MATH 3423: Statistical Inference

HU-HTAKM
Website: https://htakm.github.io/htakm_test/

Last major change: October 31, 2024
Last small update: October 31, 2024

This is MATH 3423 lecture note made by me. Note that some of the notations are slightly different with what is used in the course for better clarity. For example:

Name	My notation	Ycw's notation
Transpose	A^T	A'
Fisher Information	$\mathcal{I}_X(\theta)$	$I_X(\theta)$

Contents

1 Preliminary **5**
1.1 Random sample and parametric distribution 5
1.2 Moments 6
1.3 Moment generating function 8
1.4 Limit Theorems 10
1.5 Commonly used distribution 15

2 Point Estimation **23**
2.1 Methods of Moments Estimation 23
2.2 Maximum Likelihood Estimation 26

Chapter 1

Preliminary

Statistical inference is a statistical process which investigates how to use the information from the data to make an inference about the distribution of the random variable of our interest. In MATH 3423, we will focus on two core concepts of statistical inference: point estimation and hypothesis testing.

1.1 Random sample and parametric distribution

To make a statistical inference about the distribution of the random variable X of our interest, we need to draw a sample of data.

Definition 1.1. Denote the first observation by X_1, the second by X_2, and so on. A set of random variables $\{X_1, \cdots, X_n\}$ are called a **random sample** of size n from the common distribution of X with a PMF $p_X(x)$ or PDF $f_X(x)$ if they are independent and identically distributed (i.i.d.).

Remark 1.1.1. Random variables $X_1, \cdots, X_n$ are assumed to be observable with known actual values $x_1, \cdots, x_n$ respectively.

Under the random sampling setting, it is easy to get the following lemma.

Lemma 1.2. Given a random sample $\{X_1, \cdots, X_n\}$ of a common distribution X.

1. If the random sample is discrete with a common PMF $p_X(x)$, then the joint PMF of random sample can be obtained by:
$$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$
2. If the random sample is continuous with a common PDF $f_X(x)$, then the joint PDF of random sample can be obtained by:
$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

The underlying distribution of X is assumed to be unknown or partially known in practice. In most situation, it is reasonable to assume that the form of the PMF p_X or PDF f_X of the distribution is known but it contains some unknown parameters θ.

Definition 1.3. Parametric distribution is a distribution which the PMF p_X and PDF f_X contains some unknown parameters θ. The PMF or PDF is said to be **parametric**.

Remark 1.3.1. Instead of using parametric distributions of the data, we may assume that the form of the distribution is unknown but that the distribution has some properties. E.g. A distribution is continuous. This distribution is called a **non-parametric distribution** and the corresponding statistical method is called **non-parametric statistical approach**. If parameters are involved but the form of the distribution is unknown, then such a distribution is called **semi-parametric distribution** and the corresponding method is called **semi-parametric statistical approach**.

Example 1.1. Data are usually assumed from the normal distribution with mean μ and variance σ^2, where the parameter:

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

is unknown but fixed.

Lemma 1.4. Given a random sample $\{X_1, \cdots, X_n\}$ of a common distribution X.

1. If the random sample is discrete with a common parametric PMF $p_X(x|\theta)$, then the joint PMF of random sample can be obtained by:
$$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} p_X(x_i|\theta)$$
2. If the random sample is continuous with a common parametric PDF $p_X(x|\theta)$, then the joint PDF of random sample can be obtained by:
$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

Under parametric setting, we can see that uncertainty of the distribution is the uncertainty of its parameters. One of the central problem in statistics is how to determine which function of data is the best one to estimate θ.

Definition 1.5. If $\mathbf{X} = (X_1 \ \cdots \ X_n)^T$ is a random vector and $T(\cdot)$ is a real-valued or vector-valued function such that for all $\mathbf{X} \in \Omega$, $T(\mathbf{X})$ does not contain any unknown parameters, then $T(\mathbf{X})$ is called a **statistic**.

1. If we use the statistic $T(\mathbf{X})$ to estimate an unknown parameter θ, then we call $T(\mathbf{X})$ and $T(\mathbf{x})$ an **estimator** and an **estimate** of θ respectively, where $\mathbf{x}$ is an observed value of $\mathbf{X}$.

Remark 1.5.1. We usually denote an estimator of θ by $\hat{\theta}(\mathbf{X})$ or simply $\hat{\theta}$.

Remark 1.5.2. Since $T(\mathbf{X})$ is also random, we have a distribution of $T(\mathbf{X})$ called **sampling distribution**.

1.2 Moments

The population moments of a distribution play an important role in theoretical and applied statistics.

Definition 1.6. For each positive integer k, the **k-th population moment** of X about 0, denoted by μ'_k, is defined as:

$$\mu'_k = \mathrm{E}(X^k)$$

if the expectation exists.
When $k = 1$, μ'_1 is the population mean $\mu = E(X)$ of X.

Definition 1.7. The **k-th population central moment** of X, denoted by μ_k, is defined as:

$$\mu_k = \mathrm{E}(X - \mu)^k$$

if the expectation exists.
When $k = 2$, μ_2 is the **population variance** σ^2 of X.

Remark 1.7.1. Don't confuse the population mean μ and the k-th population central moment μ_k!

Example 1.2. We have terminologies for some useful population moments.

1. **Skewness**: μ_3, a measure of symmetry or skewness.
   (a) If $\mu_3 < 0$, then the curve is left-skewed (tail is on the left).
   (b) If $\mu_3 > 0$, then the curve is right-skewed (tail is on the right).
   (c) If $\mu_3 = 0$, the the curve is symmetrical.

   The ratio $\frac{\mu_3}{\sigma^3}$ is called the **coefficient of skewness**.
2. **Kurtosis**: μ_4, a measure of excess or kurtosis, which is the degree of flatness of a density near its center. We called $\frac{\mu_4}{\sigma^4} - 3$ the **coefficient of kurtosis**.
   (a) If $\frac{\mu_4}{\sigma^4} - 3 > 0$, then the density has a sharper peak than the density of the normal curve.
   (b) If $\frac{\mu_4}{\sigma^4} - 3 < 0$, then the density has a flatter peak than the density of the normal curve.

We usually use sample moments to estimate the population moments.

Definition 1.8. Let $X_1, \cdots, X_n$ be a random sample of size n. For each positive integer k,

1. the k**-th sample moment** about 0, denoted by $\overline{X^k}$, is defined as:
$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$
2. the k**-th sample moment** about $\overline{X}$, denoted by S_n^k, is defined as:
$$S_n^k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

Example 1.3. When $k = 1$, $\overline{X}$ is called the **sample mean** of X. When $k = 2$, S_n^2 is called the **sample variance**. However, we usually don't use this version of sample variance. Instead, we use another sample variance, denoted by S_{n-1}^2, which is defined by:
$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
We do this because S_{n-1}^2 is unbiased while S_n^2 is not unbiased.

Lemma 1.9. Let $X_1, \cdots, X_n$ be a random sample of size n. We have:
$$\mathrm{E}(\overline{X^k}) = \mu_k'$$
if μ_k' exists. We also have:
$$\mathrm{Var}(\overline{X^k}) = \frac{1}{n}[\mu_{2k}' - (\mu_k')^2]$$

Proof.
Since $X_1, \cdots, X_n$ have the same distribution, we have:
$$\mathrm{E}(X_1^k) = \cdots = \mathrm{E}(X_n^k) = \mathrm{E}(X^k) = \mu_k'$$
Therefore, we have:
$$\mathrm{E}(\overline{X^k}) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^k\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i^k) = \frac{n\mu_k'}{n} = \mu_k'$$
Since $X_1^k, \cdots, X_n^k$ are independent,
$$\begin{aligned}\mathrm{Var}(\overline{X^k}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i^k) &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathrm{E}(X_i^{2k}) - (\mathrm{E}(X_i^k)^2)\right] \\ &= \frac{1}{n}[\mu_{2k}' - (\mu_k')^2]\end{aligned}$$
□

1.3 Moment generating function

It would be useful if we have a function that could generate all moments.

Definition 1.10. The **moment generating function** (MGF) of a random variable X, denoted by $M_X(t)$, is:

$$M_X(t) = \mathrm{E}(e^{tX})$$

if the expectation exists for t in some neighbourhood of 0.

Remark 1.10.1. More precisely, there exists $h > 0$ such that for all t in $(-h, h)$, $\mathrm{E}(e^{tX})$ exists.

Remark 1.10.2. MGF of X may not always exist. However, if it exists, then $M_X(t)$ is continuously differentiable in some neighbourhood of the origin.

Remark 1.10.3. If we replace e^{tX} by its taylor series, then we would get:

$$M_X(t) = \mathrm{E}\left[\sum_{i=0}^{\infty} \frac{(tX)^i}{i!}\right] = \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathrm{E}(X^i) = \sum_{i=0}^{\infty} \frac{t^i}{i!}\mu_i'$$

Lemma 1.11. If $M_X(t)$ is the MGF of a random variable X, then:

$$\frac{\partial^k}{\partial t^k} M_X(t)\bigg|_{t=0} = \mathrm{E}(X^k) = \mu_k'$$

Proof.
From the taylor series of MGF, we would see that:

$$\frac{\partial^k}{\partial t^k} M_X(t) = \sum_{i=k}^{\infty} \frac{t^{i-k}}{(i-k)!}\mathrm{E}(X^i)$$

Therefore, by having $t = 0$, we would have:

$$\frac{\partial^k}{\partial t^k} M_X(t)\bigg|_{t=0} = \mathrm{E}(X^k)$$

□

Example 1.4. What is the MGF of $X \sim \mathrm{Bern}(p)$? We have:

$$M_X(t) = \mathrm{E}(e^{tX}) = e^{t(0)}(1-p) + e^{t(1)}(p) = pe^t + 1 - p$$

Lemma 1.12. If random variables X and Y are independent, then:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

Proof.
Since X and Y are independent,

$$M_{X+Y}(t) = \mathrm{E}(e^{t(X+Y)}) = \mathrm{E}(e^{tX})\,\mathrm{E}(e^{tY}) = M_X(t)M_Y(t)$$

