30}} \leq \mu \leq \overline{X} + z_{0.025}\frac{\frac{1}{2}}{\sqrt{30}}\right) = 95\%$$

From Z-table, we have $z_{0.025} = 1.96$, then we have

$$\mathbb{P}\left(\overline{X} - z_{0.025}\frac{\frac{1}{2}}{\sqrt{30}} \leq \mu \leq \overline{X} + z_{0.025}\frac{\frac{1}{2}}{\sqrt{30}}\right) = 95\%$$
$$\mathbb{P}\left(\overline{X} - 1.96\frac{\frac{1}{2}}{\sqrt{30}} \leq \mu \leq \overline{X} + 1.96\frac{\frac{1}{2}}{\sqrt{30}}\right) = 95\%$$
$$\mathbb{P}\left(\overline{X} - 0.18 \leq \mu \leq \overline{X} + 0.18\right) = 95\%$$

Then, we can say that we are 95% confident that the actual mean will fall into this interval.

> **Remark.** For a given sample, we can find $\overline{x}$. However, the change of $\overline{x}$ will not affect the width of the interval but only shift it to the left or right. Only the change of size n will change the interval.

Example. How many $\mathcal{N}(\mu, 25^2)$ samples are needed for a 95% confidence with width = 10 intervals?

Solution: As $1-\alpha = 95\%$, thus $\frac{\alpha}{2} = 2.5\% = 0.025$.

Then the corresponding confidence interval:

$$\left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \Longrightarrow \left[\overline{x} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{0.025}\frac{\sigma}{\sqrt{n}}\right]$$

Then we have

$$2 \times z_{0.025}\frac{\sigma}{\sqrt{n}} = 10 \Longrightarrow 2 \times 1.96\frac{25}{\sqrt{n}} = 10$$
$$n = \left(\frac{2 \times 1.96 \times 25}{10}\right)^2 \approx 96$$

Remark. We can also use the probability function to do the calculation:

$$\begin{aligned}\mathbb{P}\left(\overline{X}-\varepsilon\le\mu\le\overline{X}+\varepsilon\right)&=\mathbb{P}\left(-\varepsilon\le\overline{X}-\mu\le\varepsilon\right)\\&=\mathbb{P}\left(-\varepsilon\le\frac{\sigma}{\sqrt{n}}\mathcal{N}(0,1)\le\varepsilon\right)\\&=\mathbb{P}\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\le\mathcal{N}(0,1)\le\frac{\varepsilon\sqrt{n}}{\sigma}\right)\end{aligned}$$

$$\begin{aligned}\frac{\varepsilon\sqrt{n}}{\sigma}&=1.96\\\frac{5\times\sqrt{n}}{25}&=1.96\\n&=\left(\frac{25\times1.96}{5}\right)^2=96\end{aligned}$$

Example. 34 out of 100 Bernoulli(p) samples came out positive. Given a 95% confidence interval, what are the upper and lower limits?

Solution: As $1-\alpha=95\%$, thus $\frac{\alpha}{2}=2.5\%=0.025$.

Since it is a Bernoulli random variable, we have

$$\overline{x}=\hat{p}=\frac{34}{100}=0.34$$

Although σ is unknown, since $n=100$ is large, we can use $s=\sqrt{\hat{p}(1-\hat{p})}\approx0.47$ to approximate σ:

$$\left[\overline{x}-z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\overline{x}+z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]\Longrightarrow\left[\overline{x}-z_{0.025}\frac{s}{\sqrt{n}},\overline{x}+z_{0.025}\frac{s}{\sqrt{n}}\right]$$

Then we have

$$\left[\overline{x}-z_{0.025}\frac{s}{\sqrt{n}},\overline{x}+z_{0.025}\frac{s}{\sqrt{n}}\right]=\left[0.34-1.96\times\frac{0.47}{10},0.34+1.96\times\frac{0.47}{10}\right]=[0.248,0.432]$$

3.3 One Sided Confidence Intervals

Provides a bound for a population parameter with a specified confidence level $(1-\alpha)$. A one-sided interval addresses situations where only one direction is of interest.

3.3.1 Lower One-sided Confidence Interval

For the lower one-sided confidence interval, we have $[\hat{\theta}_n^{\min},+\infty]$, where we are confident that θ is at least as large as $\hat{\theta}_n^{\min}$.

To find $\hat{\theta}_n^{\min}$ such that $\mathbb{P}(\mu\ge\hat{\theta}_n^{\min})=1-\alpha$, we proceed similarly to the two-sided case. However, here we focus only on one side, which leads to:

$$\mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\le z_\alpha\right)=1-\alpha\quad\Longrightarrow\quad\mathbb{P}\left(\mu\ge\overline{X}-z_\alpha\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Therefore, given a specific sample mean $\overline{x}$, the lower one-sided confidence interval is:

$$\left[\overline{x}-z_\alpha\frac{\sigma}{\sqrt{n}},+\infty\right]$$

3.3.2 Upper One-sided Confidence Interval

For the upper one-sided confidence interval, we have $[-\infty, \hat{\theta}_n^{\max}]$, where we are confident that θ is at most as large as $\hat{\theta}_n^{\max}$.

To find $\hat{\theta}_n^{\max}$ such that $\mathbb{P}(\mu \leq \hat{\theta}_n^{\max}) = 1 - \alpha$, we follow a similar approach, focusing on just one side of the distribution:

$$\mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha\right) = 1 - \alpha \quad \Longrightarrow \quad \mathbb{P}\left(\mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Therefore, given a specific sample mean $\overline{x}$, the upper one-sided confidence interval is:

$$\left[-\infty, \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

3.4 Student's t-distribution

Remark. Some discussions in this section are not included in the lecture notes of ENGG2780 but are helpful for understanding. They are marked with ⊛.

Previously, we talked about the case when σ^2 is known, and with a large sample n, we can find the $(1-\alpha)$ confidence interval for the mean μ. Now, if we have an unknown σ^2 and a large n ($n \geq 30$), then we have a $1 - \alpha$ confidence interval for the mean μ:

$$\overline{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \quad s^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}.$$

Intuition. As σ^2 is unknown, we cannot utilize $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$ to construct the confidence interval. Instead, we need to analyze the distribution of $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$, where the unbiased sample standard deviation S is a random variable.

Here we consider $X_1, \cdots, X_n$ as independent random variables with a large n.

Case I: For $X_i \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1) \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

Then we have

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1) \Rightarrow \mathbb{P}\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Given a specific sample mean and standard deviation $\overline{x}, s$, the interval becomes fixed.

Case II: If the distribution of X_i is unknown, then based on the Central Limit Theorem, we have

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

As $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim t(n-1)$, we have

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \xrightarrow{P} \sigma^2$$

Then, from Slutsky's Theorem, we have

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{S} = Z \sim \mathcal{N}(0,1)$$

As n is large, the distribution of $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$ is assumed to be approximated by $\mathcal{N}(0,1)$.

