Probability Theory

SUZUKI, Atsushi

The whole contents

Outline

Outline

Sample and sample statistics

In real applications, we **rarely know the true distribution**, behind the data.

On the other hand, we often **have many data points** that we can assume follow the same distribution (often independently). Such a series of data points is called ***sample*** of the distribution.

Statistics, data science, machine learning, etc., aim to **extract information about the true distribution from available data points. Sample statistics are the basis of those pieces of technology**.

Learning outcomes

By the end of this section, you should be able to:

- Explain the difference between summary statistics and sample statistics,
- Estimate the true mean of an unknown distribution by finite size sample,
- Explain why many random variables in the real world follow a normal distribution, and
- Estimate an unknown distribution using a parametric model and maximum likelihood estimator.

Outline

Population and sample

In the context of statistics,

- The true distribution is often called the ***population***.
- A series of data points that we can assume follow the same distribution is called ***sample***. If it has many data points, we say that the sample is large, and if it has few data, we say that the sample is small.

Summary statistics and sample statistics

- ***Summary statistics*** aims to describe characteristics of a (known or true) distribution by a few values.
- ***Sample statistics*** aims to estimate some information about the true distribution from finite sample data.

We only have **finite** data points in real applications, so sample statistics are practically necessary to handle probability.

Outline

1 **Sample Statistics**

Sample mean, law of large numbers, and central limit theorem

Sample mean

One principal summary statistic is the expectation.

For data points $X_1, X_2, \ldots, X_m$, we can easily calculate the ***sample mean***

$$\overline{X}_m = \frac{1}{m}(X_1 + X_2 + \cdots + X_m), \tag{1}$$

the mean of the data points.

If we can assume that those data points are the values of random variables following the same distribution with a true mean μ, we expect $\overline{X}_m$ to approximate the true mean μ, which is unknown.

Is it correct? The answer is YES, according to the **law of large numbers**.

Law of large numbers

Theorem ((Strong) law of large numbers)

Let $X_1, X_2, \dots$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables and assume that the mean of the distribution is $\mu \in \mathbb{R}$.
Let $\overline{X}_m$ be the sample mean

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{2}$$

Then $\overline{X}_m$ converges to μ in probability 1.

Thus, the sample mean tells us some information about the unknown true distribution!

How the sample mean behaves?

The sample mean converges to the expectation. Now,

- How close to the expectation will the sample mean get as we increase the data points?
- What does the distribution of the sample mean look like?

The answer is

- The difference between the sample mean and the true expectation is proportional to the standard deviation σ of the true distribution and $\frac{1}{\sqrt{m}}$,
- With appropriate scaling, the distribution of the sample mean converges to a ***normal distribution (Gaussian distribution)***,

according to the ***central limit theorem***.

What is the standard normal distribution?

The ***standard normal distribution***, also known as the ***standard Gaussian distribution*** is the distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{3}$$

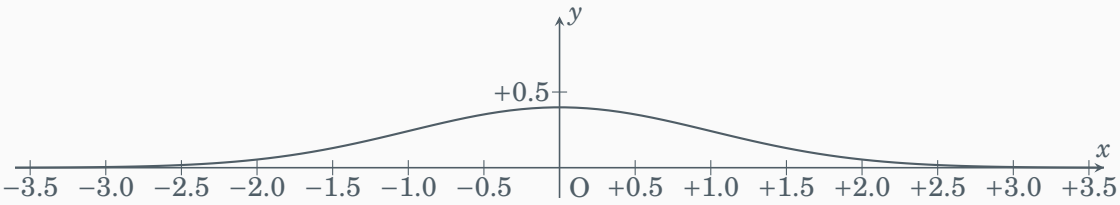

Figure: The standard normal distribution's PDF.

Mean: 0, Variance: 1. The PDF is symmetric about $x = 0$ and it is dense around $x = 0$.

Central limit theorem (CLT)

Theorem (Central limit theorem (CLT))

Let $X_1, X_2, \dots$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables and assume that the mean and variance of the distribution are $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{\geq 0}$, respectively.
Let $\overline{X}_m$ be the sample mean

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{4}$$

Then, the CDF of $\sqrt{m}\frac{\overline{X}_m - \mu}{\sigma}$ converges to the CDF of the standard normal distribution at any point in $\mathbb{R}$.

The standard normal distribution's CDF

By definition, the CDF $F : \mathbb{R} \to [0, 1]$ of the standard normal distribution is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x'^2}{2}\right) \mathrm{d}x' . \tag{5}$$

It is known that this function is not elementary.

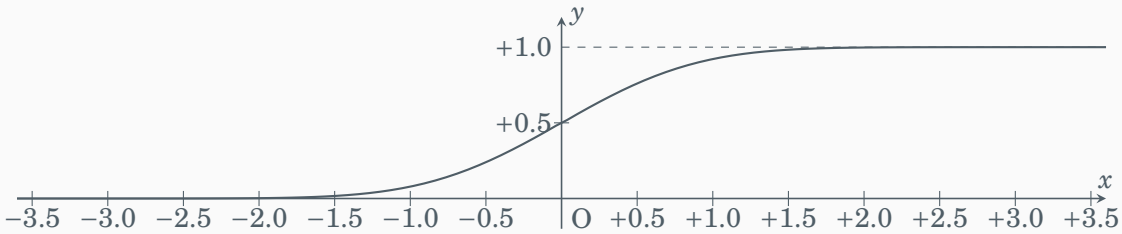

Figure: The standard normal distribution's CDF.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