□

Example 1.5. By definition, if $Y = \mathrm{Bin}(n, p)$, then $Y = X_1 + \cdots + X_n$ where $X_i \sim \mathrm{Bern}(p)$ for all i and they are independent. Therefore,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = (pe^t + 1 - p)^n$$

Example 1.6. Consider $X \sim \text{Poisson}(\lambda)$, The MGF of X can be obtained by:

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{\lambda(e^t-1)}$$

Example 1.7. Consider $X \sim \text{Exp}(\lambda)$. If $t < \lambda$, we have:

$$M_X(t) = \text{E}(e^{tX}) = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}\, dx = \lambda \int_0^{\infty} e^{-(\lambda-t)x}\, dx = \frac{\lambda}{\lambda - t}$$

Example 1.8. What is the MGF of $X \sim \text{N}(\mu, \sigma^2)$? We may first find the MGF of $Z \sim \text{N}(0, 1)$.

$$\begin{aligned} M_Z(t) = \text{E}(e^{tZ}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2tz)}\, dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}((z-t)^2-t^2)}\, dz \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}\, dz \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

Therefore, by having $X = \sigma Z + \mu$, we have:

$$M_X(t) = \text{E}(e^{tX}) = e^{\mu t} \,\text{E}(e^{t\sigma Z}) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Example 1.9. If $X \sim \text{NBin}(r, p)$, then for $t < -\ln(1-p)$:

$$M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t} \right)^r$$

If $X \sim \text{U}[a, b]$, then:

$$M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$$

If $X \sim \text{Gamma}(\alpha, \beta)$, then for $t < \beta$:

$$M_X(t) = \left(\frac{\beta}{\beta - t} \right)^{\alpha}$$

Ultimately, the reason why we use moment generating function is the following fact.

Theorem 1.13. (**Uniqueness of MGF**) Let X and Y be two random variables. Suppose that their MGFs exist and are equal for all $t \in (-h, h)$ for some $h > 0$, then we have the distribution function F_X and F_Y are equal.

This means that by knowing the MGF of a particular random variable X, we can know its distribution.

Example 1.10. Assume that $X_1, \cdots, X_n$ are independent and $X_i \sim \text{Bin}(m_i, p)$ for all $i = 1, \cdots, n$. Then we have:

$$M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (pe^t + 1 - p)^{m_i} = (pe^t + 1 - p)^{\sum_{i=1}^n m_i}$$

Therefore, we have $X_1 + \cdots + X_n \sim \text{Bin}(\sum_{i=1}^n m_i, p)$.

Example 1.11. Assume that $X_1, \cdots, X_n$ are independent and $X_i \sim \text{Poisson}(\lambda_i)$ for all $i = 1, \cdots, n$. Then we have:

$$M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{\sum_{i=1}^n \lambda_i(e^t-1)}$$

Therefore, we have $X_1 + \cdots + X_n \sim \text{Poisson}(\sum_{i=1}^n \lambda_i)$.

Example 1.12. Similarly, given a set of independent random variables $\{X_1, \cdots, X_n\}$.

1. If $X_i \sim \text{NBin}(r_i, p)$, then $X_1 + \cdots + X_n \sim \text{NBin}(\sum_{i=1}^n r_i, p)$.
2. If $X_i \sim \text{N}(\mu_i, \sigma_i^2)$, then $X_1 + \cdots + X_n \sim \text{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.
3. If $X_i \sim \text{Gamma}(\alpha_i, \beta)$, then $X_1 + \cdots + X_n \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

Remark 1.13.1. Not all sum of distributions will result in the same type of distribution.

More generally, we would deal with problems of limiting distribution.

Theorem 1.14. Suppose the $\{X_n\}$ is a sequence of random variables, each with MGF $M_{X_n}(t)$. If we have:

$$\lim_{n\to\infty} M_{X_n}(t) = M_Y(t)$$

for all t in a neighbourhood of 0, where $M_Y(t)$ is a MGF for some random variables Y, then there is an unique distribution function F_Y with corresponding $M_Y(t)$ such that:

$$\lim_{n\to\infty} F_{X_n}(y) = F_Y(y)$$

for all y where $F_Y(y)$ is continuous. We denote by $X_n \to Y$ or $X_n \xrightarrow{D} X$.

Remark 1.14.1. Simply, the limiting distribution of X_n is equal to the distribution of Y.

We may define the limiting convergence in truly theoretical way.

Definition 1.15. A sequence of random variables $\{X_n\}$ **converges in distribution** to a random variable X, denoted by $X_n \xrightarrow{D} X$, if for all continuity point x of F_X, as $n \to \infty$,

$$F_{X_n}(x) \to F_X(x)$$

We also have a more strict convergence.

Definition 1.16. A sequence of random variables $\{X_n\}$ **converges in probability** to a random variable X, denoted by $X_n \xrightarrow{\text{P}} X$, if for any $\varepsilon > 0$, as $n \to \infty$,

$$\text{P}(|X_n - X| < \varepsilon) \to 1 \qquad \text{P}(|X_n - X| \geq \varepsilon) \to 0$$

Remark 1.16.1. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$. The converse is not necessarily true.

1.4 Limit Theorems

Using the last two theorems, the next two theorems are very useful in both statistics and probability theory by giving us approximate distribution of an average without a lot of distributional assumption.

Theorem 1.17. (**Weak Law of Large Numbers (WLLN)**) Let $\{X_n\}$ be a sequence of i.i.d. random variables. Let $\text{E}(X_i) = \mu$ for all $i = 1, 2, \cdots$. Define by $\overline{X}$ the sample mean of the random variables. Then as $n \to \infty$

$$\overline{X} \xrightarrow{D} \mu$$

Theorem 1.18. (**Classical Central Limit Theorem (CLT)**) Let $\{X_n\}$ be a sequence of i.i.d. random variables whose MGFs exist in a neighbourhood of 0. Let $\text{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 > 0$ for all $i = 1, 2, \cdots$. Define by $\overline{X}$ the sample mean of the random variables. Then as $n \to \infty$,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \to \text{N}(0, 1)$$

Remark 1.18.1. This is just an abuse of notations.

This works generally to most of the distribution. However, it is probably very tedious to find the MGF. We cam apply the following version of CLT instead.

Theorem 1.19. (**Lévy-Linderberg Central Limit Theorem**) Let $\{X_n\}$ be a sequence of i.i.d. random variables with common population means μ and common population variance σ^2. Assume that $0 < \sigma^2 < \infty$. Define by $\overline{X}$ the sample mean of the random variables. Then as $n \to \infty$,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \to \mathrm{N}(0, 1)$$

Sometimes, we will deal with a function of multiple random variables. We must establish how they converges.

Theorem 1.20. (**Slutsky's Theorem**) If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{\mathrm{P}} c$, then:

1. $X_n + Y_n \xrightarrow{D} X + c$
2. $X_n Y_n \xrightarrow{D} cX$
3. $\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}$ if $c \neq 0$.

Example 1.13. Assume that $X_i \sim \mathrm{Bern}(p)$ for all i. We want to estimate the unknown p. We have common mean $\mu = p$ and common variance $\sigma^2 = p(1-p)$. By applying CLT, as $n \to \infty$,

$$\overline{X} \to \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$$

Therefore, we can use normal distribution to approximate the unknown parameter. We want an estimation that we can confident about, and commonly we use probability 0.95.

$$0.95 = \mathrm{P}\left(-z_{0.025} \leq \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0.025}\right) = \mathrm{P}\left((\overline{X} - p)^2 \leq z_{0.025}^2 \frac{p(1-p)}{n}\right)$$

Solving the inequality, we would find an interval that estimates the parameter p. However, this is highly inconvenient. We may use another method. Let us replace $\sqrt{\frac{p(1-p)}{n}}$ with $\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$. As $n \to \infty$,

$$\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{p(1-p)}{\overline{X}(1-\overline{X})}} \to \sqrt{\frac{p(1-p)}{n}}$$

since by Slutsky's Theorem, $\sqrt{\frac{p(1-p)}{\overline{X}(1-\overline{X})}} \to 1$ as $\overline{X} \to p$ by WLLN. We have:

$$0.95 = \mathrm{P}\left(-z_{0.025} \leq \frac{\overline{X} - p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \leq z_{0.025}\right) = \mathrm{P}\left(\overline{X} - z_{0.025}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \leq p \leq \overline{X} + z_{0.025}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right)$$

Example 1.14. In a survey before an election, a poll was taken of 300 potential voters. Among them, 120 said that they would vote for candidate A. Determine a 95% confidence interval for the population proportion p_A of voters who would vote for candidate A in the election.
From the poll, we have a point estimate $\overline{x} = \hat{p}_A = \frac{120}{300} = 0.4$. From the last example, we have found that the 95% confidence interval is:

$$\left(\overline{x} - z_{0.025}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}, \overline{x} + z_{0.025}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right) \approx (0.3446, 0.4554)$$

Equivalently, the percentage of voters of candidate A would be from 34.46% to 45.54%, with a margin of error 5.54%.

Example 1.15. Following the previous example. Assume that we have been given a margin of error D. How many data should we collect in order to to have the margin of error?
From how we find the margin of error:

$$z_{0.025}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} = D \implies n = p(1-p)\frac{z_{0.025}^2}{D^2}$$

Since $p(1-p) \leq \frac{1}{4}$, if we specify that $D = 0.05$, we have:

$$n \leq \frac{z_{0.025}^2}{4D^2} = \frac{1.96^2}{4(0.05)^2} \leq \frac{2^2}{4(0.05)^2} = 400$$

We may use this to determine whether we have obtained enough data.
Assume that we have n^* respondents. Is it enough? The number of required respondents is obtained by:

$$n_{\text{required}} = \frac{\overline{x^*}(1-\overline{x^*})z_{0.025}^2}{D^2}$$

If $n^* < n_{\text{required}}$, then the current number of data is not enough. We would need to find more respondents.
If $n^* \geq n_{\text{required}}$, then it is enough.