However, when σ is unknown and we have a small n, how can we find the confidence interval? This brings us to the Student's t-distribution.

3.4.1 Chi-squared Random Variable

We first take a look at some normal algebra.

Suppose we have $X_1 \sim \text{Normal}(\mu_1, \sigma_1)$ and $X_2 \sim \text{Normal}(\mu_2, \sigma_2)$, where X_1, X_2 are independent. Then, we have $X_1 + X_2 \sim \text{Normal}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$ and $X_1 - X_2 \sim \text{Normal}(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Then, the adjusted sample variance S^2 of two independent $\text{Normal}(\mu, \sigma)$ samples is

$$\begin{aligned} S^2 &= (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 \\ &= (X_1 - \frac{X_1 + X_2}{2})^2 + (X_2 - \frac{X_1 + X_2}{2})^2 \\ &= \left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_1 - X_2}{2}\right)^2 \\ &= \frac{1}{2}(X_1 - X_2)^2 \\ &= \frac{1}{2}\text{Normal}(0, \sqrt{2}\sigma)^2 = \text{Normal}(0, \sigma^2). \end{aligned}$$

Now, consider a sample $X_1, X_2, \ldots, X_n$ drawn from a normal distribution:

$$X_i \sim \text{Normal}(\mu, \sigma^2), \quad \text{for } i = 1, 2, \ldots, n.$$

The sample mean is given by:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample variance is defined as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

We can then find

$$\mathbb{E}[X_i - \overline{X}] = \mathbb{E}[X_i] - \mathbb{E}[\overline{X}] = \mu - \mu = 0$$

$$\text{Var}(X_i - \overline{X}) = \left(1 - \frac{1}{n}\right)^2 \sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \frac{n-1}{n}\sigma^2.$$

Then, we have

$$X_i - \overline{X} \sim \text{Normal}\left(0, \frac{n-1}{n}\sigma^2\right).$$

The sum of squares of these normal deviations forms a chi-squared distributed random variable with $n-1$ degrees of freedom:

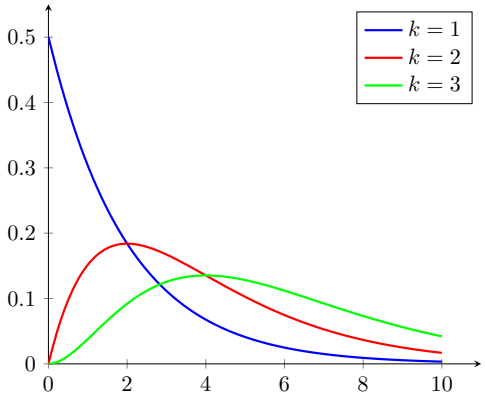
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

If $X_1, \cdots, X_n$ are independent standard normal random variables $\mathcal{N}(0,1)$, then we have

$$(X_1^2 + \cdots + X_n^2) \sim \chi^2(n),$$

where n is the degrees of freedom (df), and it has the probability density function (PDF)

$$f(x) \propto x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$$



Remark. Note that although X_i are independent Normal(μ, σ), the variables $X_i - \overline{X}$ are not independent. Suppose $X_1 - \overline{X} = z_1, X_2 - \overline{X} = z_2$, then

$$z_1 + z_2 + \cdots + z_n = (X_1 + X_2 + \cdots + X_n) - n\overline{X} = 0.$$

Theorem 3.4.1. If $X_1, \cdots, X_n$ are independent random variables following $\mathcal{N}(\mu, 1)$, then

$$(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2 \sim \chi^2(n-1),$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean.

Corollary 3.4.1. If $X_1, \cdots, X_n$ are independent random variables following $\mathcal{N}(\mu, \sigma^2)$, then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is the sample variance.

⊛ Then, to find the confidence interval for σ, we have

$$\begin{aligned}
\mathbb{P}(z^- \le \chi^2(n-1) \le z^+) &= \mathbb{P}(z^- \le \chi^2(n-1) \le z^+) \\
&= \mathbb{P}(z^- \le \frac{(n-1)S^2}{\sigma^2} \le z^+) \\
&= \mathbb{P}(\frac{(n-1)S^2}{z^+} \le \sigma^2 \le \frac{(n-1)S^2}{z^-}) \\
&= \mathbb{P}(\sqrt{\frac{(n-1)S^2}{z^+}} \le \sigma \le \sqrt{\frac{(n-1)S^2}{z^-}})
\end{aligned}$$

Now, we can go back to the discussion.

3.4.2 Student's t Random Variable

Since we consider the case where σ is unknown, we can no longer use the previous method, i.e., $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, to find the confidence interval for the mean. Instead, we use the sample variance S to estimate the confidence interval. However, in this case, the distribution is no longer normal.

For example, we have

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\text{Normal}(0, \frac{\sigma}{\sqrt{n}})}{\sqrt{\frac{\sigma^2\chi^2(n-1)}{n(n-1)}}} = \frac{\text{Normal}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{(n-1)}}}$$

Since Normal$(0,1)$ and $\sqrt{\frac{\chi^2(n-1)}{n-1}}$ are independent, we can define a random variable for this distribution. This is what we call the Student's t-distribution.

If $X_1, \cdots, X_n$ are independent random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, then a $(1-\alpha)$-confidence interval for the mean μ is given by

$$\overline{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}},$$

where $t_{\frac{\alpha}{2}}$ is the t-score such that the area to the right of it under the t-distribution curve with $n-1$ degrees of freedom is $\frac{\alpha}{2}$.

Consider two independent random variables Y and Z, such that Y has the χ^2 distribution with n degrees of freedom ($\chi^2(n)$) and Z has the standard normal distribution $\mathcal{N}(0,1)$. Suppose a random variable T is defined as

$$T = \frac{Z}{\sqrt{\frac{Y}{n}}} = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n)}{n}}}.$$

The distribution of T is called the t-distribution or Student's t-distribution with $n-1$ degrees of freedom.

Theorem 3.4.2. If $X_1, \cdots, X_n$ are independent random variables following $\mathcal{N}(\mu, \sigma^2)$, regardless of whether n is large or not, then we have

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1),$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the sample variance.

Proof. As $X_1, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$ random variables, we have

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Denote $Y = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. Then,

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t(n-1) \Longleftrightarrow T = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{S}{\sigma}} = \frac{(\overline{X} - \mu)}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

This holds true when Z and Y are independent.