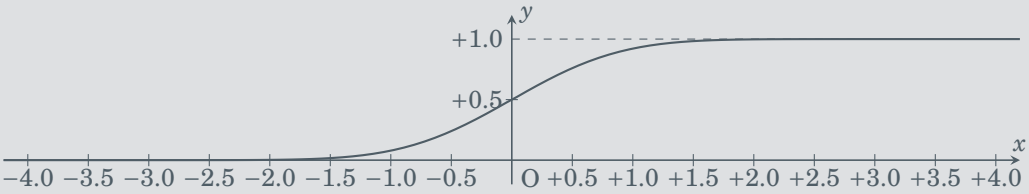

Figure: Dashed: the standard normal distribution's CDF.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

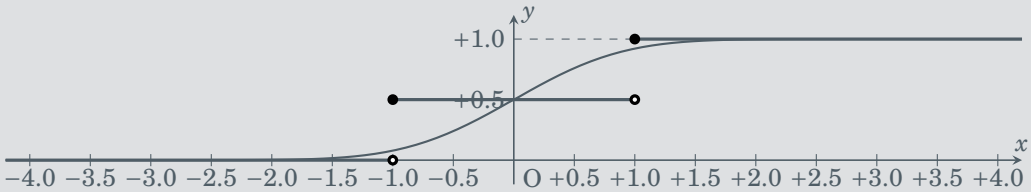

Figure: Dashed: the standard normal distribution's CDF. Solid: the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$, where $m = 1$.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

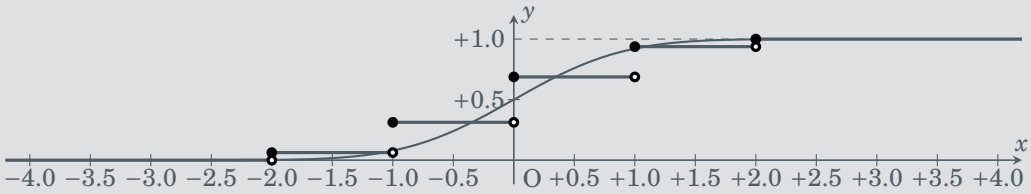

Figure: Dashed: the standard normal distribution's CDF. Solid: the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$, where $m = 4$.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.

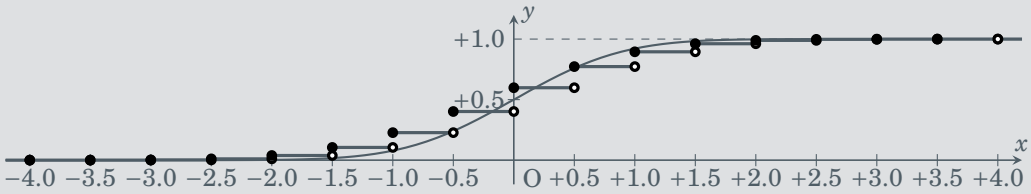

Figure: Dashed: the standard normal distribution's CDF. Solid: the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$, where $m = 16$.

Example of the convergence by the CLT

Example

Let $X_1, X_2, \ldots$ be an infinite sequence of independently identically distributed RVs, where X_i takes 0 or $+1$ with probability $\frac{1}{2}$ for each.
Then the mean and the variance of X_i are $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
According to the CLT, the CDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution $\mathcal{N}(0,1)$.
Note that the CLT is about the CDF, but **NOT about the PDF**. The convergence of the PDF does not always hold. Specifically, in the above case, $\overline{X_m}$ is a discrete random variable since each X_i is. Hence, the random variable $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ does not have a PDF. Therefore, we **CANNOT** say that the PDF of $2\sqrt{m}\left(\overline{X_m} - \frac{1}{2}\right)$ converges to that of the standard normal distribution.

The implications of the CLT

- The error $\overline{X}_m - \mu$ in estimating the true mean μ is almost proportional to $\frac{1}{\sqrt{m}}$. In particular, the more data points, the more accurate the estimate is.
- The sum of sufficiently many independent random variables approximately follows a normal distribution. In particular, various types of random variables decomposable to many independent factors follow a normal distribution. This is why **the normal distribution appears everywhere in the real world**.

Outline

1 **Sample Statistics**

Estimation of distribution and parametric model

Estimation of a distribution

We have estimated the expectation only. In real applications, we might want to estimate the distribution itself. However, if the support of the distribution is an infinite set[1], it is not practical to determine a PMF or PDF from finite data points with no assumptions.

We often assume that the distribution is in a parametric model, which is a set of distributions parametrized by a few values.

[1] This is almost always the case if we consider a continuous RV

Parametric model

Definition (A parametric model)

- ***A discrete parametric model*** on support $\mathscr{X} \subset \mathbb{R}^n$ is a pair of a parameter set $\Theta \subset \mathbb{R}^k$ and a parametrized PMF $P : \mathscr{X} \times \Theta \to [0,1]$ such that $P(\boldsymbol{x};\boldsymbol{\theta})$ is a PMF on $\mathscr{X}$ as a function of $\boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$.
- ***A continuous parametric model*** on support $\mathbb{R}^n$ is a pair of a parameter set $\Theta \subset \mathbb{R}^k$ and a parametrized PDF $p : \mathbb{R}^n \times \Theta \to \mathbb{R}_{\geq 0}$ such that $p(\boldsymbol{x};\boldsymbol{\theta})$ is a PDF on $\mathbb{R}^n$ as a function of $\boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$.