Example 1.16. We try to use Poisson random variables to prove that as $n \to \infty$,

$$e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} \to \frac{1}{2}$$

Let $\{X_n\}$ be a sequence of i.i.d. random variables and $X_i \sim \text{Poisson}(1)$ for $i = 1, 2, \cdots$. Let $Y_n = \sum_{i=1}^{n} X_i$. By CLT, we have:

$$\frac{Y_n - n}{\sqrt{n}} \to \text{N}(0,1)$$

Therefore, since $Y_n \sim \text{Poisson}(n)$, we have:

$$e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} = \text{P}(Y_n \leq n) = \text{P}(\frac{Y_n - n}{\sqrt{n}} \leq 0) \to \frac{1}{2}$$

Example 1.17. Given a sequence of i.i.d. random variables $\{X_n\}$. We want to find the asymptotic distribution for the k-th sample moment $\overline{X^k}$ as $n \to \infty$. Notice that X_i^k are independent for $i = 1, 2, \cdots$. By CLT,

$$\frac{\sqrt{n}(\overline{X^k} - \mu_k')}{\sqrt{\mu_{2k}' - (\mu_k')^2}} \to \text{N}(0,1)$$

Therefore, the asymptotic distribution for $\overline{X^k}$ when $n \to \infty$ is $\text{N}(\mu_k', \frac{1}{n}(\mu_{2k}' - (\mu_k')^2))$.

Central Limit Theorem can provide us a limiting standard normal distribution of sample mean. However, we usually deal with some functions of the sample mean.

Theorem 1.21. (**Continuous mapping theorem**) Let $\{X_n\}$ be a sequence of random variables and X be a random variable. Suppose there is a function g has a set of discontinuity points D_g such that $\text{P}(X \in D_g) = 0$, then:

1. If $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$.
2. If $X_n \xrightarrow{\text{P}} X$, then $g(X_n) \xrightarrow{\text{P}} g(X)$.

Theorem 1.22. (**Delta method**) Let $\{X_n\}$ be a sequence of random variables such that for constants a and $b > 0$, as $n \to \infty$,

$$\sqrt{n}(X_n - a) \to \text{N}(0, b^2)$$

Then for a given function g, suppose that $g'(a)$ exists and not 0, as $n \to \infty$:

$$\sqrt{n}(g(X_n) - g(a)) \to \text{N}(0, [g'(a)b]^2)$$

Corollary 1.23. If $\overline{X}$ is the sample mean of a random sample $X_1, \cdots, X_n$ of size n from a distribution with a finite mean μ and finite variance $\sigma^2 > 0$. For a given function g, suppose that $g'(\mu)$ exists and is not 0, as $n \to \infty$:

$$\sqrt{n}(g(\overline{X}) - g(\mu)) \to \mathrm{N}(0, [g'(\mu)\sigma]^2)$$

Example 1.18. Assume that there are 70 respondents, 68 of which would vote for one candidate.
If we use the same process from previous examples, we would find that the 95% confidence interval is $(0.9324, 1.0105)$, which is out of the range. In fact, if the point estimate $\hat{p}$ is quite close to 0 or 1, the resulting interval guess may include values that is out of the range of p. This is a poor interval guess.
We take a transformation, say $g(p)$, such that $g(p) \in (-\infty, \infty)$. We may find that since $0 < p < 1$, $\ln p < 0$. Therefore, we can find that:

$$g(p) = \ln(-\ln p) \in (-\infty, \infty)$$

By Delta method,

$$\frac{g(\overline{X}) - g(p)}{g'(p)\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}} \to \mathrm{N}(0, 1)$$

By WLLN and continuous mapping theorem, we can replace $g'(p)$ with $g'(\overline{X})$. Therefore, we have:

$$0.95 = \mathrm{P}\left(-z_{0.025} \leq \frac{g(\overline{X}) - g(p)}{g'(p)\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}} \leq z_{0.025}\right)$$

Solving the formula and we would get the a good 95% confidence interval of p.

Example 1.19. Let $\{X_n\}$ be a sequence of i.i.d. random variables and $X_i \sim \mathrm{Bern}(\theta)$ for $i = 1, 2, \cdots$. Show that:

$$Z_n = 2\sqrt{n}\left(\sin^{-1}\sqrt{\overline{X}} - \sin^{-1}\sqrt{\theta}\right) \to \mathrm{N}(0, 1)$$

Let $g(x) = \sin^{-1}\sqrt{x}$. We may obtain that:

$$g'(x) = \frac{1}{2\sqrt{x}\sqrt{1-x}}$$

We can find that the derivative is well-defined and non-zero for $0 < \theta < 1$ by substituting $x = \theta$. Note that $\mathrm{E}(X_i) = \theta$ and $\mathrm{Var}(X_i) = \theta(1-\theta)$ for $i = 1, \cdots, n$. By Corollary 1.23,

$$\sqrt{n}(g(\overline{X}) - g(\theta)) \to \mathrm{N}\left(0, \frac{1}{4}\right)$$

Since $Z_n = 2\sqrt{n}(g(\overline{X}) - g(\theta))$, we may find that as $n \to \infty$,

$$Z_n \to \mathrm{N}(0, 1)$$

Example 1.20. Let $\{X_n\}$ be a sequence of i.i.d. random variables and $X_i \sim \mathrm{Exp}(\theta)$ for $i = 1, 2, \cdots$. We want to find a variance-stabilizing transformation, which is to find a function $g(x)$ such that the limiting distribution of:

$$Y_n = \sqrt{n}[g(\overline{X_n}) - g(\theta)]$$

does not depend on θ.
We may find that $\mathrm{E}(X_i) = \frac{1}{\theta}$ and $\mathrm{Var}(X_i) = \frac{1}{\theta^2}$ for $i = 1, 2, \cdots$. We claim that $g(x) = \ln x$ is what we want. We can find that:

$$g'(x) = \frac{1}{x}$$

By substituting $x = \frac{1}{\theta}$, we can see that the derivative is non-zero. Applying Corollary 1.23,

$$\sqrt{n}\left(g(\overline{X}) - g\left(\frac{1}{\theta}\right)\right) \to \mathrm{N}(0, 1)$$

Therefore, $g(x) = \ln x$ is the variance-stabilizing transformation.

However, we usually deal with more than 1 variables. Before we extend the theorems into multivariate case, we must first introduce the multivariate normal distribution.

Example 1.21. (**Multivariate Normal Distribution**) $\mathbf{X} \sim \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Given a random vector $\mathbf{X}$. Let the $k \times 1$ vector $\boldsymbol{\mu}$ be the expected value of $\mathbf{X}$ and the $k \times k$ matrix $\boldsymbol{\Sigma}$ be its variance-covariance matrix. Assume that $\boldsymbol{\Sigma}$ is positive-definite. (For all non-zero vectors $\mathbf{z}$ with real entries, we have $\mathbf{z}^T\boldsymbol{\Sigma}\mathbf{z} > 0$) The random vector $\mathbf{X}$ is k-dimensional normal if its PDF is:

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Remark 1.23.1. The i-th row and j-th column of the $k \times k$ variance-covariance matrix $\boldsymbol{\Sigma}$ is the element a_{ij} found by:

$$a_{ij} = \mathrm{cov}(X_i, X_j)$$

Note that if $i = j$, then $\mathrm{cov}(X_i, X_i) = \mathrm{Var}(X_i)$.

Example 1.22. If $k = 2$, then $X \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is bivariate normal.

Lemma 1.24. If $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for any $q \times p$ matrix $\mathbf{A}$, we have:

$$\mathbf{AX} \sim \mathrm{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

Example 1.23. Using this lemma, one can isolate some of the random variables that made up the random vector $\mathbf{X} = (X_1 \ \cdots \ X_p)^T \sim \mathrm{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. For example, setting $(p-1) \times p$ matrix $\mathbf{A}$ as:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} = \left(\begin{array}{c|c} 0 & \\ \vdots & \mathbf{I}_{(p-1)\times(p-1)} \\ 0 & \end{array}\right)$$

We may find that:

$$\mathbf{AX} = \begin{pmatrix} X_2 \\ \vdots \\ X_p \end{pmatrix} \sim \mathrm{N}_{p-1}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

where $\boldsymbol{\mu}_2$ is the mean vector of $(X_2 \ \cdots \ X_p)^T$ and $\boldsymbol{\Sigma}_2$ is the variance-covariance matrix of $(X_2 \ \cdots \ X_p)^T$.

Lemma 1.25. If we have:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right),$$

then X_1 and X_2 are independent if and only if $\sigma_{12} = \sigma_{21} = 0$.

Proof.
From the properties of covariance,

$$\sigma_{12} = \mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_2, X_1) = \sigma_{21}$$

Assume that X_1 and X_2 are independent. We have:

$$\mathrm{cov}(X_1, X_2) = \mathrm{E}((X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2))) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2) = \mathrm{E}(X_1)\,\mathrm{E}(X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2) = 0$$

Therefore, $\sigma_{12} = \sigma_{21} = 0$.
Assume that $\sigma_{12} = \sigma_{21} = 0$. Then:

$$\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\,\mathrm{E}(X_2)$$

Therefore, X_1 and X_2 are uncorrelated. Since X_1 and X_2 are also bivariate normal, they are independent. □

Remark 1.25.1. Two random variables are uncorrelated does not mean they are independent. It is only true if they are bivariate normal.

We may extend the CLT to multivariate case.

Theorem 1.26. (**Multivariate Central Limit Theorem**) Let $\{\mathbf{X}_n = (X_{n1} \ \cdots \ X_{nk})^T \in \mathbb{R}^k\}$ be a sequence of i.i.d. random vectors with variance-covariance matrix $\mathbf{\Sigma}$. We assume that $\mathrm{E}(X_{ij}^2) < \infty$ for $i = 1, 2, \cdots$ and $j = 1, \cdots, k$. Define by $\overline{\mathbf{X}}$ the sample mean of the random vectors. Then as $n \to \infty$,

$$\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{D} \mathrm{N}_k(\mathbf{0}, \mathbf{\Sigma})$$

We may extend the Delta method to multivariate cases.