To show that Z and Y are independent, we prove that $\overline{X}$ and S^2 are independent. Since

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

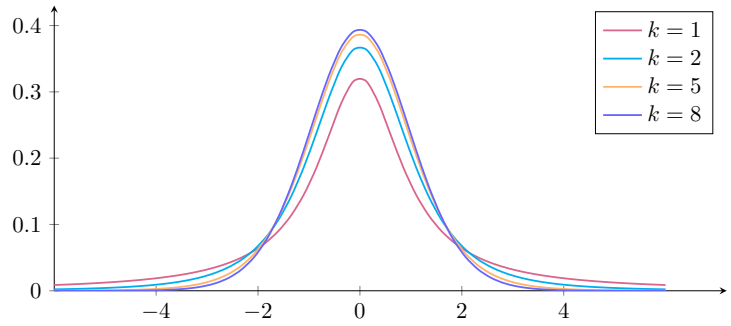
it suffices to show that $\overline{X}$ and $X_i - \overline{X}$ are independent. We know that

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad (X_i - \overline{X}) \sim \mathcal{N}\left(0, \frac{n-1}{n}\sigma^2\right).$$

We now calculate the covariance between $\overline{X}$ and $X_i - \overline{X}$:

$$\begin{aligned}
\operatorname{Cov}(\overline{X}, X_i - \overline{X}) &= \operatorname{Cov}(\overline{X}, X_i) - \operatorname{Cov}(\overline{X}, \overline{X}) \\
&= \frac{1}{n}\operatorname{Cov}(X_i, X_i) - \operatorname{Cov}(\overline{X}, \overline{X}) \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.
\end{aligned}$$

Thus, $\overline{X}$ and $X_i - \overline{X}$ are independent, and therefore, Z and Y are independent. ■



As shown in the graph, we can see that with a larger n, the distribution becomes more similar to the standard normal distribution $\mathcal{N}(0,1)$, since the sample variance becomes a more accurate estimate of the population variance:

$$t(n) = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n)}{n}}} \overset{n\to\infty}{\Longrightarrow} \mathcal{N}(0,1)$$

Now, if $X_1, \cdots, X_n$ are independent $\mathcal{N}(\mu, \sigma^2)$, then a $(1-\alpha)$-confidence interval for the mean μ is

$$\overline{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

where the t-value gives us the area to the right of it under the t-distribution curve with degrees of freedom $n-1$ equal to $\frac{\alpha}{2}$.

Proof. Given $X_1, \cdots, X_n$ are independent random variables following $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is unknown, and n is small ($n < 30$), we have

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

Therefore, the probability that T lies between the critical values $\pm t_{\frac{\alpha}{2}}$ is

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

By multiplying through by $\frac{S}{\sqrt{n}}$, we get

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \overline{X} - \mu \leq t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Finally, solving for μ, we obtain the confidence interval for μ:

$$\mathbb{P}\left(\overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

■

Example. 5 random athletes are 152, 163, 188, 201, and 192 cm tall. Given a 95% confidence interval for μ.

Solution: Here we have $n = 5$, so the degrees of freedom are $n - 1 = 4$.

$$\overline{x} = \frac{152 + 163 + 188 + 201 + 192}{5} = 179.2$$

$$s = \sqrt{\frac{\sum_{i=1}^{5} (x_i - 179.2)^2}{4}} = 20.73$$

For $\alpha = 5\%$, we have $\frac{\alpha}{2} = 2.5\% = 0.025 \Rightarrow t_{\frac{\alpha}{2}} = 2.78$.

$$\overline{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 179.2 \pm 2.78 \times \frac{20.73}{\sqrt{5}} = 179.2 \pm 25.78 \Rightarrow [153.43, 204.97]$$

3.4.3 One-sided Confidence Interval

Consider $X_1, \cdots, X_n$ as independent $\mathcal{N}(\mu, \sigma^2)$ random variables, where σ^2 is unknown and $n < 30$. We have:

Lower one-sided: Find $\hat{\theta}_n^{\min}$ such that $\mathbb{P}(\mu \geq \hat{\theta}_n^{\min}) = 1 - \alpha$

$$\mathbb{P}(T \leq t_\alpha) = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_\alpha\right) = 1 - \alpha \Longrightarrow \mathbb{P}\left(\mu \geq \overline{X} - t_\alpha \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Given specific values of $\overline{x}$ and s, we have

$$[\overline{x} - t_\alpha, +\infty]$$

Upper one-sided: Find $\hat{\theta}_n^{\max}$ such that $\mathbb{P}(\mu \leq \hat{\theta}_n^{\max}) = 1 - \alpha$

$$\mathbb{P}(T \geq -t_\alpha) = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \geq t_\alpha\right) = 1 - \alpha \Longrightarrow \mathbb{P}\left(\mu \leq \overline{X} + t_\alpha \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Given specific values of $\overline{x}$ and s, we have

$$[-\infty, \overline{x} + t_\alpha]$$

3.5 Summary

In summary, if $X_1, \cdots, X_n$ are independent samples with the same PMF or PDF, for a $(1-\alpha)$-confidence interval, we consider the following cases:

1. **Known** σ^2

If n is large, i.e. $n \geq 30$, or if $n < 30$ with the PDF $\mathcal{N}(\mu, \sigma^2)$, we have:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1), \quad \overline{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

2. **Unknown** σ^2

Case 1: If n is large, then we have:

$$\sigma \approx s, \quad Z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{N}(0, 1), \quad \overline{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Case 2: If $n < 30$, with the PDF $\mathcal{N}(\mu, \sigma^2)$, then we have:

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1), \quad \overline{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Chapter 4

Hypothesis Testing

Previously, we talked about hypothesis testing in the context of Bayesian statistics, where we use it to check how likely the estimated parameter is to be the actual one given the observed data and a prior. The same applies to classical statistics, but instead of having a prior, we only have observations.

So, how do we determine how likely the estimated distribution is to be the actual one, or to what extent it has errors? This is where hypothesis testing comes in.

4.1 Terminology

Before we dive into the context, we first define some terminology.

Recall that in Bayesian statistics, we have a parameter Θ that takes m possible values $\theta_1, \theta_2, \ldots, \theta_m$, and we use Bayesian inference to estimate the most likely value given observed data. Now, we consider a special case where Θ can take only two values, 0 and 1. This scenario is known as binary hypothesis testing.

We denote the hypothesis $\Theta = 0$ by H_0, called the null hypothesis, which is considered the default assumption. We denote the hypothesis $\Theta = 1$ by H_1, called the alternative hypothesis.

For example, we can claim that a new drug has no effect, which is the default claim H_0, while the claim that "the new drug has an effect" corresponds to the alternative hypothesis H_1.

Note that under each hypothesis, the data follows a specific probability distribution. By default, we assume the sample follows the distribution defined by H_0, and hypothesis testing determines whether there is sufficient evidence to reject H_0 in favor of H_1.

After observing n independent samples $X_1, \ldots, X_n$ with the same PMF or PDF, which depends on the hypothesis, we denote by $f_{X|\Theta}$ the PMF or PDF defined by the hypothesis.

A binary decision rule can be represented by two disjoint regions of all possible observations. We have R, called the rejection region, where hypothesis H_0 is rejected if the observed data fall within this region. Simply put, the samples suggest that the data follow the distribution under hypothesis H_1.

