Lecture 4

Fundamentals of Statistical Analysis

Contents

Chemical data are never exact. Whether obtained from experiments, simulations, or databases, they inevitably reflect (1) *variability*, (2) *uncertainty (noise)*, and (3) *incomplete information*.

Probability provides a mathematical framework for describing and reasoning about such imperfect data.

Probability, however, is more than a mathematical concept. It offers a philosophical perspective on scientific problems that departs from the usual deterministic viewpoint and can therefore appear counterintuitive at first. Probability is also deeply connected to fundamental physical concepts. For example, joint probability distributions describe dependencies among multiple variables, analogous to coupled or interacting subsystems. Covariance quantifies correlations between physical observables. In quantum mechanics, the state of a system is itself described probabilistically through a wavefunction. For these reasons, this lecture should be viewed not only as a prerequisite for machine learning, but also as training in probabilistic thinking that is central to modern physical and chemical sciences.

Statistical analysis provides systematic methods for interpreting probabilistic descriptions. In this sense, probability can be viewed as a *representation* of uncertainty in a problem, while statistical analysis serves as an *interpretation* of that representation based on observed data.

In this lecture, we will introduce the fundamentals of probability and statistical analysis and discuss numerical methods for making predictions from noisy data. These methods can be viewed as precursors to ML models and form the basis for key concepts in probabilistic data analysis, including variance and covariance, likelihood, and confidence.

1 Probability Theory

Formally, probability is a mapping that assigns a real number between 0 and 1 to events associated with data of interest. In this section, we give a rigorous definition of this mapping. We can use experimental measurements to represent the action of drawing a data with uncertainty. We call the act of doing a single experiment a ***trial***.

1.1 Sample Space, Events, and Random Variables

There are three conceptual layers used to describe experimental measurements: the **sample space**, **events** and **random variables**.

- **Sample space** Ω: the set of *all possible outcomes* of a random experiment.
  E.g., For flipping two coins: $\Omega = \{\text{HH, HT, TH, TT}\}$.
- **Event** A: a subset of the sample space, $A \subseteq \Omega$.
  E.g., An event of for flipping two coins could be $A = \{\text{HH, TH}\}$.
- **Random variable** X: a function that assigns a numerical value to each outcome in the sample space, translating *real-world* outcomes into *mathematical* quantities.

E.g., One possible random variable for flipping two coins is

$$X(\mathrm{HH}) = 0, \quad X(\mathrm{HT}) = 1, \quad X(\mathrm{TH}) = 2, \quad X(\mathrm{TT}) = 3$$

We now clarify these concepts in more detail.

1. ***Why is an event A a subset of the sample space Ω?***
   You should regard the sample space as a *theoretical* concept, while an event is a *real-life* concept tied to actual experimental practice. For example, suppose we decide to flip two coins repeatedly, say 10 times. This sequence of 10 trials constitutes an event. In practice, we cannot guarantee that all possible outcomes in the sample space will appear within these 10 repetitions. If we collect all distinct outcomes observed in these trials, the resulting set will contain a number of elements that is less than or equal to the total number of possible outcomes in the sample space.
   Therefore, an event can be understood as the set of distinct outcomes realized in a *finite number* of repeated experiments, which should be a subsets of the sample space rather than the sample space itself.
2. ***Is the mapping between outcomes in Ω and random variables one-to-one?***
   Not necessarily. The only requirement for a random variable is that every possible outcome in the sample space Ω must be assigned a numerical value. There is no requirement that this assignment be one-to-one. In other words, different outcomes may be mapped to the same value of a random variable.
   For example, suppose that in the two-coin experiment we are only interested in the number of heads, regardless of the order in which they appear. In this case, we can define the random variable as
   $$X(\mathrm{HH}) = 2, \quad X(\mathrm{HT}) = 1, \quad X(\mathrm{TH}) = 1, \quad X(\mathrm{TT}) = 0$$
   Here, the outcomes HT and TH are distinct elements of the sample space, but they correspond to the same value of the random variable because they represent the same physical quantity of interest.
   We may therefore use a simpler variable x to denote the value of $X(\Omega)$: $x = 2$ correpsonds to the outcome HH, $x = 1$ corresponds to either HT or TH, and $x = 0$ corresponds to TT.