Here, the nonnegative integer k is the dimension of the parameter.

When we have a parametric model, estimating a parameter corresponds to estimating a distribution.

Parametric model example 1: Bernoulli distribution

Example (Bernoulli distribution)

The Bernoulli distribution[2] is a discrete parametric model with a sole parameter, which is usually denoted by θ. The support and the parameter set are $\mathscr{X} = \{0, 1\}$ and $\Theta = [0, 1]$, respectively. The parametrized PMF $P(x; \theta)$ is given by $P(1; \theta) = \theta$. Thus, we have $P(0; \theta) = 1 - \theta$.

Theorem

The mean and the variance of a RV following the Bernoulli distribution with the parameter θ are θ and $\theta(1 - \theta)$, respectively.

[2] A parametric model is often called like the XXX distribution, but it is, indeed, a parametrized **set** of distributions.

Parametric model example 2: normal distribution

Example (Normal distribution)

The normal distribution, also known as the ***Gaussian distribution***, is a continuous parametric model, which has mean parameter $\mu \in \mathbb{R}$ and variance parameter $\sigma^2 \in \mathbb{R}_{>0}$. That is, the parameter set is $\Theta = \mathbb{R} \times \mathbb{R}_{>0}$. The parametrized PDF $p(x; \mu, \sigma^2)$ is given by $p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Theorem

The mean and the variance of a RV following the normal distribution with mean parameter μ and variance parameter σ^2 are μ and σ^2, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

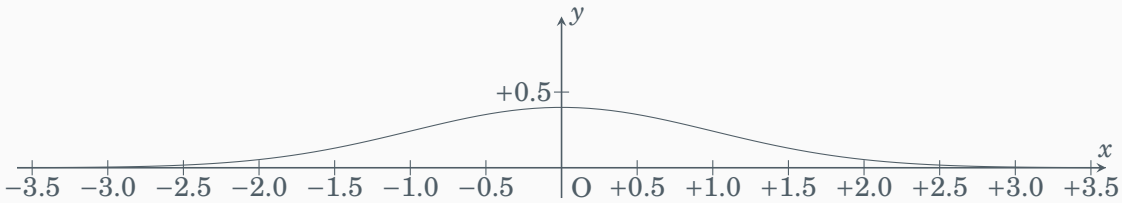

Figure: Normal distributions' PDF ($\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

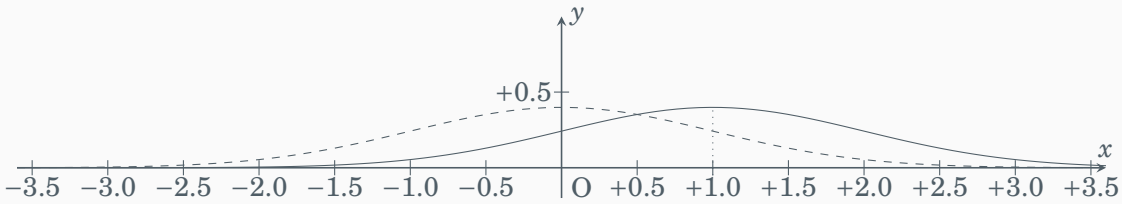

Figure: Normal distributions' PDF (Solid: $\mu = 1, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

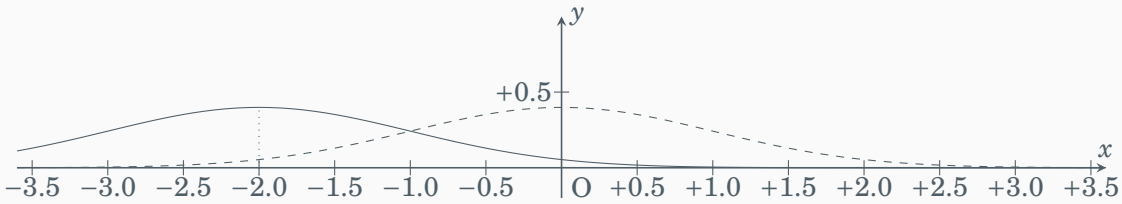

Figure: Normal distributions' PDF (Solid: $\mu = -2, \sigma = 1$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

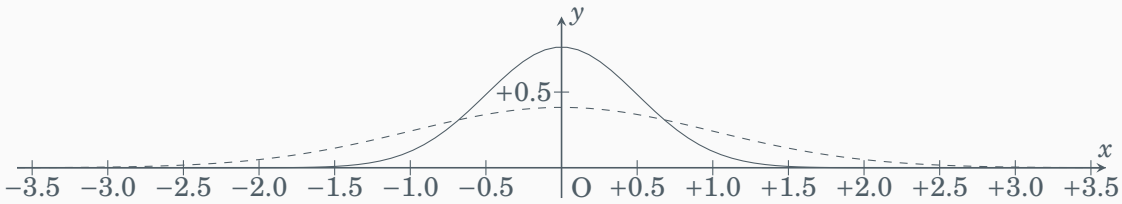

Figure: Normal distributions' PDF (Solid: $\mu = 0, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