The complement R^c is called the acceptance region, where hypothesis H_0 is accepted if the observed data fall within this region. In other words, the data are consistent with the distribution under H_0.

For a particular choice of the rejection region, we have two types of errors.

For a **Type I error**, also known as a false rejection or false positive, we reject hypothesis H_0 even though H_0 is true. In other words, the data suggest that it follows H_1, while the actual distribution is H_0. We define the probability of a Type I error as:

$$\alpha(R) = \mathbb{P}(X \in R | H_0).$$

Similarly, we have a **Type II error**, also known as a false acceptance or false negative. This occurs when we accept hypothesis H_0 even though H_0 is false. In other words, the data suggest that it follows

H_0, while the actual distribution is H_1. We define the probability of a Type II error as:

$$\beta(R) = \mathbb{P}(X \notin R | H_1).$$

4.2 Likelihood Ratio

4.2.1 Likelihood Ratio Test

Now, let's look at an example. Suppose we have a coin that can either be fair (H_0) or loaded (H_1), where $p_H = \frac{3}{4}$. If we decide that when the number of heads observed is greater than or equal to 14, we would conclude that it is more likely to be H_1.

We can now calculate the error as follows:

$$\mathbb{P}(H_1|H_0) = \mathbb{P}(\text{Binomial}(20, 0.5) \geq 14) = 0.057$$

$$\mathbb{P}(H_0|H_1) = \mathbb{P}(\text{Binomial}(20, \frac{3}{4}) \leq 13) = 0.214$$

Note that here the Type I error is much smaller than the Type II error, which is desirable, since Type I errors are usually more costly.

One way to think about it is that the null hypothesis represents the default case, which we only reject if there is sufficient evidence. However, a Type I error means that even though the null hypothesis is true, we still reject it. This is not considered a "safe" decision, so we typically aim to minimize the Type I error, even though doing so may increase the Type II error.

That said, there are cases where a Type II error is more costly—for example, failing to detect a disease—but such cases are less common. But in general-purpose testing, Type I is treated more seriously because it's the one we explicitly control.

However, we still want to make a decision. So how can we ensure that the decision we make is optimal? One approach is to use the **Likelihood Ratio Test**.

Suppose $X_1, \cdots, X_n$ are independent random variables with the same PDF or PMF. Then we define the **likelihood ratio** as

$$L(x_1, \cdots, x_n) = \frac{f_X(x_1, \cdots, x_n|H_1)}{f_X(x_1, \cdots, x_n|H_0)}.$$

We use the following general decision rule: we accept H_1 ($\Theta = 1$) if $L(x_1, \cdots, x_n) > \xi$, where $\xi > 0$ is a critical threshold. Otherwise, we accept H_0 ($\Theta = 0$).

> **Remark.** Notice that in Bayesian statistics, we use the MAP (Maximum A Posteriori) rule, where the threshold is given by
>
> $$\xi = \frac{\mathbb{P}_\Theta(\theta = 0)}{\mathbb{P}_\Theta(\theta = 1)},$$
>
> and the posterior ratio becomes
>
> $$\frac{f_{\Theta|X}(1|x)}{f_{\Theta|X}(0|x)} = \frac{f_{X|\Theta}(x|1)\mathbb{P}_\Theta(\theta = 1)}{f_{X|\Theta}(x|0)\mathbb{P}_\Theta(\theta = 0)} > 1.$$
>
> In classical statistics, we use the likelihood ratio test without a prior:
>
> $$\frac{f_X(x|H_1)}{f_X(x|H_0)} > 1.$$

For example, given a six-sided die, there are two hypotheses: fair (H_0) or loaded (H_1), with the following PMFs:

$$P_X(x|H_0) = \frac{1}{6} \quad \text{for } x = 1, \ldots, 6$$

$$P_X(x|H_1) = \begin{cases} \frac{1}{4}, & \text{if } x = 1, 2 \\ \frac{1}{8}, & \text{if } x = 3, 4, 5, 6 \end{cases}$$

Given a single roll x of the die, consider the likelihood ratio:

$$L(x) = \begin{cases} \frac{\frac{1}{4}}{\frac{1}{6}} = \frac{3}{2}, & \text{if } x = 1, 2 \\ \frac{\frac{1}{8}}{\frac{1}{6}} = \frac{3}{4}, & \text{if } x = 3, 4, 5, 6 \end{cases}$$

There are three possibilities to consider for the critical threshold ξ:

$$\begin{cases} \xi < \frac{3}{4}, & \text{Reject } H_0 \text{ for all } x \\ \frac{3}{4} < \xi < \frac{3}{2}, & \text{Reject } H_0 \text{ if } x = 1, 2; \text{ Accept } H_0 \text{ otherwise} \\ \xi > \frac{3}{2}, & \text{Accept } H_0 \text{ for all } x \end{cases}$$

Then we can compute the probabilities of Type I and Type II errors:

Type I error (false positive):

$$\alpha(\xi) = \mathbb{P}(x \in R | H_0) = \begin{cases} 1, & \text{if } \xi < \frac{3}{4} \\ \frac{2}{6} = \frac{1}{3}, & \text{if } \frac{3}{4} < \xi < \frac{3}{2} \\ 0, & \text{if } \xi > \frac{3}{2} \end{cases}$$

Type II error (false negative):

$$\beta(\xi) = \mathbb{P}(x \notin R | H_1) = \begin{cases} 0, & \text{if } \xi < \frac{3}{4} \\ \frac{4}{8} = \frac{1}{2}, & \text{if } \frac{3}{4} < \xi < \frac{3}{2} \\ 1, & \text{if } \xi > \frac{3}{2} \end{cases}$$

Also, as ξ increases, the rejection region becomes smaller. Thus, the probability of false rejection, $\alpha(R)$, decreases, while the probability of false acceptance, $\beta(R)$, increases.

Then, how do we choose the trade-off threshold ξ?

One popular approach to choosing ξ is the **Likelihood Ratio Test**. We first specify a target value α for the false rejection probability (Type I error). Then, we select a value of ξ such that the false rejection probability equals α, i.e.,

$$\mathbb{P}(L(x) > \xi | H_0) = \alpha.$$

Once the value $X = x$ is observed, we reject H_0 if $L(x) > \xi$.

Here, α is called the *significance level*. Typical choices for α are 0.01, 0.05, and 0.10.

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? How should we choose ξ?