Theorem 1.27. (**Multivariate 1st-order Delta Method**) Let $\{\mathbf{X}_n \in \mathbb{R}^k\}$ be a sequence of random vectors such that for constant vector $\mathbf{a} \in \mathbb{R}^k$, as $n \to \infty$,

$$\sqrt{n}(\mathbf{X}_n - \mathbf{a}) \xrightarrow{D} \mathbf{U}$$

where $\mathbf{U}$ is a random vector in $\mathbb{R}^k$. If a function $h : \mathbb{R}^k \to \mathbb{R}$ has a derivative $\nabla h(\mathbf{a}) \neq \mathbf{0}$, then as $n \to \infty$:

$$\sqrt{n}(h(\mathbf{X}_n) - h(\mathbf{a})) \xrightarrow{D} \nabla h(\mathbf{a})\mathbf{U}$$

where

$$\nabla h = \left(\frac{\partial}{\partial t_1} h(t_1, \cdots, t_k), \cdots, \frac{\partial}{\partial t_k} h(t_1, \cdots, t_k) \right)$$

1.5 Commonly used distribution

Let us recall some useful distributions.

Example 1.24. (**Binomial distribution**) $X \sim \mathrm{Bin}(n, p)$
Random variable X is a Binomial random variable with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ if it has a PMF for $x = 0, \cdots, n$:

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \mathrm{E}(X) = np \qquad \mathrm{Var}(X) = np(1-p)$$

Example 1.25. (**Geometric distribution**) $X \sim \mathrm{Geom}(p)$
Random variable X is geometric with parameter $p \in [0, 1]$ if it has a PMF for $x = 1, 2, \cdots$:

$$p_X(x) = p(1-p)^{x-1} \qquad \mathrm{E}(X) = \frac{1}{p} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

Example 1.26. (**Poisson distribution**) $X \sim \mathrm{Poisson}(\lambda)$
Random variable X is a Poisson random variable with parameter λ if it has a PMF for $x = 0, 1, \cdots$:

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \mathrm{E}(X) = \lambda \qquad \mathrm{Var}(X) = \lambda$$

Example 1.27. (**Negative Binomial distribution**) $X \sim \mathrm{NBin}(r, p)$
Assume that $X_1, \cdots, X_r$ are independent and $X_i \sim \mathrm{Geom}(p)$ for $i = 1, \cdots, r$. Let $Y = \sum_{i=1}^n X_i$. Random variable Y is negative Binomial with parameters $r > 0$ and $p \in [0, 1]$ if for $x > r$:

$$p_X(x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r \qquad \mathrm{E}(X) = \frac{r}{p} \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

Example 1.28. (**Cauchy distribution**) $X \sim \mathrm{Cauchy}(\theta)$
Random variable X is a Cauchy random variable with parameter θ if it has a PDF:

$$f_X(x) = \frac{1}{\pi(1 + (x-\theta)^2)} \qquad \mathrm{E}(X) \text{ DNE} \qquad \mathrm{Var}(X) \text{ DNE}$$

Example 1.29. (**Uniform distribution**) $X \sim \mathrm{U}[a,b]$
Random variable X is uniform if given $a < b$, it has a PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{Otherwise} \end{cases} \qquad \mathrm{E}(X) = \frac{a+b}{2} \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Example 1.30. (**Exponential distribution**) $X \sim \mathrm{Exp}(\lambda)$
Random variable X is exponential with parameter λ if it has a PDF:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad \mathrm{E}(X) = \frac{1}{\lambda} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Example 1.31. (**Normal distribution / Gaussian distribution**) $X \sim \mathrm{N}(\mu, \sigma^2)$
Random variable X is normal if it has two parameters μ and σ^2, and its PDF and CDF is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \qquad \mathrm{E}(X) = \mu \qquad \mathrm{Var}(X) = \sigma^2$$

Random variable X is standard normal if $\mu = 0$ and $\sigma^2 = 1$. ($X \sim \mathrm{N}(0,1)$)

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad F_X(x) = \Phi(x) = \int_{-\infty}^{x} \phi(u)\,du \qquad \mathrm{E}(X) = 0 \qquad \mathrm{Var}(X) = 1$$

Example 1.32. (**Gamma distribution**) $X \sim \mathrm{Gamma}(\alpha, \beta)$
Random variable X is a gamma random variable with parameters $\alpha > 0$ and $\beta > 0$ if its PDF is:

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x}, & x \geq 0 \\ 0, & \text{Otherwise} \end{cases} \qquad \mathrm{E}(X) = \frac{\alpha}{\beta} \qquad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}$$

Remark 1.27.1. Gamma function $\Gamma(z)$ has the following properties:

1. If z is a positive integer, then $\Gamma(z) = (z-1)!$.
2. For all z, $\Gamma(z+1) = z\Gamma(z)$.
3. If $\Re(z) > 0$, then $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\,dt$.
4. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Example 1.33. (**Beta distribution**) $X \sim \mathrm{Beta}(\alpha, \beta)$
Random variable X is a beta random variable with parameters $\alpha > 0$ and $\beta > 0$ if its PDF is:

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, & x \in (0,1) \\ 0, & \text{Otherwise} \end{cases} \qquad \mathrm{E}(X) = \frac{\alpha}{\alpha+\beta} \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Remark 1.27.2. Beta function $B(z_1, z_2)$ has the following properties:

1. $B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$
2. $B(z_1, z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,dt$

We have some more distributions that is associated with normal distribution. For example, Chi-squared distribution, which is a special case of gamma distribution.

Example 1.34. (**Chi-squared distribution**) $Y \sim \chi^2(n)$
Assume that $X_1, X_2, \cdots, X_n$ are independent and $X_i \sim \mathrm{N}(0,1)$ for $i = 1, \cdots, n$. Let $Y = \sum_{i=1}^{n} X_i^2$. Random variable Y has a χ^2-distribution with n degree of freedom if:

$$f_Y(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})} 2^{-\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, \text{Otherwise} \end{cases} \qquad \mathrm{E}\, Y = n \qquad \mathrm{Var}(Y) = 2n$$

Theorem 1.28. If the random variable $X \sim \mathrm{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$, then the random variable $V = \frac{(X-\mu)^2}{\sigma^2} \sim \chi^2(1)$.

Proof.
By the properties of normal distribution, we get that:

$$\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$$

Therefore, by the definition of chi-squared distribution,

$$V = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$$

□

Example 1.35. Given $Y \sim \chi^2(r)$. How do we find the MGF of Y?
We can find that chi-squared distribution is actually a special case of gamma distribution. We have $\chi^2(r) = \Gamma(\frac{r}{2}, \frac{1}{2})$. Therefore, by substituting, we may get:

$$M_Y(t) = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{r}{2}} = (1-2t)^{-\frac{r}{2}}$$

Example 1.36. Given that $Y \sim \chi^2(r)$. How do we find $\mathrm{E}(Y)$ without using the MGF of Y?
By definition, we may let $Y = \sum_{i=1}^{r} X_i^2$, where $X_i \sim \mathrm{N}(0,1)$. Therefore,

$$\mathrm{E}(Y) = \sum_{i=1}^{r} \mathrm{E}(X_i^2) = \sum_{i=1}^{r} \frac{\mathrm{d}^2}{\mathrm{d}t^2} e^{\frac{1}{2}t^2}\bigg|_{t=0} = r(1+t^2)e^{\frac{1}{2}t^2}\bigg|_{t=0} = r$$

Theorem 1.29. Given a set of random variables $\{X_1, \cdots, X_k\}$. Let $Y = \sum_{i=1}^{k} X_i$ and $X_i \sim \chi^2(r_i)$ for all $i = 1, \cdots, k$. If they are independent, then $Y \sim \chi^2(r_1 + \cdots + r_k)$.

Proof.
It suffices to prove the relation for two random variables Z_1 and Z_2. If $Z_1 \sim \chi^2(n_1)$ and $Z_2 \sim \chi^2(n_2)$, then $Z_1 + Z_2 \sim \chi^2(n_1 + n_2)$. By repeating applying the relation for two random variables, one can easily have the desired relation for k random variables.
From definition, $Z_1 = X_{11}^2 + \cdots + X_{1n_1}^2$ and $Z_2 = X_{21}^2 + \cdots + X_{2n_2}^2$, where $X_{1i} \sim \mathrm{N}(0,1)$ and $X_{2j} \sim \mathrm{N}(0,1)$ for $i = 1, \cdots, n_1$ and $j = 1, \cdots, n_2$. Therefore,

$$Z_1 + Z_2 = X_{11}^2 + \cdots + X_{1n_1}^2 + X_{21}^2 + \cdots + X_{2n_2}^2 \sim \chi^2(n_1 + n_2)$$

□

Theorem 1.30. If $\{X_1, \cdots, X_n\}$ is a random sample of size $n > 1$ of a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$, then we have:

1. Sample mean $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$
2. Sample mean $\overline{X}$ and sample variance S_{n-1}^2 are independent.
3. $$\frac{(n-1)S_{n-1}^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$$

Proof.

1. From definition,
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
Since $X_i \sim \mathrm{N}(\mu, \sigma^2)$ for $i = 1, \cdots, n$, we can find that $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.