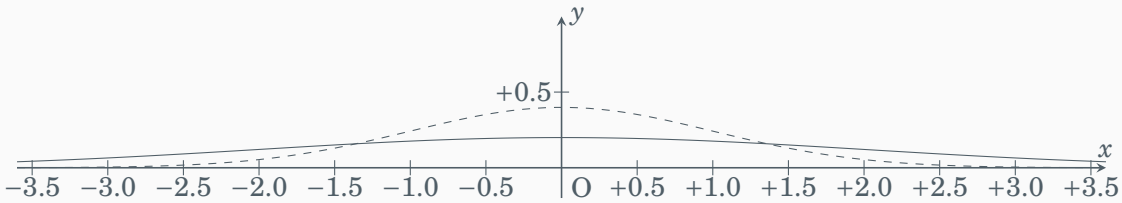

Figure: Normal distributions' PDF (Solid: $\mu = 0, \sigma = 2.0$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

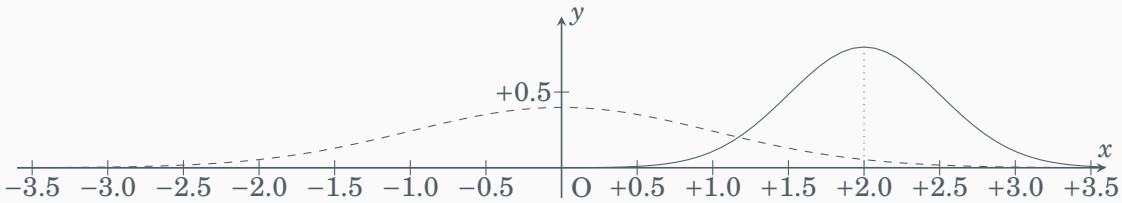

Figure: Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

The mean, the variance, and the standard deviation are μ, σ^2, and $\sigma := \sqrt{\sigma^2}$, respectively.

PDF of the normal distribution

The ***normal distribution***, also known as the ***Gaussian distribution*** with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}_{>0}$ is a distribution with the following PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

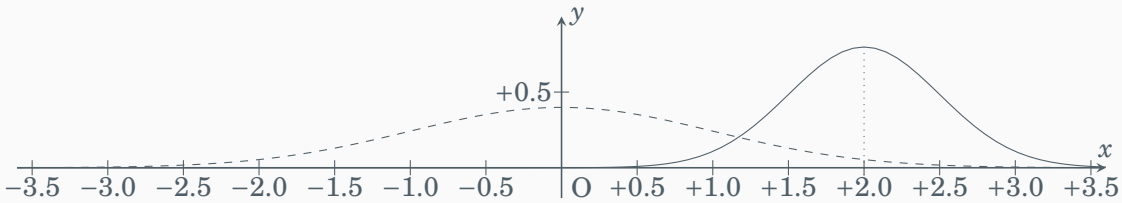

Figure: Normal distributions' PDF (Solid: $\mu = 2, \sigma = 0.5$, Dashed: $\mu = 0, \sigma = 1$).

The PDF is symmetric about $x = \mu$ and it is dense around $x = \mu$.

Outline

1 **Sample Statistics**

Likelihood

Likelihood

To determine a parameter of a parametric model from data points, we quantify how "likely" the distribution indicated by a parameter is correct.

When we have a PMF or PDF of a distribution, we simply define the value of the PMF or PDF of the data points as the ***likelihood*** of the distribution.

Definition (Likelihood of a discrete parametric model)

Let $P(\cdot;\cdot)$ be a discrete parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
Then the ***likelihood*** of $P(\cdot;\boldsymbol{\theta})$ (or often called the likelihood of the parameter $\boldsymbol{\theta}$) is defined as the following product.

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{7}$$

Likelihood

To determine a parameter of a parametric model from data points, we quantify how "likely" the distribution indicated by a parameter is correct.

When we have a PMF or PDF of a distribution, we simply define the value of the PMF or PDF of the data points as the ***likelihood*** of the distribution.

Definition (Likelihood of a continuous parametric model)

Let $p(\cdot;\cdot)$ be a continuous parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
Then the ***likelihood*** of $p(\cdot;\boldsymbol{\theta})$ (or often called the likelihood of the parameter $\boldsymbol{\theta}$) is defined as the following product.

$$p(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot p(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot p(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{7}$$

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1 - \theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{8}$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{8}$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{8}$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{8}$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.
- The likelihood of $\theta = \frac{3}{4}$ is $\frac{3}{4} \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{3}{4} = \frac{27}{256}$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{8}$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.
- The likelihood of $\theta = \frac{3}{4}$ is $\frac{3}{4} \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{3}{4} = \frac{27}{256}$.
- The likelihood of $\theta = 1$ is $1 \cdot 1 \cdot (1-1) \cdot 1 = 0$.

Examples of likelihood calculation

Suppose that we have data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.

The likelihood of the Bernoulli distribution with θ on the data is given by

$$P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta) = P(1;\theta)P(1;\theta)P(0;\theta)P(1;\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta. \tag{8}$$

- The likelihood of $\theta = 0$ is $0 \cdot 0 \cdot (1-0) \cdot 0 = 0$.
- The likelihood of $\theta = \frac{1}{4}$ is $\frac{1}{4} \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{256}$.
- The likelihood of $\theta = \frac{1}{2}$ is $\frac{1}{2} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{16} = \frac{16}{256}$.
- The likelihood of $\theta = \frac{3}{4}$ is $\frac{3}{4} \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{3}{4} = \frac{27}{256}$.
- The likelihood of $\theta = 1$ is $1 \cdot 1 \cdot (1-1) \cdot 1 = 0$.