Solution: Let H_0 denote "no intruder" and H_1 denote "intruder present". Then, the densities are:

$$f_X(x|H_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad f_X(x|H_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}.$$

The likelihood ratio is:

$$L(x) = \frac{f_X(x|H_1)}{f_X(x|H_0)} = \frac{e^{-\frac{(x-1)^2}{2}}}{e^{-\frac{x^2}{2}}} = e^{\frac{2x-1}{2}}.$$

Suppose we set the significance level $\alpha = 0.025$, the probability of a Type I error. Then we solve:

$$\mathbb{P}(L(x) > \xi | H_0) = 0.025.$$

Substituting the likelihood ratio:

$$\mathbb{P}(e^{\frac{2X-1}{2}} > \xi | H_0) = 0.025,$$

where $X \sim \mathcal{N}(0,1)$. Taking logarithms:

$$\mathbb{P}\left(X > \ln \xi + \frac{1}{2}\right) = 0.025.$$

Using the standard normal quantile $z_{0.975} = 1.96$, we solve:

$$\ln \xi + \frac{1}{2} = 1.96 \quad \Longrightarrow \quad \ln \xi = 1.46 \quad \Longrightarrow \quad \xi = e^{1.46}.$$

4.2.2 Neyman-Pearson Lemma

Consider a particular choice of ξ in the likelihood ratio test, which results in error probabilities

$$\mathbb{P}(L(x) > \xi \mid H_0) = \alpha \quad \text{and} \quad \mathbb{P}(L(x) \leq \xi \mid H_1) = \beta.$$

Suppose an alternative test with rejection region R' satisfies

$$\mathbb{P}(X \in R' \mid H_0) \leq \alpha,$$

then we have

$$\mathbb{P}(X \notin R' \mid H_1) \geq \beta,$$

with strict inequality

$$\mathbb{P}(X \notin R' \mid H_1) > \beta \quad \text{if} \quad \mathbb{P}(X \in R' \mid H_0) < \alpha.$$

This lemma shows that no test exists such that $\mathbb{P}(X \in R' \mid H_0) = \alpha$ while $\mathbb{P}(X \notin R' \mid H_1) < \beta$.

Chapter 5

Composite Hypothesis

5.1 Overview

We talked about binary hypothesis in the previous chapter. However, in real-world settings, hypothesis testing problems do not always involve two well-specified parameters as alternatives. We still have two disjoint hypotheses, but they are not as well-specified as in the binary case.

In the previous chapter, we discussed simple hypotheses, where the distribution or the parameters are completely specified. For example, we have a drug that is either effective or not, giving us $H_0 = 0$ and $H_1 = 1$. However, for composite hypotheses, the distribution or the parameters are not completely specified.

For example, consider the claim that the average monthly income of residents of a city is more than or equal to 20,000 dollars. Now we have $H_0 : \mu \geq 20,000$, while $H_1 : \mu < 20,000$ is not completely specified.

Therefore, we need a more general statistical test of hypothesis. To do this, we first specify the null hypothesis H_0 and complement H_1. We then choose a test statistic based on the random samples for the parameter in the hypotheses. For example, we can choose $\overline{X}$ as a test statistic for μ.

Then, assuming H_0 is true, we look for evidence from observations to support H_1. We make a conclusion: either reject H_0 if there is strong evidence from the test that indicates the assumption H_0 is true does not hold, or accept H_0 if there is no strong statistical evidence from observations to refute the assumption.

5.2 Composite Hypothesis on Population Mean

Here we consider the composite hypothesis on population mean μ.

There are two categories. The first one is the two-sided test, where we have

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

The second category is the one-sided test, where we have

$$\text{Right-sided: } H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0$$

or

$$\text{Left-sided: } H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

In the critical value approach, we define a significance level α, which is the largest false rejection probability we can accept. Then we find the rejection region of H_0 such that $\alpha = \mathbb{P}(H_1|H_0)$. We estimate the sample mean $\overline{x}$ from the observed data. If $\overline{x}$ is in the rejection region, then we consider there is strong statistical evidence to reject H_0. Otherwise, we accept H_0.

5.2.1 Two-Sided test

When $n \geq 30$, we can have the two-sided test of μ. Suppose $X_1, \cdots, X_n$ are independent samples with the same PDF or PMF (μ, σ^2, etc.). We assume that $H_0 : \mu = \mu_0$ is true, and we have the test statistic, which is the sample mean $\overline{X}$. The rejection region will then be $|\overline{X} - \mu_0| \geq \xi$, where $\xi > 0$. ξ is determined by solving $\alpha = \mathbb{P}(H_1|H_0)$, then we have

$$\alpha = \mathbb{P}(|\overline{X} - \mu_0| \geq \xi | \mu = \mu_0)$$

Remark. The further away the sample mean $\overline{X}$ is from μ_0, the stronger the evidence points toward $H_1 : \mu \neq \mu_0$.

Then, based on the central limit theorem, assuming H_0 is true, as n is large, we have

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\alpha = \mathbb{P}\left(|\overline{X} - \mu_0| \geq \xi\right) = \mathbb{P}\left(|Z| \geq \frac{\xi}{\sigma/\sqrt{n}}\right) = \mathbb{P}\left(|Z| \geq z_{\alpha/2}\right)$$

Then, when σ is known, given a specific estimate $\overline{x}$, if

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} \geq z_{\alpha/2},$$

we reject H_0. Otherwise, we accept it.

When σ is unknown, we approximate μ with s. Given a specific estimate $\overline{x}$, if

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \geq z_{\alpha/2},$$

we again reject H_0. Otherwise, we accept it.

However, when $n < 30$, as n is small, CLT is not applicable here. But if $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then we still have

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Then, when σ is known, we have

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

When σ is unknown, we have

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

Then, from the same formula, we have

$$\alpha = \mathbb{P}\left(|\overline{X} - \mu_0| \geq \xi\right) = \mathbb{P}\left(|T| \geq \frac{\xi}{S/\sqrt{n}}\right) = \mathbb{P}\left(|T| \geq t_{\alpha/2}\right)$$

Given a specific estimate $\overline{x}$, if

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \geq t_{\alpha/2},$$

then we reject H_0; otherwise, we accept it.

Example. The average temperature of Hong Kong in February is 18°C. Has this year been unusual? Assume temperature in February follows $\mathcal{N}(\mu, \sigma^2)$ and $\sigma = 3$°C. Suppose $\alpha = 0.05$.

Day	1	6	11	16	21	26
Temp (°C)	15	15	19	18	8	17

Solution: Here we have $H_0 : \mu = 18$, $H_1 : \mu \neq 18$.

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

By calculation, we have

$$\overline{x} = \frac{15+15+19+18+8+17}{6} = 15.33, \quad \mu_0 = 18$$

Then,

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} = \frac{|15.33-18|}{3/\sqrt{6}} \approx 2.18 > z_{0.025} = 1.96$$

Thus, we reject H_0.

However, if σ is unknown, we have

$$s^2 = \frac{(15-15.33)^2 + \cdots + (17-15.33)^2}{6-1} \approx 15.47, \quad s \approx 3.93$$

Then,

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} = \frac{|15.33-18|}{3.93/\sqrt{6}} \approx 1.71 < t_{0.025} \approx 2.57$$

Thus, we accept H_0.