2. Let $\mathbf{X} = (X_1 \ \cdots \ X_n)^T$. We may find that:
$$\begin{pmatrix} \overline{X} \\ X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n \\ \left(1 - \frac{1}{n}\right)X_1 - \frac{1}{n}X_2 - \cdots - \frac{1}{n}X_n \\ -\frac{1}{n}X_1 + \left(1 - \frac{1}{n}\right)X_2 - \cdots - \frac{1}{n}X_n \\ \vdots \\ -\frac{1}{n}X_1 - \frac{1}{n}X_2 - \cdots + \left(1 - \frac{1}{n}\right)X_n \end{pmatrix} = \mathbf{AX} \qquad \mathbf{A} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & 1 - \frac{1}{n} \end{pmatrix}$$
By Lemma 1.24, we have $\mathbf{AX} \sim \mathrm{N}_{n+1}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\sigma^2\mathbf{I}_{n\times n}\mathbf{A}^T)$, where $\boldsymbol{\mu} = (\mu \ \cdots \ \mu)^T$.
Let $\mathbf{X}^* = (X_1 - \overline{X} \ \cdots \ X_n - \overline{X})^T$ and $\boldsymbol{\Sigma}^*$ be the variance-covariance matrix of $\widetilde{\mathbf{X}}$. We may notice that:
$$\mathbf{A}\sigma^2\mathbf{I}_{n\times n}\mathbf{A}^T = \left(\begin{array}{c|c} \mathrm{Var}(\overline{X}) & \mathrm{cov}(\mathbf{X}^*, \overline{X}) \\ \hline \mathrm{cov}(\mathbf{X}^*, \overline{X}) & \boldsymbol{\Sigma}^* \end{array}\right)$$
We prove that $\mathrm{cov}(X_i - \overline{X}, \overline{X}) = 0$ for $i = 1, \cdots, n$. Since X_i are independent for all i,
$$\mathrm{cov}(X_i - \overline{X}, \overline{X}) = \mathrm{cov}(X_i, \overline{X}) - \mathrm{Var}(\overline{X}) = \frac{1}{n}\mathrm{Var}(X_i) - \frac{\sigma^2}{n} = 0$$
Therefore, we can find that $\mathrm{cov}(\mathbf{X}^*, \overline{X}) = 0$. By Lemma 1.25, $\overline{X}$ and $\mathbf{X}^*$ are independent.
Since S_n^2 is a function of $\mathbf{X}^*$, we can conclude that $\overline{X}$ and S_{n-1}^2 are independent.

3. We have:
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n(\overline{X} - \mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right) - \left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right)^2$$
Let $U = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)$ and $V = \left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right)^2$. The distribution we are finding is $U - V$.
From definition, we know that $U \sim \chi^2(n)$. From Theorem 1.28, we find the $V \sim \chi^2(1)$. From Part 2, since functions of $\mathbf{X}^*$ and $\overline{X}$ are independent,
$$\begin{aligned} M_U(t) &= M_V(t)M_{U-V}(t) \\ (1-2t)^{-\frac{n}{2}} &= (1-2t)^{-\frac{1}{2}}M_{U-V}(t) \\ M_{U-V}(t) &= (1-2t)^{-\frac{n-1}{2}} \end{aligned}$$
Therefore, we can conclude that:
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$$

□

Remark 1.30.1. From the same proof of above theorem part 2, we can also find that $\overline{X}$ and S_n^2 are independent.

Example 1.37. (**Student's t-distribution**) $T \sim t(r)$
Assume that $X \sim \mathrm{N}(0,1)$ and $Y \sim \chi^2(r)$. Let:

$$T = \frac{X}{\sqrt{\frac{Y}{r}}}$$

Then T has a t-distribution with t degree of freedom and:

$$f_T(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{t^2}{r}\right)^{-\frac{r+1}{2}} \qquad \mathrm{E}(T) = \begin{cases}\text{Undefined}, & r \le 1\\ 0, & n > 1\end{cases} \qquad \mathrm{Var}(T) = \begin{cases}\text{Undefined}, & r \le 1\\ \infty, & 1 < r \le 2\\ \frac{r}{r-2}, & r > 2\end{cases}$$

Remark 1.30.2. As $r \to \infty$, $T \to \mathrm{N}(0,1)$ by CLT.

Remark 1.30.3. If we fix $Y = y$, then we can find that $T \sim \mathrm{N}(0, \frac{r}{y})$.

The t-distribution has the following properties.

Theorem 1.31. If $\{X_1, \cdots, X_n\}$ is a random sample of size $n > 1$ of random variable $X \sim \mathrm{N}(\mu, \sigma^2)$, then we have:

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S_{n-1}} \sim t(n-1)$$

Proof.
From Theorem 1.30, $\overline{X}$ and S_{n-1}^2 are independent, and:

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim \mathrm{N}(0,1) \qquad\qquad \frac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2(n-1)$$

Therefore, from definition,

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S_{n-1}} = \frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{1}{n-1}\left(\frac{(n-1)S_{n-1}^2}{\sigma^2}\right)}} \sim t(n-1)$$

□

Example 1.38. Assume that we want to find the 95% confidence interval of μ without knowing the population variance σ^2. Then we can find:

$$\begin{aligned} 0.95 &= \mathrm{P}\left(-t_{0.025,n-1} \le \frac{\sqrt{n}(\overline{X}-\mu)}{S_{n-1}} \le t_{0.025,n-1}\right) \\ &= \mathrm{P}\left(\overline{X} - t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}} \le \mu \le \overline{X} + t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}}\right) \end{aligned}$$

Therefore, we can find the 95% confidence interval:

$$\left(\overline{X} - t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}}, \overline{X} + t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}}\right)$$

Usually when $n > 30$, $t_{0.025,n-1} \approx z_{0.025}$. Therefore, we may find the 95% confidence interval:

$$\left(\overline{X} - z_{0.025}\frac{S_{n-1}}{\sqrt{n}}, \overline{X} + z_{0.025}\frac{S_{n-1}}{\sqrt{n}}\right)$$

Example 1.39. (**F distribution**) $F \sim F(r_1, r_2)$
Assume that X and Y are independent random variables with $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Let:

$$F = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}}$$

Then F has a F-distribution with r_1 and r_2 degrees of freedom with:

$$f_F(w) = \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)}\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} w^{\frac{r_1}{2}-1}\left(1+\frac{r_1 w}{r_2}\right)^{-\frac{r_1+r_2}{2}}$$

where $0 < w < \infty$. We define $F_\alpha(r_1, r_2)$ by:

$$\mathrm{P}(F \geq F_\alpha(r_1, r_2)) = \alpha$$

Lemma 1.32. Let $U \sim F(r_1, r_2)$. The F-distribution has the following properties:

1. $\frac{1}{U} \sim F(r_2, r_1)$
2. If $F_\alpha(r_1, r_2)$ is defined by $\mathrm{P}(U \geq F_\alpha(r_1, r_2)) = \alpha$, then:

$$\frac{1}{F_\alpha(r_1, r_2)} = F_{1-\alpha}(r_2, r_1)$$

Proof.

1. By definition,

$$U = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}}$$

where $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Therefore,

$$\frac{1}{U} = \frac{\frac{Y}{r_2}}{\frac{X}{r_1}} \sim F(r_2, r_1)$$

2. With $\mathrm{P}(U \geq F_\alpha(r_1, r_2)) = \alpha$, since $f_U(w)$ is only defined for $w > 0$, we can get that:

$$\mathrm{P}\left(\frac{1}{U} \leq \frac{1}{F_\alpha(r_1, r_2)}\right) = \alpha$$
$$\mathrm{P}\left(\frac{1}{U} > \frac{1}{F_\alpha(r_1, r_2)}\right) = 1 - \alpha$$

From Part 1, we find that $\frac{1}{U} \sim F(r_2, r_1)$. Therefore,

$$\mathrm{P}\left(\frac{1}{U} \geq F_{1-\alpha}(r_2, r_1)\right) = 1 - \alpha$$

We find that $\frac{1}{F_\alpha(r_1, r_2)} = F_{1-\alpha}(r_2, r_1)$

□

Example 1.40. Assume that we try to compare two populations. Let $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ represents the random variable of the first population and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ represents the random variable of the second population. We want to find an interval guess of their ratio of variance $\frac{\sigma_1^2}{\sigma_2^2}$.
Let $\{X_{11}, \cdots, X_{1n}\}$ be a random sample of size n from X_1 and $\{X_{21}, \cdots, X_{2m}\}$ be a random sample of size m from X_2. We can find that:

$$\frac{(n-1)S_{n-1,1}^2}{\sigma_1^2} \sim \chi^2(n-1) \qquad \frac{(m-1)S_{m-1,2}^2}{\sigma_2^2} \sim \chi^2(m-1)$$

We can also find that $S_{n-1}, 1$ and $S_{m-1}, 2$ are independent since they are from different population. Therefore, we have:

$$\frac{\sigma_1^2}{\sigma_2^2}\left(\frac{S_{m-1,2}^2}{S_{n-1,1}^2}\right) = \frac{\frac{1}{m-1}\left(\frac{(m-1)S_{m-1,2}^2}{\sigma_2^2}\right)}{\frac{1}{n-1}\left(\frac{(n-1)S_{n-1,1}^2}{\sigma_1^2}\right)} \sim F(m-1, n-1)$$

Then we can find the 95% confidence interval by:

$$\begin{aligned} 0.95 &= \mathrm{P}\left(f_{0.975,m-1,n-1} \leq \frac{\sigma_1^2}{\sigma_2^2}\left(\frac{S_{m-1,2}^2}{S_{n-1,1}^2}\right) \leq f_{0.025,m-1,n-1}\right) \\ &= \mathrm{P}\left(\frac{S_{n-1,1}^2}{S_{m-1,2}^2} f_{0.975,m-1,n-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_{n-1,1}^2}{S_{m-1,2}^2} f_{0.025,m-1,n-1}\right) \end{aligned}$$

Chapter 2

Point Estimation

In this chapter, we will study two different general approaches to estimate an unknown parameters of any given parametric distribution.
The basic idea of point estimation is that we use a statistic T, an estimate $T(\mathbf{x})$ or an estimator $T(\mathbf{X})$ to estimate the unknown parameter $g(\theta)$, where $\mathbf{x} = (x_1 \ \cdots \ x_n)^T$ is a realization of random vector $\mathbf{X} = (X_1 \ \cdots \ X_n)^T$ with a PDF $f(x|\theta)$ or PMF $p(x|\theta)$ and θ in parameter space Θ.

Remark 2.0.1. Most often, the parameters of our interest to be estimated (**estimand**) is a function of the unknown distribution parameters θ. E.g. $\mu^2, \frac{\sigma}{\mu}$.

Remark 2.0.2. We only estimate an unknown parameters. There is no point to estimate an already known parameters.