Hence, among the above three, the distribution given by $\theta = \frac{3}{4}$ most likely generates the data sequence $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$.

Probability and likelihood

The value of the product

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}) \tag{9}$$

can be interpreted as either

- the probability of the random variable sequence taking the value sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$, i.e., a function of a value sequence, or
- the likelihood of the distribution determined by the parameter $\boldsymbol{\theta}$, i.e., a function of a distribution (or parameter).

In other words, the above product is the probability (or the probability density for continuous distribution case) if we interpret it as a function of a value sequence, and the likelihood if we interpret it as a function of a distribution (or a parameter).

Outline

Maximum likelihood estimator

Once we define the likelihood of a distribution, all we need to do is find a parameter that maximizes the likelihood.

The parameter vector that maximizes the likelihood is called the ***maximum likelihood estimator (MLE)***.

Definition (Maximum likelihood estimator)

Let $P(\cdot;\cdot)$ be a discrete parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
The parameter vector $\boldsymbol{\theta}$ is called a maximum likelihood estimator (MLE) if it maximizes the likelihood

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{10}$$

If there is a unique MLE, we often denote it by $\hat{\boldsymbol{\theta}}$.

MLE maximizes the score and minimizes the negative log likelihood

For a parameter vector $\boldsymbol{\theta}$, the following is equivalent[3].

- The parameter vector $\boldsymbol{\theta}$ maximizes the likelihood function

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{11}$$

- The parameter vector $\boldsymbol{\theta}$ maximizes the ***log-likelihood*** function

$$\log P(\boldsymbol{x}_1;\boldsymbol{\theta}) + \log P(\boldsymbol{x}_2;\boldsymbol{\theta}) + \cdots + \log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{12}$$

- The parameter vector $\boldsymbol{\theta}$ minimizes the ***negative log likelihood*** function

$$-\log P(\boldsymbol{x}_1;\boldsymbol{\theta}) - \log P(\boldsymbol{x}_2;\boldsymbol{\theta}) - \cdots - \log P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{13}$$

[3] It follows since $\log$ is an increasing function. It holds regardless of the base of the logarithm.

Why do we consider the logarithm of the likelihood?

- The likelihood is a product and its logarithm is a sum. When we maximize it in a computer, we rely on its derivative (gradient descent methods). Differentiation of a sum is much easier than that of a product, so the (negative) log-likelihood has an advantage over the original likelihood from the optimization viewpoint.
- If the data size m is large, the absolute value of the likelihood, the product of many small values, tends to be too small to represent in a computer (underflow). Since the logarithm sees the power index, it can handle extremely small likelihood.
- The negative log-likelihood can be interpreted as the sum of the errors. For example, we can interpret the negative log-likelihood of the normal distribution as the squared error.

The MLE of the normal distribution minimizes the square error.

Let $p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then, the negative (natural) log-likelihood of the data sequence is given by

$$\begin{aligned} &\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\ &= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right]. \end{aligned} \tag{14}$$

The MLE of the normal distribution minimizes the square error.

Let $p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then, the negative (natural) log-likelihood of the data sequence is given by

$$\begin{aligned} &\log\left(2\pi\sigma^2\right) + \frac{\left(x_1-\mu\right)^2}{2\sigma^2} + \log\left(2\pi\sigma^2\right) + \frac{\left(x_2-\mu\right)^2}{2\sigma^2} + \cdots + \log\left(2\pi\sigma^2\right) + \frac{\left(x_m-\mu\right)^2}{2\sigma^2} \\ &= m\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\left(x_1-\mu\right)^2 + \left(x_2-\mu\right)^2 + \cdots + \left(x_m-\mu\right)^2\right]. \end{aligned} \tag{14}$$

When we minimize the above with respect to μ, we can ignore the gray parts.

In this sense, the MLE of the mean parameter of the normal distribution model is equivalent to minimizing the squared error.

MLE example: Bernoulli case

Example

Suppose that we have data points $x_1, x_2, \ldots, x_m$, and consider the Bernoulli distribution $P(0;\theta) = 1-\theta, P(1;\theta) = \theta$.
The negative log-likelihood of the Bernoulli distribution with θ on the data is given by

$$-\log P(x_1;\theta)P(x_2;\theta)\ldots P(x_m;\theta) = m_0 \log(1-\theta) + m_1 \log\theta, \tag{15}$$

where m_0 and m_1 are the numbers of zeros and ones in the data sequence. Obviously, $m_0 + m_1 = m$, and the sample mean $\bar{x} = \frac{m_1}{m}$. Let l denote the above negative log-likelihood. Suppose that $m_0 \neq 0$ and $m_1 \neq 0$, then l takes the minimum[4] if and only if $\theta = \frac{m_1}{m} = \bar{x}$. Hence, the MLE $\hat{\theta} = \frac{m_1}{m} = \bar{x}$.
For example, if $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, then $\hat{\theta} = \frac{m_1}{m} = \bar{x} = \frac{3}{4}$.

[4] To prove it, differentiate the loss by θ and apply the first derivative test.

Why can we justify the maximum likelihood estimator (MLE)?

Similar to the sample mean, if data points are generated by a distribution indicated by a parameter vector in the parameter set of a parametric vector, the MLE has the following properties:

- ***Consistency***: The MLE converges to the true parameter as $m \to \infty$.
- ***Asymptotic normality***: An appropriately scaled MLE's distribution converges to a normal distribution, and its error is proportional to $\frac{1}{\sqrt{m}}$ for sufficiently large m.