5.2.2 One-Sided Test

In category II, we have one-sided tests, where

$$\text{Right-sided: } H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0$$

or

$$\text{Left-sided: } H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

We can transform these hypotheses as

$$\text{Right-sided: } H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0$$

or

$$\text{Left-sided: } H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

After transformation, $H_0 : \mu = \mu_0$ is rejected only if there is strong evidence to support $H_1 : \mu > \mu_0$. In this case, we also reject $H_0 : \mu \leq \mu_0$. $H_0 : \mu = \mu_0$ is not rejected only if there is no strong evidence to support $H_1 : \mu > \mu_0$. In this case, $H_0 : \mu \leq \mu_0$ should not be rejected either.

We can apply the same approach by assuming H_0 is true, and then setting the false rejection probability α to find the rejection region for H_0. Then we have

$$\alpha = \mathbb{P}(H_1 | H_0 : \mu = \mu_0) \geq \mathbb{P}(H_1 | H_0 : \mu \leq \mu_0)$$

Right-Sided Test

For a right-sided test of μ, we have

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0$$

Here we set

$$\alpha = \mathbb{P}(\overline{X} - \mu_0 \geq \xi | \mu = \mu_0)$$

When $n \geq 30$, based on the CLT and assuming H_0 is true, we have

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\alpha = \mathbb{P}\left(\overline{X} - \mu_0 \geq \xi | \mu = \mu_0\right) = \mathbb{P}\left(Z \geq \frac{\xi}{\sigma/\sqrt{n}}\right) = \mathbb{P}\left(Z \geq z_\alpha\right)$$

Same as before, when σ is known, given a specific estimate $\overline{x}$, if

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha,$$

we reject H_0; otherwise, we accept it.

When σ is unknown, we approximate σ with s. Given a specific estimate $\overline{x}$, if

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \geq z_\alpha,$$

we reject H_0; otherwise, we accept it.

When $n < 30$, since the sample size is small, the CLT is not applicable. But if $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, then we still have

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Then, when σ is known, we have

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

When σ is unknown, we have

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

Then, given a specific estimate $\overline{x}$, if

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha \quad \text{or} \quad \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \geq t_\alpha,$$

then we reject H_0; otherwise, we accept it.

Left-Sided Test

For a left-sided test of μ, we have

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

Here we set

$$\alpha = \mathbb{P}(\overline{X} - \mu_0 \leq \xi | \mu = \mu_0)$$

When $n \geq 30$ and σ is known, given a specific estimate $\overline{x}$, if

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha,$$

we reject H_0; otherwise, we accept it.

When σ is unknown, we approximate it with s. If

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq -z_\alpha,$$

we reject H_0; otherwise, we accept it.

When $n < 30$, if $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, and σ is known, then if

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha \quad \text{or} \quad \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq -t_\alpha,$$

we reject H_0; otherwise, we accept it.

Example. The average temperature of Hong Kong in February is 18°C. Has this year been colder? Assume temperature in February follows $\mathcal{N}(\mu, \sigma^2)$ and $\sigma = 3$°C. Suppose $\alpha = 0.05$.

Day	1	6	11	16	21	26
Temp (°C)	15	15	19	18	8	17

Solution: Here we have $H_0 : \mu = 18$, $H_1 : \mu \neq 18$.

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

By calculation, we have

$$\overline{x} = \frac{15 + 15 + 19 + 18 + 8 + 17}{6} = 15.33, \quad \mu_0 = 18$$

Then,

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{15.33 - 18}{3/\sqrt{6}} \approx -2.18 < -z_{0.05} \approx -1.645$$

Thus, we reject H_0.

However, if σ is unknown, we have

$$s^2 = \frac{(15 - 15.33)^2 + \cdots + (17 - 15.33)^2}{6 - 1} \approx 15.47, \quad s \approx 3.93$$

Then,

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{15.33 - 18}{3.93/\sqrt{6}} \approx -1.71 > -t_{0.05} \approx -2.015$$

Thus, we accept H_0.

5.3 The p-value

The p-value is the smallest probability of committing a type I error for which the null hypothesis H_0 would be rejected, given a specific test statistic. Consider the two-sided hypothesis test:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0,$$

where

$$\alpha = \mathbb{P}(\overline{X} \in R \mid \mu = \mu_0), \quad R = \left\{ \overline{X} \;\middle|\; |\overline{X} - \mu_0| \geq \xi \right\}.$$

Given a specific observation $\overline{x}$, we reject H_0 if $|\overline{x} - \mu_0| \geq \xi$. However, instead of using a fixed rejection threshold ξ, we aim to compute the smallest significance level α for which H_0 would be rejected—this is the p-value. A smaller ξ enlarges the rejection region R, increasing α, while a larger ξ shrinks R, decreasing α. Thus, the smallest possible rejection region consistent with the observed $\overline{x}$ is:

$$R = \left\{ \overline{X} \;\middle|\; |\overline{X} - \mu_0| \geq |\overline{x} - \mu_0| \right\},$$

and the p-value is given by

$$\mathbb{P}\left(|\overline{X} - \mu_0| \geq |\overline{x} - \mu_0| \;\middle|\; \mu = \mu_0 \right).$$

For large sample sizes ($n \geq 30$) and known σ, the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

follows a standard normal distribution, so the p-value becomes

$$\mathbb{P}\left(|Z| \geq \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right) = \mathbb{P}\left(Z \geq \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right) + \mathbb{P}\left(Z \leq -\left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right).$$

If σ is unknown, we estimate it using the sample standard deviation s, and approximate:

$$\mathbb{P}\left(|Z| \geq \left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right|\right) \approx \mathbb{P}\left(Z \geq \left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right|\right) + \mathbb{P}\left(Z \leq -\left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right|\right).$$

For a **right-tailed** test:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0,$$

we compute:

$$\mathbb{P}\left(\overline{X} - \mu_0 \geq \overline{x} - \mu_0 \mid \mu = \mu_0\right) = \underbrace{\mathbb{P}\left(Z \geq \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)}_{\sigma \text{ known}} \approx \underbrace{\mathbb{P}\left(Z \geq \frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right)}_{\sigma \text{ unknown}}.$$

For a **left-tailed** test:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0,$$

we compute:

$$\mathbb{P}\left(\overline{X} - \mu_0 \leq \overline{x} - \mu_0 \mid \mu = \mu_0\right) = \underbrace{\mathbb{P}\left(Z \leq \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)}_{\sigma \text{ known}} \approx \underbrace{\mathbb{P}\left(Z \leq \frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right)}_{\sigma \text{ unknown}}.$$

Now suppose $X_1, \ldots, X_n$ are i.i.d. normal random variables, and $n < 30$. Then $\overline{X}$ is normally distributed, and we proceed as follows:

For the two-sided test:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0,$$

when σ is known:

$$\mathbb{P}\left(|Z| \geq \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right) = \mathbb{P}\left(Z \geq |\cdot|\right) + \mathbb{P}\left(Z \leq -|\cdot|\right).$$

When σ is unknown, we instead use the t-distribution:

$$\mathbb{P}\left(|T| \geq \left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right|\right) = \mathbb{P}\left(T \geq |\cdot|\right) + \mathbb{P}\left(T \leq -|\cdot|\right),$$

where $T \sim t_{n-1}$.