2.1 Methods of Moments Estimation

Methods of moments estimation is one of the most popular methods in statistics to estimate an unknown parameters. As the name suggests, it is related to moments. The motivation is that in some situations, the parameter of interest can be written as a function of population moments about 0.

Definition 2.1. Suppose that there are k unknown parameters $\theta_1, \cdots, \theta_k$. If we can write them in terms of k or more moments, i.e.:

$$\begin{cases} \theta_1 = g_1(\mu_1', \mu_2', \cdots, \mu_k', \cdots) \\ \theta_2 = g_2(\mu_1', \mu_2', \cdots, \mu_k', \cdots) \\ \vdots \\ \theta_k = g_k(\mu_1', \mu_2', \cdots, \mu_k', \cdots) \end{cases}$$

then the **method of moments estimator** (MME), denoted by $(\widetilde{\theta}_1, \widetilde{\theta}_2, \cdots, \widetilde{\theta}_k)$, of $(\theta_1, \theta_2, \cdots, \theta_k)$ is:

$$\begin{cases} \widetilde{\theta}_1 = g_1(\overline{X}, \overline{X^2}, \cdots, \overline{X^k}, \cdots) \\ \widetilde{\theta}_2 = g_2(\overline{X}, \overline{X^2}, \cdots, \overline{X^k}, \cdots) \\ \vdots \\ \widetilde{\theta}_k = g_k(\overline{X}, \overline{X^2}, \cdots, \overline{X^k}, \cdots) \end{cases}$$

Remark 2.1.1. This estimation is quick and easy, but the MME obtained are often biased and heavily relies on the existence of the required population moments.

Remark 2.1.2. Do not mix up method of moment estimator or method of moment estimate.

Remark 2.1.3. Do not write the MME as $(\theta_1, \theta_2, \cdots, \theta_k)$. This is wrong.

Example 2.1. Consider a random sample of size n from $X \sim \mathrm{N}(10, \sigma^2)$. We want to estimate σ^2.
We have $k = 1$, $\theta_1 = \sigma^2$. We can write it in terms of moments:

$$\sigma^2 = \mathrm{E}(X^2) - 100$$

Therefore, the MME of σ^2 is:

$$\widetilde{\sigma}^2 = \overline{X^2} - 100$$

Example 2.2. Consider a random sample of size n from $X \sim \mathrm{N}(\mu, \sigma^2)$. We want to estimate μ and σ^2.
We have $k = 2$, $(\theta_1, \theta_2) = (\mu, \sigma^2)$. We can write them in terms of moments:

$$\begin{cases} \mu = E(X) \\ \sigma^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \end{cases}$$

Therefore, the MME of μ and σ^2 are:

$$\begin{cases} \widetilde{\mu} = \overline{X} \\ \widetilde{\sigma}^2 = \overline{X^2} - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{cases}$$

Remark 2.1.4. MME may not be unique because the parameter can be written as different functions of moments. To fix this problem, we usually prefer using fewer or lower moments to get MME.

Example 2.3. Consider a random sample of size n from $X \sim \mathrm{Poisson}(\lambda)$. We want to estimate λ. We have $k = 1$, $\theta_1 = \lambda$. We have multiple ways to write it in terms of moments. It can be $\lambda = \mathrm{E}(X)$, $\lambda = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$ or any other combinations. From the remark, we would choose the one with fewer or lower moments, which is:

$$\lambda = \mathrm{E}(X)$$

Therefore, the MME of λ is:

$$\lambda = \overline{X}$$

Example 2.4. Consider a random sample of size n from $X \sim \mathrm{Gamma}(\alpha, \beta)$. Assume that we know that $\mathrm{E}(X) = 3423$. We have $k = 2$, $(\theta_1, \theta_2) = (\alpha, \beta)$. We can write them in terms of moments:

$$\begin{cases} 3423 = \frac{\alpha}{\beta} \\ \mathrm{E}(X^2) = \frac{\alpha}{\beta^2} + 3423^2 \end{cases} \implies \begin{cases} \alpha = \frac{3423^2}{\mathrm{E}(X^2) - 3423^2} \\ \beta = \frac{3423}{\mathrm{E}(X^2) - 3423^2} \end{cases}$$

Therefore, the MME of α and β is:

$$\begin{cases} \widetilde{\alpha} = \frac{3423^2}{\overline{X^2} - 3423^2} \\ \widetilde{\beta} = \frac{3423}{\overline{X^2} - 3423^2} \end{cases}$$

Lemma 2.2. (**Invariance property of MME**) If $\widetilde{\theta}_i$ is the MME for θ_i for $i = 1, \cdots, k$, then $h(\widetilde{\theta}_1, \cdots, \widetilde{\theta}_k)$ is the MME for $h(\theta_1, \cdots, \theta_k)$, where h is a known function.

Theorem 2.3. A sequence of MME $\{\widetilde{\theta}_n \in \mathbb{R}^k\}$ is consistent and asymptotically unbiased for θ. It is also asymptotically normally distributed. More precisely, under certain assumption like $\mathrm{E}\,|X|^{2k} < \infty$, as $n \to \infty$, we have:

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \to \mathrm{N}_k(\mathbf{0}, \mathbf{G}\mathbf{H}\mathbf{G}^T)$$

where $\mathbf{G}$ is a $k \times k$ matrix with $\frac{\partial g_i}{\partial \mu_j'}$ as its (i, j)-th entry and $\mathbf{H}$ is a $k \times k$ matrix with $\mu_{i+j}' - \mu_i'\mu_j'$ as its (i, j)-th entry, for $i = 1, \cdots, k$ and $j = 1, \cdots, k$.

Remark 2.3.1. In the theorem, "consistent" means convergence in probability. For any $\varepsilon > 0$, as $n \to \infty$,

$$\mathrm{P}(|\widetilde{\theta}_n - \theta| > \varepsilon) \to 0$$

Remark 2.3.2. Also in the theorem, "asymptotically unbiased" means that we have:

$$\lim_{n\to\infty} \mathrm{E}(\widetilde{\theta}_n) = \theta$$

Note that $\mathrm{E}(\widetilde{\theta}_n) \neq \theta$ for some n.

Example 2.5. Consider a random sample of size n from a random variable X with $\mathrm{E}\,|X|^4 < \infty$. We take:

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mu_1' \\ \mu_2' - (\mu_1')^2 \end{pmatrix}$$

We have:

$$\mathbf{G} = \begin{pmatrix} 1 & 0 \\ -2\mu_1' & 1 \end{pmatrix} \qquad \mathbf{H} = \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - \mu_2'\mu_1' & \mu_4' - (\mu_2')^2 \end{pmatrix}$$

Therefore,

$$\begin{aligned} \mathbf{GHG}^T &= \begin{pmatrix} 1 & 0 \\ -2\mu_1' & 1 \end{pmatrix} \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - \mu_2'\mu_1' & \mu_4' - (\mu_2')^2 \end{pmatrix} \mathbf{G}^T \\ &= \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 & \mu_4' - 2\mu_1'\mu_3' - (\mu_2')^2 + 2(\mu_1')^2\mu_2' \end{pmatrix} \begin{pmatrix} 1 & -2\mu_1' \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 \\ \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 & \mu_4' - 4\mu_1'\mu_3' - (\mu_2')^2 + 8\mu_2'(\mu_1')^2 - 4(\mu_1')^4 \end{pmatrix} \end{aligned}$$

Using the fact that:

$$\begin{aligned} \mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 \\ \mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \\ \sigma^4 &= (\mu_2')^2 - 2\mu_2'(\mu_1')^2 + (\mu_1')^4 \end{aligned}$$

We can find the resultant matrix:

$$\mathbf{GHG}^T = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}$$

Using Theorem 2.3, denote:

$$\widetilde{\theta}_n = \begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix}$$

As $n \to \infty$,

$$\sqrt{n}\left(\begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right) \to \mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix} \right)$$

Based on the properties of variance-covariance matrix, we may find that as $n \to \infty$:

$$\sqrt{n}(S_n^2 - \sigma^2) \to \mathrm{N}(0, \mu_4 - \sigma^4)$$

By Delta Method, in condition that $\sigma^2 > 0$,

$$\sqrt{n}(S_n - \sigma) \to \mathrm{N}\left(0, \frac{\mu_4 - \sigma^4}{4\sigma^2}\right)$$

2.2 Maximum Likelihood Estimation

The method of maximum likelihood is by far the most popular technique for deriving estimators, popularized by Ronald Aylmer Fisher in 1922. Currently, there are still a lot of research studying the properties of this estimation method.

Definition 2.4. Consider a random sample of size n from a population with a PDF $f(\mathbf{x}|\theta)$ or a PMF $p(\mathbf{x}|\theta)$. Given a realization $\mathbf{x} = (x_1 \ \cdots \ x_n)^T$. The **likelihood function** is defined by:

$$L(\theta) = L(\theta_1, \cdots, \theta_k|\mathbf{x}) = \begin{cases} \prod_{i=1}^n f(x_i|\theta), & \text{Continuous case} \\ \prod_{i=1}^n p(x_i|\theta), & \text{Discrete case} \end{cases}$$

The likelihood function can be used to quantify how the observed data is likely to occur.

Remark 2.4.1. The likelihood function $L(\theta)$ is a function of θ with fixed $\mathbf{x}$.

Remark 2.4.2. Do not replace x_i with x.

$$L(\theta) = \begin{cases} \prod_{i=1}^n f(x_i|\theta) \neq \prod_{i=1}^n f(x|\theta), & \text{Continuous case} \\ \prod_{i=1}^n p(x_i|\theta) \neq \prod_{i=1}^n p(x|\theta), & \text{Discrete case} \end{cases}$$

The idea is that for each realization of $\mathbf{x}$, we want to estimate a value of $\theta \in \Theta$ at which $L(\theta)$ attains its maximum.