Outline

Exercise (Normal Distribution Arising from Physical Phenomena)

In the following, we consider the left-right directional velocity v (positive for rightward direction) of object A after being collided by m particles from either side, assuming that object A's initial velocity is 0.
Each particle collides with object A from the left with a probability of 0.5, increasing object A's velocity by c, and from the right with a probability of 0.5, decreasing object A's velocity by c. Moreover, the behavior of each particle is independent of others. Assuming that Newton's first law of motion holds ideally, and no other factors alter object A's velocity except for the collisions of the particles. (1) For $m = 2$, find $\Pr(v = -2c)$, $\Pr(v = +2c)$, and $\Pr(v = 0)$.

Exercise (Normal Distribution Arising from Physical Phenomena)

In the following, we consider the left-right directional velocity v (positive for rightward direction) of object A after being collided by m particles from either side, assuming that object A's initial velocity is 0.
Each particle collides with object A from the left with a probability of 0.5, increasing object A's velocity by c, and from the right with a probability of 0.5, decreasing object A's velocity by c. Moreover, the behavior of each particle is independent of others. Assuming that Newton's first law of motion holds ideally, and no other factors alter object A's velocity except for the collisions of the particles. (2) Next, we examine situations where particles are extremely small and numerous, akin to molecules in the atmosphere. Mathematically, this corresponds to scenarios where c is small and m is large. When $c = \frac{4}{\sqrt{m}}$, according to the central limit theorem, the cumulative distribution function (CDF) of v converges in the limit as $m \to +\infty$ to the CDF of a normal distribution with expectation μ and variance σ^2. Here, find μ and σ^2.

Example answer:

Define the discrete random variable X_i as $X_i = +1$ when the ith particle increases the velocity of object A (positive for rightward direction) and $X_i = -1$ when it decreases the velocity.

(1) $\Pr(v = -2c) = \Pr(X_1 = -1 \wedge X_2 = -1) = \Pr(X_1 = -1)\Pr(X_2 = -1) = 0.5 \cdot 0.5 = 0.25$. The second equality follows from the independence of X_1 and X_2. Similarly, $\Pr(v = +2c) = \Pr(X_1 = +1 \wedge X_2 = +1) = \Pr(X_1 = +1)\Pr(X_2 = +1) = 0.5 \cdot 0.5 = 0.25$.

$v = 0$ occurs either when $X_1 = -1$ and $X_2 = +1$, or when $X_1 = +1$ and $X_2 = -1$. Therefore, $\Pr(v = 0) = \Pr((X_1 = -1 \wedge X_2 = +1) \vee (X_1 = +1 \wedge X_2 = -1)) = \Pr(X_1 = -1)\Pr(X_2 = +1) + \Pr(X_1 = +1)\Pr(X_2 = -1) = 0.5 \cdot 0.5 + 0.5 \cdot 0.5 = 0.5$.

Example answer:

(2) According to the central limit theorem, for an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$ with expectation μ and variance σ^2, the cumulative distribution function (CDF) of $Y_m = \sqrt{m}\dfrac{\overline{X}_m - \mu}{\sigma}$ converges to the CDF of a standard normal distribution at every point. Here, $\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ represents the sample mean. Restating the claim of the central limit theorem, the CDF of $Y'_m = \sqrt{m}(\overline{X}_m - \mu)$ converges at every point to the CDF of a normal distribution with expectation 0 and variance σ^2.

Transforming v into a form conducive to the application of the central limit theorem yields

$$v = \sum_{i=1}^{m} cX_i = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} 4X_i = \sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m} 4X_i - \mu\right),$$

where $\mu = 0$ is the expectation of $4X_i$, and the variance of $4X_i$ is 16. Thus, by applying the central limit theorem to the sequence of random variables $4X_1, 4X_2, \ldots$, it can be inferred that the CDF of v converges to the CDF of a normal distribution with expectation 0 and variance 16 as $m \to +\infty$.

Example answer:

This example illustrates how, in the presence of numerous small-scale phenomena, their sum approximates a normal distribution. This is why **the normal distribution holds a special place** in statistics for applications in the natural and social sciences. Furthermore, due to the significance of the normal distribution, **the descriptive statistics characterizing it, such as expectation and variance, are of particular importance** compared to other descriptive statistics.

Exercise (CLT)

Consider an infinite sequence of independently and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$, each with an expected value μ and variance σ^2. Define the sample mean $\overline{X}_m = \frac{\sum_{i=1}^{m} X_i}{m}$ and the standardized variable $Y_m = \frac{\sqrt{m}(\overline{X}_m - \mu)}{\sigma}$.
Additionally, let Z be a standard normal variable, independent of $X_1, X_2, \ldots$, with its probability density function denoted by p_Z, and its cumulative distribution function by F_Z.
Assess whether the following statements are correct or wrong, based on the central limit theorem.

(1) Regardless of the probability distribution of X_i, for $m = 1, 2, \ldots$, the variable Y_m has a probability density function, denoted as $p_{Y_m}(x)$, and choosing it appropriately, for any $x \in \mathbb{R}$, $\lim_{m \to +\infty} p_{Y_m}(x) = p_Z(x)$. **Correct or wrong?**

Exercise (CLT)

Consider an infinite sequence of independently and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$, each with an expected value μ and variance σ^2. Define the sample mean $\overline{X}_m = \frac{\sum_{i=1}^{m} X_i}{m}$ and the standardized variable $Y_m = \frac{\sqrt{m}(\overline{X}_m - \mu)}{\sigma}$.
Additionally, let Z be a standard normal variable, independent of $X_1, X_2, \ldots$, with its probability density function denoted by p_Z, and its cumulative distribution function by F_Z.
Assess whether the following statements are correct or wrong, based on the central limit theorem.