For the one-sided tests with small n:

Right-tailed:

$$\mathbb{P}\left(\overline{X} - \mu_0 \geq \overline{x} - \mu_0 \mid \mu = \mu_0\right) = \underbrace{\mathbb{P}\left(Z \geq \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)}_{\sigma \text{ known}} \approx \underbrace{\mathbb{P}\left(T \geq \frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right)}_{\sigma \text{ unknown}}.$$

Left-tailed:

$$\mathbb{P}\left(\overline{X} - \mu_0 \leq \overline{x} - \mu_0 \mid \mu = \mu_0\right) = \underbrace{\mathbb{P}\left(Z \leq \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)}_{\sigma \text{ known}} \approx \underbrace{\mathbb{P}\left(T \leq \frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right)}_{\sigma \text{ unknown}}.$$

We can then use the p-value approach to conduct a statistical test. Assume that H_0 is true. We estimate the sample mean $\overline{x}$ from the observed data, compute the corresponding p-value, and compare it with a predefined significance level α. If the p-value is smaller than α, we reject H_0; otherwise, we fail to reject H_0.

Example. The average temperature of Hong Kong in February is 18°C. Has this year been unusual? Has this year been colder (assume σ unknown)? Assume temperature in February follows $\mathcal{N}(\mu, \sigma^2)$ and $\sigma = 3$°C. Suppose $\alpha = 0.05$.

Day	1	6	11	16	21	26
Temp (°C)	15	15	19	18	8	17

Solution: Here we have $H_0 : \mu = 18$, $H_1 : \mu \neq 18$.

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

By calculation, we have

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{15.33 - 18}{3/\sqrt{6}} \approx 2.18$$

$$\begin{aligned} p\text{-value} &= \mathbb{P}\left(Z \geq \left|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) + \mathbb{P}\left(Z \leq -\left|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \\ &= \mathbb{P}(Z \geq 2.18) + \mathbb{P}(Z \leq -2.18) \\ &= (1 - 0.9854) + (1 - 0.9854) \\ &= 0.0292 \\ &< \alpha = 0.05 \end{aligned}$$

Thus, we reject H_0.

If σ is unknown, for $H_0 : \mu = 18$, $H_1 : \mu < 18$, we have

$$s^2 = \frac{(15 - 15.33)^2 + \cdots + (17 - 15.33)^2}{6 - 1} \approx 15.47, \quad s \approx 3.93$$

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{15.33 - 18}{3.93/\sqrt{6}} \approx -1.71$$

$$p\text{-value} = \mathbb{P}\left(T \leq \frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right) = \mathbb{P}(T > 1.71) = 0.07 > \alpha = 0.05$$

Thus, we accept H_0.

Chapter 6

Comparing Population

Suppose we want to explore whether male and female college students have different driving behaviors in terms of the **mean fastest driving speed**. Based on a survey of 18 male students and 20 female students, we find that the mean fastest speeds driven by male and female students are 105 kph and 90 kph, respectively. Can we claim that the mean fastest speed driven by male college students is different from that of female college students? Or, more specifically, can we claim that the mean fastest speed driven by male students is **faster** than that of female students?

This requires a different technique, which will be introduced in this chapter.

Suppose $X_1, \cdots, X_{n_x}$ are independent random variables with common mean μ_x and variance σ_x^2, and $Y_1, \cdots, Y_{n_y}$ are independent with common mean μ_y and variance σ_y^2. We also assume that the X_i's and Y_i's are mutually independent.

Suppose $n_x, n_y \geq 30$, and we are testing at a significance level α. Then, as discussed in the previous chapter, we may consider:

Two-sided test:

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x \neq \mu_y$$

One-sided tests:

$$\text{Right-sided: } H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x > \mu_y$$

$$\text{Left-sided: } H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x < \mu_y$$

We can also rewrite these hypotheses in terms of the difference in means:

$$H_0 : \mu_x - \mu_y = 0 \quad \text{vs.} \quad H_1 : \mu_x - \mu_y \neq 0$$

By the Central Limit Theorem (CLT), we have:

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n_x}\right), \quad \overline{Y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma_y^2}{n_y}\right) \Longrightarrow \overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

Assuming $H_0 : \mu_x - \mu_y = 0$ is true, then the test statistic is:

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sigma_D} = \frac{\overline{X} - \overline{Y}}{\sigma_D} \sim \mathcal{N}(0, 1) \Longrightarrow \sigma_D = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Then, we can proceed using the p-value approach or rejection region approach as before.

Critical Value Approach

$$\alpha = \mathbb{P}\left(\left|(\overline{X} - \overline{Y}) - 0\right| \geq \xi\right) = \mathbb{P}\left(\left|\frac{\overline{X} - \overline{Y}}{\sigma_D}\right| \geq \frac{\xi}{\sigma_D}\right) = \mathbb{P}\left(|Z| \geq z_{\frac{\alpha}{2}}\right)$$

When σ_x and σ_y are known, and given observed sample means $\overline{x}$ and $\overline{y}$, if

$$\left|\frac{\overline{x}-\overline{y}}{\sigma_D}\right| \geq z_{\frac{\alpha}{2}},$$

then we **reject** H_0; otherwise, we **fail to reject** H_0.

When σ_x and σ_y are unknown, we approximate them with the sample standard deviations s_x and s_y. If

$$\left|\frac{\overline{x}-\overline{y}}{s_D}\right| \geq z_{\frac{\alpha}{2}},$$

then we **reject** H_0; otherwise, we **fail to reject** it.

Note that

$$s_D = \sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

p-Value Approach

Given specific values $\overline{x}$ and $\overline{y}$, the p-value is computed as

$$\mathbb{P}\left(|Z| \geq \left|\frac{\overline{x}-\overline{y}}{\sigma_D}\right|\right) = \mathbb{P}\left(Z \geq \left|\frac{\overline{x}-\overline{y}}{\sigma_D}\right|\right) + \mathbb{P}\left(Z \leq -\left|\frac{\overline{x}-\overline{y}}{\sigma_D}\right|\right).$$

If σ is unknown, we estimate it using the sample standard deviations s_x and s_y, and approximate:

$$\mathbb{P}\left(|Z| \geq \left|\frac{\overline{x}-\overline{y}}{s_D}\right|\right) = \mathbb{P}\left(Z \geq \left|\frac{\overline{x}-\overline{y}}{s_D}\right|\right) + \mathbb{P}\left(Z \leq -\left|\frac{\overline{x}-\overline{y}}{s_D}\right|\right).$$

If the p-value is smaller than α, we reject H_0; otherwise, we fail to reject H_0.