Definition 2.5. The **maximum likelihood estimate** (MLE), denoted by $\hat{\theta}$, is obtained by:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$$

Remark 2.5.1. In some cases, especially when differentiation is used, it is easier to work with **log likelihood** defined by:

$$l(\theta) = \log L(\theta)$$

We can do this because $l(\theta)$ and $L(\theta)$ are strictly increasing and they have the same maxima.

Example 2.6. Consider a random sample of size $n = 10$ from Bern(θ), where θ is unknown. Therefore,

$$L(\theta) = \prod_{i=1}^n p(x_i|\theta) = \theta^{n\overline{x}}(1-\theta)^{n-n\overline{x}}$$

Suppose that there are only two possible values of θ. We have either $\theta = 0.1$ or $\theta = 0.5$.
From the data observed, assumed that we have found that $\overline{x} = 0.4$. Substituting gets:

$$L(0.1) = (0.1)^4(0.9)^6 = 0.0000531441 \qquad L(0.5) = (0.5)^4(0.5)^6 = 0.0009765625$$

Therefore, the MLE of θ is $\hat{\theta} = 0.5$.

Example 2.7. In the case when $L(\theta)$ is differentiable on the interior of θ. One possible way of finding an MLE of $\theta = (\theta_1 \ \cdots \ \theta_k)^T$ is to solve the first order equation for $i = 1, \cdots, k$:

$$\frac{\partial}{\partial \theta_i} L(\theta) = 0 \text{ or } \frac{\partial}{\partial \theta_i} l(\theta) = 0$$

and check all the extrema.

Remark 2.5.2. Solving the first-order likelihood only gives you the maximum at critical points. You need to also check the extreme values too.

Example 2.8. Consider a random sample of size n from $\mathrm{N}(\theta, 1)$, where θ is unknown. We may obtain the log likelihood:

$$l(\theta) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\theta)^2}\right) = -\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{n}{2}\ln(2\pi)$$

We find the critical points by solving:

$$0 = \frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n}(x_i - \theta)$$

This has the solution $\hat{\theta} = \overline{x}$. To check that the solution is in fact a global maximum, we can check that:

$$\frac{\partial^2}{\partial\theta^2} l(\theta) = -n < 0$$

Therefore, the MLE of θ is $\hat{\theta} = \overline{x}$.

Example 2.9. Continues the previous example. Alternatively, we may find that for any $\theta \in \Theta$,

$$\sum_{i=1}^{n}(x_i-\theta)^2 \geq \sum_{i=1}^{n}(x_i-\overline{x})^2$$

Thus, for any $\theta \in \Theta$,

$$L(\theta) \leq L(\overline{x})$$

Therefore, the MLE of θ is $\hat{\theta} = \overline{x}$.

Example 2.10. Consider a random sample of size n from $\mathrm{N}(\theta, 1)$, where θ is unknown. Previously, we have found that $\hat{\theta} = \overline{x}$, which maximize the log likelihood. Let us restrict that $\theta \geq 0$.
If $\overline{x} \geq 0$, then it satisfies the constraint $\theta \geq 0$. Therefore, the MLE would be:

$$\hat{\theta} = \overline{x} \tag{2.1}$$

If $\overline{x} < 0$, then it does not satisfy the constraint $\mu \geq 0$. We analyse the log likelihood again:

$$l(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{n}{2}\ln(2\pi) = -\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2 - \frac{n}{2}(\overline{x}-\theta) - \frac{n}{2}\ln(2\pi)$$

The term $(\overline{x}-\theta)^2$ is minimized while satisfying the constraint is when $\theta = 0$.
Therefore, if we restrict $\theta \geq 0$, the MLE of θ would be:

$$\hat{\theta} = \max(\overline{x}, 0)$$

Remark 2.5.3. Remember, when we estimate a parameter, we must use the data we have obtained.

Example 2.11. Consider a random sample of size n from $\mathrm{U}[0, \theta]$, where $\theta \in (0, \infty)$ is unknown. The likelihood function is:

$$L(\theta) = \frac{1}{\theta^n}\mathbf{1}_{0\leq x_{(1)}\leq\cdots\leq x_{(n)}\leq\theta}$$

where $x_{(i)}$ represents the i-th smallest data for $i = 1, \cdots, n$. Therefore, the MLE is:

$$\hat{\theta} = x_{(n)}$$

Remark 2.5.4. MLE may be biased and it may not exist in Θ, especially when Θ is an open set.

Remark 2.5.5. MLE defined may not be unique.

Example 2.12. Consider a random sample of size n from $\mathrm{U}[\theta-1,\theta+1]$, where θ is unknown. The likelihood function is:

$$L(\theta)=\frac{1}{2^n}\mathbf{1}_{\theta-1\le x_{(1)}\le\cdots\le x_{(n)}\le\theta+1}=\frac{1}{2^n}\mathbf{1}_{x_{(n)}-1\le\theta\le x_{(1)}+1}$$

where $x_{(i)}$ represents the i-th smallest data for $i=1,\cdots,n$. We may fins that any estimates in $[x_{(n)}-1,x_{(1)}+1]$ maximize $L(\theta)$. Therefore, there are infinitely many MLEs of θ.

Lemma 2.6. (**Invariance property of MLE**) If $\hat{\theta}_i$ is the MLE of θ_i for $i=1,\cdots,k$, then $h(\hat{\theta}_1,\cdots,\hat{\theta}_k)$ is the MLE for $h(\theta_1,\cdots,\theta_k)$, where h is a known function.

Theorem 2.7. A sequence of MLE $\{\hat{\theta}_n\in\mathbb{R}^k\}$ is consistent, asymptotically unbiased for θ, asymptotically efficient and asymptotically normally distributed. More precisely, under regularity assumption, as $n\to\infty$, we have:

$$\sqrt{n}(\hat{\theta}_n-\theta)\to\mathrm{N}_k(\mathbf{0},\mathcal{I}_X^{-1}(\theta))$$

where $\mathcal{I}_X(\theta)$ is known as the **Fisher Information matrix** and is a $k\times k$ matrix with the (i,j)-th entry defined as:

$$\begin{cases}\mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\ln f_X(X|\theta)\right)\left(\frac{\partial}{\partial\theta_j}\ln f_X(X|\theta)\right)\right], & \text{Continuous case}\\ \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\ln p_X(X|\theta)\right)\left(\frac{\partial}{\partial\theta_j}\ln p_X(X|\theta)\right)\right], & \text{Discrete case}\end{cases}$$

for $i=1,\cdots,k$ and $j=1,\cdots,k$.

Remark 2.7.1. In the theorem, "asymptotically efficient" means that the limiting variance is the smallest possible. *If you want to know why reciprocal of Fisher Information is the lowest possible variance of any unbiased estimator, search Cramér-Rao bound.*

Notice that we have used a special matrix called "Fisher Information Matrix". What is Fisher Information?

Definition 2.8. Given a set of random variables $\{X_1,\cdots,X_n\}$. The **Fisher Information**, or **Fisher Information matrix** if more than one unknown parameter is considered, of the set is defined by:

$$\mathcal{I}_{X_1,\cdots,X_n}(\theta)=\begin{cases}\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta)\right]^2, & \text{Continuous case}\\ \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln p_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta)\right]^2, & \text{Discrete case}\end{cases}$$

Remark 2.8.1. Fisher Information is a measure of amount of information about an unknown parameter θ that a random variable or data carries. It is very important because we do want to know how to quantify this amount appropriately.

Example 2.13. If $X\sim\mathrm{N}(\mu,\sigma^2)$, where σ^2 is known but $\mu\in(-\infty,\infty)$ is unknown, then the Fisher Information about μ contained in X is:

$$\mathcal{I}_X(\mu)=\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\mu}\ln f_X(X|\mu)\right]^2=\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\mu}\left(-\frac{1}{2\sigma^2}(X-\mu)^2-\frac{1}{2}\ln(2\pi\sigma^2)\right)\right]^2=\mathrm{E}\left[\frac{1}{\sigma^2}(X-\mu)\right]^2=\frac{1}{\sigma^2}$$

Example 2.14. If $X\sim\mathrm{Bern}(p)$, where $p\in(0,1)$ is unknown, then the Fisher Information about p contained in X is:

$$\begin{aligned}\mathcal{I}_X(p)=\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}p}\ln f_X(X|p)\right]^2&=\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}p}\left(X\ln p+(1-X)\ln(1-p)\right)\right]^2\\&=\mathrm{E}\left[\frac{X}{p}-\frac{1-X}{1-p}\right]^2\\&=\mathrm{E}\left[\frac{X-p}{p(1-p)}\right]^2\\&=\frac{p(1-p)}{p^2(1-p)^2}=\frac{1}{p(1-p)}\end{aligned}$$

We will see some properties of the Fisher Information. *For simplicity, we will only discuss continuous random variables.* Notice that we used something called "regularity assumption"? The following are the regularity conditions that we need.

1. $\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)$ exists for all $x_1,\cdots,x_n$ and all $\theta \in \Theta$
2. For any statistic $T(x_1,\cdots,x_n)$,
$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}\int\cdots\int T(x_1,\cdots,x_n) f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)\,dx_1\,\cdots\,dx_n\\ =&\int\cdots\int T(x_1,\cdots,x_n)\frac{\mathrm{d}}{\mathrm{d}\theta} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)\,dx_1\,\cdots\,dx_n\end{aligned}$$
3. $0 < \mathcal{I}_{X_1,\cdots,X_n}(\theta) < \infty$ for all $\theta \in \Theta$

Condition 2 can be satisfied when the support of X does not depend on θ, where the support of X is defined by $\{x : f_X(x|\theta) > 0\}$. Note that $U[0,\theta]$ violates this condition.