(2) Regardless of the probability distribution of X_i, for $m = 1, 2, \ldots$, the variable Y_m has a cumulative distribution function, denoted as F_{Y_m}, and for any $x \in \mathbb{R}$, $\lim_{m \to +\infty} F_{Y_m}(x) = F_Z(x)$.
Correct or wrong?

Exercise (CLT)

Consider an infinite sequence of independently and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$, each with an expected value μ and variance σ^2. Define the sample mean $\overline{X}_m = \frac{\sum_{i=1}^{m} X_i}{m}$ and the standardized variable $Y_m = \frac{\sqrt{m}(\overline{X}_m - \mu)}{\sigma}$.
Additionally, let Z be a standard normal variable, independent of $X_1, X_2, \ldots$, with its probability density function denoted by p_Z, and its cumulative distribution function by F_Z.
Assess whether the following statements are correct or wrong, based on the central limit theorem.

(3) Regardless of the probability distribution of X_i, for any $a, b \in \mathbb{R}$ with $a < b$, $\lim_{m \to +\infty} \Pr(a < Y_m \leq b) = \Pr(a < Z \leq b)$. **Correct or wrong?**

Example answer:

(1) **Wrong**. The Central Limit Theorem does not guarantee that the sum or average of i.i.d. random variables, $\overline{X}_m$ or Y_m, have probability density functions. In fact, when X_i are discrete random variables, $\overline{X}_m$ and Y_m also remain discrete and thus do not possess probability density functions.

(2) **Correct**. This is a standard expression of the Central Limit Theorem using the cumulative distribution function (CDF). It is worth noting that a CDF is defined for any random variable.

(3) **Correct**. By definition of the CDF, $\Pr(a < Y_m \leq b) = F_{Y_m}(b) - F_{Y_m}(a)$, and this difference converges to $F_Z(b) - F_Z(a)$ due to the Central Limit Theorem (using the CDF expression), which is equal to $\Pr(a < Z \leq b)$.

Exercise (Cauchy distribution)

In this problem, we define the inverse function of $\tan$ restricted to the open interval $(-\frac{\pi}{2}, +\frac{\pi}{2})$ as $\arctan$. That is, for $y \in \mathbb{R}$, $y = \arctan x$ is defined as the unique $y \in (-\frac{\pi}{2}, +\frac{\pi}{2})$ that satisfies $\tan y = x$.
It is known that $\frac{d}{dx} \arctan(x) = \frac{1}{x^2+1}$.
Let the random variable X have the probability density function p_X defined as follows:

$$p_X(x) = \frac{1}{\pi} \frac{1}{x^2+1}$$

(1) Evaluate $\Pr(1 \le X \le \sqrt{3})$.
(2) Considering $p_X(x)$ is an even function, meaning $p_X(-x) = p_X(x)$, evaluate the median of X.

Exercise (Cauchy distribution)

Let the random variable X have the probability density function p_X defined as follows:

$$p_X(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}$$

(3) To calculate the expected value of X, we evaluate the improper integral $\int_{-\infty}^{+\infty} x p_X(x) dx$.

(3-1) Consider the following two definite integrals, keeping in mind that $x p_X(x)$ is an odd function, i.e., $(-x)p_x(-x) = -x p_x(x)$. Find the value of $\lim_{t \to +\infty} \int_{-t}^{+t} x p_X(x) dx$.

Exercise (Cauchy distribution)

Let the random variable X have the probability density function p_X defined as follows:

$$p_X(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}$$

(3) To calculate the expected value of X, we evaluate the improper integral $\int_{-\infty}^{+\infty} x p_X(x) dx$.

(3-2) Find the value of $\lim_{a \to -\infty} \int_a^0 x p_X(x) dx + \lim_{b \to +\infty} \int_0^b x p_X(x) dx$.

Exercise (Cauchy distribution)

Let the random variable X have the probability density function p_X defined as follows:

$$p_X(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}$$

(3) To calculate the expected value of X, we evaluate the improper integral $\int_{-\infty}^{+\infty} x p_X(x) dx$.

(3-3) Write down the definition of $\int_{-\infty}^{+\infty} x p_X(x) dx$.

Exercise (Cauchy distribution)

Let the random variable X have the probability density function p_X defined as follows:

$$p_X(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}$$

(3) To calculate the expected value of X, we evaluate the improper integral $\int_{-\infty}^{+\infty} x p_X(x) dx$.

(3-4) Find the expected value of X.

Exercise (Cauchy distribution)

(4) Given a sequence of mutually independent random variables $X_1, X_2, X_3, \ldots$ with the same distribution as X, define the random variable $\overline{X}_m = \frac{1}{m}\sum_1^m X_i$ for $m = 1, 2, \ldots$. Choose the correct statement from the following options:

- According to the law of large numbers (strong form), it can be stated that the event $\overline{X}_m \to 0$ occurs with probability 1.
- According to the law of large numbers (strong form), there exists some positive number c such that the event $\overline{X}_m \to c$ occurs with probability 1.
- According to the law of large numbers (strong form), there exists some negative number c such that the event $\overline{X}_m \to c$ occurs with probability 1.
- The law of large numbers does not apply to the behavior of $\overline{X}_m$.