As in the previous chapter, we consider the case where $n_x, n_y < 30$. If both X_i and Y_i are normally distributed and σ_x^2, σ_y^2 are known, then under $H_0 : \mu_x - \mu_y = 0$, we have:

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n_x}\right), \quad \overline{Y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma_y^2}{n_y}\right) \Rightarrow \overline{X}-\overline{Y} \sim \mathcal{N}\left(\mu_x-\mu_y, \frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}\right)$$

$$Z = \frac{\overline{X}-\overline{Y}}{\sigma_D} \sim \mathcal{N}(0,1), \quad \text{where } \sigma_D = \sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

Given specific $\overline{x}$ and $\overline{y}$, if

$$\left|\frac{\overline{x}-\overline{y}}{\sigma_D}\right| \geq z_{\frac{\alpha}{2}},$$

then we reject H_0; otherwise, we fail to reject it.

Suppose $X_1, \ldots, X_{n_x} \sim \mathcal{N}(\mu_x, \sigma^2)$, and $Y_1, \ldots, Y_{n_y} \sim \mathcal{N}(\mu_y, \sigma^2)$, with mutually independent samples, and σ^2 unknown but equal. For $n_x, n_y < 30$, we have:

$$T = \frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{S_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \sim t(n_x+n_y-2)$$

where the pooled variance is given by:

$$S_D^2 = \frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2}$$

and S_x^2, S_y^2 are the unbiased sample variances.

Assuming $H_0 : \mu_x - \mu_y = 0$, we have:

$$\alpha = \mathbb{P}\left(\left|\overline{X}-\overline{Y}\right| \geq \xi\right) = \mathbb{P}\left(|T| \geq \frac{\xi}{S_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right) = \mathbb{P}\left(|T| \geq t_{\frac{\alpha}{2}}\right)$$

Given specific values $\overline{x}$ and $\overline{y}$, if

$$\frac{|\overline{x}-\overline{y}|}{S_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \geq t_{\frac{\alpha}{2}},$$

then we reject H_0; otherwise, we fail to reject H_0.

We also consider the one-sided case. Suppose $X_1, \cdots, X_{n_x}$ are independent random variables with common mean μ_x and variance σ_x^2, and $Y_1, \cdots, Y_{n_y}$ are independent with common mean μ_y and variance σ_y^2. We also assume that the X_i's and Y_i's are mutually independent. Consider the following:

$$\text{Right-sided: } H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x > \mu_y \Longrightarrow H_0 : \mu_x - \mu_y = 0 \quad \text{vs.} \quad H_1 : \mu_x - \mu_y > 0$$

For large $n_x, n_y \geq 30$ or $n_x, n_y < 30$ with normal PDF and known σ_x, σ_y, we can use the critical value z_α. That is, given specific $\overline{x}$ and $\overline{y}$, if

$$\frac{\overline{x}-\overline{y}}{\sigma_D} \geq z_{\frac{\alpha}{2}},$$

then we reject H_0; otherwise, we fail to reject it. Here we have

$$\sigma_D = \sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}.$$

If σ_x, σ_y are unknown, then they are replaced by s_x^2, s_y^2.

We can also use the p-value approach, where

$$p\text{-value} = \mathbb{P}\left(Z \geq \frac{\overline{x}-\overline{y}}{\sigma_D}\right).$$

For $n_x, n_y < 30$ with normal PDF, and unknown $\sigma_x^2 = \sigma_y^2 = \sigma^2$, we use the critical value t_α. That is, given specific $\overline{x}, \overline{y}$, if

$$\frac{\overline{x}-\overline{y}}{S_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \geq t_{\frac{\alpha}{2}},$$

then we reject H_0; otherwise, we fail to reject H_0.

Here we have the p-value as

$$\mathbb{P}\left(T \geq \frac{\overline{x}-\overline{y}}{S_D\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}\right),$$

where

$$T \sim t(n_x + n_y - 2).$$

Example. A thermometer reports readings:

Mon	23.5	23.3	21.3	22.1	23.7
Tue	22.8	24.5	23.7		

Has the temperature increased? Suppose the readings follow $\mathcal{N}(\mu_M, \sigma^2)$ on Monday and $\mathcal{N}(\mu_T, \sigma^2)$ on Tuesday, and are independent, with $\sigma = 1°\text{C}$. Set $\alpha = 0.05$, and use the p-value approach.

Solution: Consider

$$H_0 : \mu_T - \mu_M = 0 \quad \text{vs.} \quad H_1 : \mu_T - \mu_M > 0$$

Denote by X_i the readings on Monday and Y_i the readings on Tuesday, where $n_x = 5$, $n_y = 3$, and the test statistic is $\overline{Y} - \overline{X}$. Then we have

$$p\text{-value} = \mathbb{P}\left(Z \geq \frac{\overline{y}-\overline{x}}{\sigma_D}\right)$$

where

$$\sigma_D = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}.$$

Plugging in $\overline{x} = 22.78$, $\overline{y} = 23.67$, $\sigma = 1$, $n_x = 5$, $n_y = 3$, we have

$$p\text{-value} = \mathbb{P}\left(Z \geq \frac{0.89}{0.73}\right) \approx \mathbb{P}(Z \geq 1.22) = 0.1112 > \alpha = 0.05$$

Thus, we do not reject H_0.

However, if σ is unknown, we have

$$T \sim t(n_x + n_y - 2) = t(6).$$

$$S_D^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} = 0.98$$

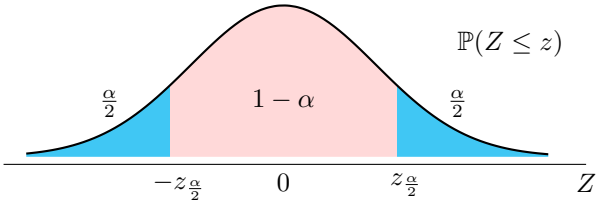
$$\frac{\overline{y} - \overline{x}}{S_D\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \approx \frac{0.89}{0.98 \times 0.73} \approx 1.23 < t_{0.05}(6) = 1.94$$

Again, we do not reject H_0.

Remark. The content of paired t test is not covered here.

Appendix A

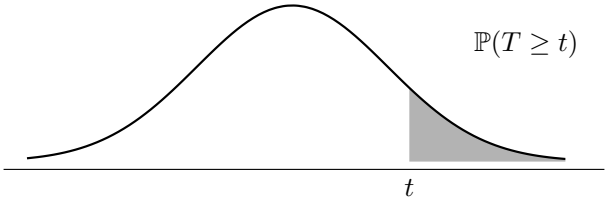
Z TABLE



	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

Appendix B

Student's t-Distribution



CI	-	-	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
df/p	0.40	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.289	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646