Lemma 2.9. Suppose the X is a random variable with PDF f_X. Under the regularity condition, we have:
$$\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right] = 0$$

Proof.
$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{\infty} f_X(x|\theta)\,dx \int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta} f_X(x|\theta)\,dx = \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(x|\theta)\right) f_X(x|\theta)\,dx = \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]$$
□

Remark 2.9.1. Using this lemma, we can find that:
$$\mathcal{I}_X(\theta) = \mathrm{Var}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right)$$

Lemma 2.10. Suppose that $\{X_1,\cdots,X_n\}$ is a set of random variables. Under the regularity conditions and the assumption that $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)$ exists for all $x_1,\cdots,x_n$ and all $\theta \in \Theta$, we have:
$$\mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2 = -\,\mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X|\theta)\right]$$

Proof.
From the proof of last Lemma,
$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{\infty}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(x|\theta)\right) f_X(x|\theta)\,dx\\ &= \int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(x|\theta)\right) f_X(x|\theta)\right]\,dx\\ &= \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(x|\theta)\right) f_X(x|\theta)\,dx + \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(x|\theta)\right)\frac{\mathrm{d}}{\mathrm{d}\theta} f_X(x|\theta)\,dx\\ &= \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(x|\theta)\right) f_X(x|\theta)\,dx + \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(x|\theta)\right)^2 f_X(x|\theta)\,dx\\ &= \mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X|\theta)\right] + \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2\\ -\,\mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X|\theta)\right] &= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2\end{aligned}$$
□

Assume that we consider two independent random variables X and Y. We can find the Fisher Information about θ contained in (X, Y) by finding the Fisher Information about θ contained in each of them.

Lemma 2.11. If X and Y are independent and their PDFs satisfy the regularity conditions, then:

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)$$

Proof.
Since X and Y are independent,

$$\begin{aligned}
\mathcal{I}_{X,Y}(\theta) &= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{X,Y}(X,Y|\theta)\right]^2 \\
&= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta) + \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_Y(Y|\theta)\right]^2 \\
&= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2 + 2\,\mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right)\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_Y(Y|\theta)\right)\right] + \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_Y(Y|\theta)\right]^2 \\
&= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2 + \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_Y(Y|\theta)\right]^2 \qquad \text{(Lemma 2.9)} \\
&= \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)
\end{aligned}$$

□

By applying the same result to a random sample of size n, we can obtain the following property.

Lemma 2.12. Suppose the $\{X_1, \cdots, X_n\}$ is a random sample of size n from a distribution. Then,

$$\mathcal{I}_{X_1,\cdots,X_n}(\theta) = \sum_{i=1}^{n} \mathcal{I}_{X_i}(\theta) = n\mathcal{I}_{X_1}(\theta)$$

Remark 2.12.1. For any $i \neq j$, $\mathcal{I}_{X_i}(\theta) = \mathcal{I}_{X_j}(\theta)$ only means that X_i and X_j carries the same amount of the information about θ. It does not mean they carry identical information.

Example 2.15. Consider a set of i.i.d. random variables $\{X_1, \cdots, X_n\}$ where for all $i = 1, \cdots, n$, $X_i \sim$ Cauchy(θ) and has a PDF:

$$f_{X_i}(x|\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}$$

We may find that:

$$\begin{aligned}
\mathcal{I}_{X_i}(\theta) = \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{X_i}(X_i|\theta)\right]^2 &= \mathrm{E}\left(\frac{\frac{2(X_i-\theta)}{\pi(1+(X_i-\theta)^2)^2}}{\frac{1}{\pi(1+(X_i-\theta)^2)}}\right)^2 \\
&= \mathrm{E}\left(\frac{2(X_i-\theta)}{1+(X_i-\theta)^2}\right)^2 \\
&= \int_{-\infty}^{\infty}\left(\frac{2(x-\theta)}{1+(x-\theta)^2}\right)^2 \frac{1}{\pi(1+(x-\theta)^2)}\,dx \\
&= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{u^2}{(1+u^2)^3}\,du && (u = x-\theta,\ du = dx) \\
&= \frac{8}{\pi}\int_{0}^{\infty}\frac{u^2}{(1+u^2)^3}\,du \\
&= \frac{4}{\pi}\int_{0}^{1}\sqrt{y}\sqrt{1-y}\,dy && (y = \tfrac{1}{1+u^2},\ dy = -\tfrac{2u}{(1+u^2)^2}\,du) \\
&= \frac{4}{\pi}\int_{0}^{1}(y)^{\frac{3}{2}-1}(1-y)^{\frac{3}{2}-1}\,dy && \text{(Beta integral)} \\
&= \frac{4\Gamma(\frac{3}{2})\Gamma(\frac{3}{2})}{\pi\Gamma(3)} = \frac{4(0.5\sqrt{\pi})^2}{2!} = \frac{1}{2} && (\tfrac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)} = \textstyle\int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,dt)
\end{aligned}$$

Therefore, $I_{X_1,\cdots,X_n}(\theta) = nI_{X_1}(\theta) = \frac{n}{2}$.

Note that a statistic or an estimator can be considered as a function for data condensation because we condense a random sample into a lower-dimensional quantity.

Lemma 2.13. Suppose that $\mathbf{X}$ is a random vector. Under the regularity conditions, for any statistic $T(\mathbf{X})$ for θ, we have:

$$\mathcal{I}_{T(\mathbf{X})}(\theta) \leq \mathcal{I}_{\mathbf{X}}(\theta)$$

Remark 2.13.1. The Fisher Information of $T(\mathbf{X})$ is defined by:

$$\mathcal{I}_{T(\mathbf{X})}(\theta) = \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{T(\mathbf{X})}(T(\mathbf{X})|\theta)\right]^2$$

We may prove Theorem 2.7 in one-parameter case:

Theorem 2.14. Consider a random sample $\{X_1, \cdots, X_n\}$ of size n from a parametric distribution with a PDF f_X. Then under the regularity and some other conditions, for $\theta \in \mathbb{R}$, a sequence of MLE $\{\hat{\theta}_n \in \mathbb{R}\}$ allows,

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathrm{N}\left(0, \frac{1}{\mathcal{I}_X(\theta)}\right)$$

Proof.
Since the MLE $\hat{\theta}_n$ is the solution to $l'(\theta) = 0$, we can apply a Taylor expansion of $l'(\hat{\theta}_n)$ at θ to find that:

$$\begin{aligned} 0 &= l'(\theta) + l''(\theta)(\hat{\theta}_n - \theta) + o((\hat{\theta}_n - \theta)) \\ \sqrt{n}(\hat{\theta}_n - \theta) &= \frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)} - o((\hat{\theta}_n - \theta)) \end{aligned}$$

We first consider the numerator. Note that $\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta), \cdots, \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta)$ are i.i.d.. By CLT, we have:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_i|\theta) - \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta)\right]\right) \to \mathrm{N}\left(0, \mathrm{Var}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta)\right]\right)$$

By Lemma 2.9, we have:

$$\frac{1}{\sqrt{n}}l'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_i|\theta) \to \mathrm{N}(0, \mathcal{I}_X(\theta))$$

Now we consider the denominator. By WLLN and Lemma 2.10, since $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X_1|\theta), \cdots, \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X_1|\theta)$ are i.i.d.,

$$-\frac{1}{n}l''(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X_i|\theta) \to -\mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X|\theta)\right] = \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2 = \mathcal{I}_X(\theta)$$

Consequently, we have:

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)} - o((\hat{\theta}_n - \theta)) \to \mathrm{N}\left(0, \frac{1}{\mathcal{I}_X(\theta)}\right)$$

□

Remark 2.14.1. Sometime, $\mathcal{I}_X(\theta)$ cannot be determined easily. We will replace it by an observed Fisher Information defined by $-\frac{1}{n}l''(\hat{\theta}_n)$. Since $\hat{\theta}_n$ is consistent for θ, $-\frac{1}{n}l''(\hat{\theta}_n)$ is also consistent for $\mathcal{I}_X(\theta)$ by continuous mapping theorem. Therefore,

$$\sqrt{-l''(\hat{\theta}_n)}(\hat{\theta}_n - \theta) = \sqrt{n}\sqrt{-\frac{1}{n}l''(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \to \mathrm{N}(0, 1)$$

Example 2.16. (**Principal of Numerical solution to likelihood equations**) Consider a random sample of size n from $X \sim \text{Cauchy}(\theta)$, similar to the previous example. We want to find the MLE of θ. We have:

$$l(\theta) = -n \ln \pi - \sum_{i=1}^{n} \ln(1 + (x_i - \theta)^2)$$

We want to find the solution of $l'(\theta) = 0$, which is the MLE. Setting:

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}$$

However, this cannot be solved explicitly in this case.

Example 2.17. (**Newton-Raphson Algorithm**) By Taylor expansion, we can get:

$$0 = \frac{1}{n} l'(\hat{\theta}) \approx \frac{1}{n} l'(\theta) + \frac{1}{n}(\hat{\theta} - \theta) l''(\theta)$$

We may modify it to find that:

$$\hat{\theta} \approx \theta - \frac{l'(\theta)}{l''(\theta)}$$

We may initially guess a number, say θ_0. By iteratively applying the procedure for $j = 0, 1, \cdots$:

$$\theta_{j+1} = \theta_j - \frac{l'(\theta_j)}{l''(\theta_j)}$$

and stop it at a certain stopping criterion, say $|\theta_{j+1} - \theta_j| < K$ for some chosen constant K. E.g. $K = 10^{-5}$. Using this algorithm, we can obtain or approximate the MLE of θ.

Example 2.18. Consider a random sample of size n from $X \sim \text{Gamma}(\alpha, \beta)$, where $\beta = 3423$ and α is unknown. The PDF is defined by:

$$f_X(x|\theta) = \begin{cases} \frac{3423^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-3423x}, & x > 0 \\ 0, & \text{Otherwise} \end{cases}$$

We can find the log likelihood:

$$l(\alpha) = \sum_{i=1}^{n} \ln f_X(x_i|\alpha) = n\alpha \ln 3423 - n \ln(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - 3423 \sum_{i=1}^{n} x_i$$

To find MLE, we try to solve the equation:

$$0 = \frac{\mathrm{d}}{\mathrm{d}\alpha} l(\alpha) = n \ln 3423 - n \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln(\Gamma(\alpha)) + \sum_{i=1}^{n} \ln x_i$$

However, we have no idea how to find $\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln(\Gamma(\alpha))$. Therefore, instead, we would use the numerical methods to approximate the MLE.