Example answer:

(1) $\Pr(1 \leq X \leq \sqrt{3}) = \int_1^{\sqrt{3}} p_X(x)dx = \int_1^{\sqrt{3}} \frac{1}{\pi}\frac{1}{x^2+1}dx$ Using the fact that $\frac{d}{dx}\arctan(x) = \frac{1}{x^2+1}$, this can be calculated as follows:
$\int_1^{\sqrt{3}} \frac{1}{\pi}\frac{1}{x^2+1}dx = \frac{1}{\pi}[\arctan x]_1^{\sqrt{3}} = \frac{1}{\pi}(\frac{\pi}{3} - \frac{\pi}{4}) = \frac{1}{12}$.

(2) Given that p_X is an even function, $\int_{-\infty}^{0} p_X(x)dx = \int_0^{+\infty} p_X(x)dx$. Thus, $\Pr(X \leq 0) = \Pr(X \geq 0)$. Therefore, the median is 0.

Example answer:

(3-1) Since $xp_X(x)$ is an odd function, for any t, $\int_{-t}^{+t} xp_X(x)dx = 0$. Therefore, $\lim_{t\to+\infty}\int_{-t}^{+t} xp_X(x)dx = 0$.

(3-2) To evaluate $\lim_{a\to-\infty}\int_a^0 xp_X(x)dx$, consider $\int_a^0 xp_X(x)dx = \frac{1}{\pi}\int_a^0 \frac{x}{x^2+1}dx$. Using the substitution $u = x^2+1$, the integral can be calculated as $\int_a^0 \frac{x}{x^2+1}dx = \int_{a^2+1}^1 \frac{1}{2u}du = [\ln u]_{a^2+1}^1 = -\ln(a^2+1)$. Therefore, $\lim_{a\to-\infty}\int_a^0 xp_X(x)dx = -\infty$.

Similarly, evaluating $\lim_{b\to+\infty}\int_0^b xp_X(x)dx$ results in $\lim_{b\to+\infty}\int_0^b xp_X(x)dx = \lim_{b\to+\infty}\ln(b^2+1) = +\infty$, diverging. Hence, $\lim_{a\to-\infty}\int_a^0 xp_X(x)dx + \lim_{b\to+\infty}\int_0^b xp_X(x)dx$ forms a $-\infty+\infty$ shape and is not defined as a finite real number.

Example answer:

(3-3) Since the improper integral $\int_{-\infty}^{+\infty} xp_X(x)dx$ is defined as $\lim_{a\to-\infty}\int_a^0 xp_X(x)dx + \lim_{b\to+\infty}\int_0^b xp_X(x)dx$, this improper integral is undefined, and thus the expected value of X is not defined as a finite real number.

(3-4) According to (3-3), X does not have an expectation.

(4) Since X_i does not have an expected value, the law of large numbers cannot be applied to describe the behavior of $\overline{X}_m$.

Exercise (Likelihood and MLE)

Given a sequence of data points $(x_1, x_2, x_3, x_4) = (1, 1, 0, 1)$, consider the likelihood under a discrete parametric probability model, the Bernoulli distribution, where $P(0;\theta) = (1-\theta)$ and $P(1;\theta) = \theta$.
Evaluate the likelihood for $\theta = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$.
Also, find the maximum likelihood estimate (MLE).

Example answer:

Discrete parametric probability models have parameters, and fixing these parameters defines a probability mass function. Representing the probability mass function determined by parameter θ as $P(\cdot;\theta)$, the likelihood for a sequence of data points (x_1,x_2,x_3,x_4) is defined as $\prod_{i=1}^{4} P(x_i;\theta) = P(x_1;\theta)P(x_2;\theta)P(x_3;\theta)P(x_4;\theta)$.

Therefore, the likelihood for $\theta = \frac{1}{4}$ is $\prod_{i=1}^{4} P(x_i;\frac{1}{4}) = \frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} = \frac{3}{256}$, the likelihood for $\theta = \frac{2}{4}$ is $\prod_{i=1}^{4} P(x_i;\frac{2}{4}) = \frac{2}{4}\cdot\frac{2}{4}\cdot\frac{2}{4}\cdot\frac{2}{4} = \frac{16}{256}$, and the likelihood for $\theta = \frac{3}{4}$ is $\prod_{i=1}^{4} P(x_i;\frac{3}{4}) = \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4} = \frac{27}{256}$.

The maximum likelihood estimator is the parameter θ that maximizes the likelihood. For this problem, we consider maximizing the likelihood function $\prod_{i=1}^{4} P(x_i;\theta) = \theta\cdot\theta\cdot(1-\theta)\cdot\theta = \theta^3(1-\theta)$. Maximizing this likelihood function is equivalent to maximizing $\sqrt[4]{\theta^3[3(1-\theta)]}$ due to the monotonicity of the fourth root function. By the inequality of arithmetic and geometric means, $\sqrt[4]{\theta^3[3(1-\theta)]} \leq \frac{1}{4}(\theta+\theta+\theta+3(1-\theta)) = \frac{3}{4}$, and equality holds when $\theta = 3(1-\theta)$, that is, when $\theta = \frac{3}{4}$. Therefore, the likelihood function reaches its maximum value when $\theta = \frac{3}{4}$. This is the maximum likelihood estimator.