Once we understand the concepts of sample space, events, and random variables, we can formally introduce probability and probability distributions. These concepts allow us to quantify how likely different outcomes or values of a random variable are to occur.

Quiz: What are all possible events for flipping two coins? Hint: it should be a combinatorial problem.

1.2 Probability

Probability is a numerical measure of how likely an event is to occur. More intuitively, if we perform a trial and observe an outcome ω, the probability of an event A is the likelihood that $\omega \in A$.

Formally, probability maps an event to a number between 0 (the event cannot occur) and 1 (the event will certainly occur):

$$P(A) = y, \quad y \in [0, 1].$$

From a mathematical perspective, a function is a mapping from one set to another. Here, the target set is $[0, 1]$, and the set being mapped is a collection of events $\mathcal{F} = \{A_1, A_2, \dots\}$. This gives:

$$P : \mathcal{F} \to [0, 1].$$

Axioms of probability

- **Non-negativity:** $P(A) \geq 0$ for any event A.
- **Normalization:** $P(\Omega) = 1$, where Ω is the sample space.
- **Additivity:** If A and B are mutually exclusive,

$$P(A \cup B) = P(A) + P(B), \quad \text{if } A \cap B = \emptyset$$

Here, the symbol $\cup$ denotes the *union* of two sets. For example, if

$$A = \{\text{HH, HT}\}, \quad B = \{\text{TT}\}, \quad \text{then } A \cup B = \{\text{HH, HT, TT}\}.$$

The symbol $\cap$ denotes the *intersection* (shared elements) of two sets. For mutually exclusive events, $A \cap B = \emptyset$.

A generalization of the additivity axiom for any events A and B is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Example: Two-coin experiment

Consider flipping two coins. Each outcome has equal probability $1/4$. Let

$$A = \{\text{HH}\}, \quad B = \{\text{TT}\}.$$

Then

$$P(A) = \frac{1}{4}, \quad P(B) = \frac{1}{4}, \quad P(A \cup B) = \frac{1}{2}.$$

1.2.1 Continuous outcomes

The coin example involves **discrete outcomes**. Many experiments, however, have **continuous outcomes**, such as concentrations, temperatures, or reaction rates. In these cases, events correspond to *continuous regions* of the sample space.

For instance, the concentration of a NaCl-in-water solution ranges from 0% to the saturation limit 26.47% by weight at room temperature. We define the sample space as

$$\Omega = [0\%, 26.47\%],$$

and an event could be a specific concentration range, e.g.,

$$A = [5\%, 10\%].$$

For continuous outcomes, probabilities are associated with *regions*, not individual points.

1.3 Probability Distribution

While probability tells us about an event, a *probability distribution* tells us how the total probability (which is 1) is spread across all possible values of a random variable X.

Since the measurement outcomes can fall into a discrete or continuous sample space, the random variables can also be discrete or continuous.

1.3.1 Discrete distributions

We use the **probability mass function** (PMF) to describe discrete distributions. The probability that X is exactly equal to a value x is simply given by the value of PMF:

$$P(X = x) = p(x). \tag{1}$$

Note that we use the lower cased p for probability distribution, and upper cased P for probability.

1.3.2 Continuous distribution

We use **probability density function** (PDF) $p(x)$ to describe continuous distributions. Just like a weight equal to the density times a certain volumn, the probability for a continuous distribution

should be combined with a "volumn", i.e., an interval (event) $[a, b]$:

$$P(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x \tag{2}$$

Note that $p(x)$ itself is not a probability, but a density!

From the axioms for probability, we can derive that

- $p(x) \geq 0$
- $\int_{x \in X(\Omega)} p(x)\mathrm{d}x = 1$
- $P([a, b] \cup [c, d]) = \int_a^b p(x)\mathrm{d}x + \int_c^d p(x)\mathrm{d}x, \quad c \geq b.$

For simplicity, we assume that $X(\Omega) = (-\infty, \infty)$.

1.3.3 Cumulative distribution functions (CDF)

You might also heard of cumulative distribution functions (CDF), they are simply the summation or integration of probability distributions up to a certain value.

- Discrete case:

$$C(x) = \sum_{z \leq x} p(x). \tag{3}$$

- Continuous case:

$$C(x) = \int_{-\infty}^{x} p(z)\mathrm{d}z. \tag{4}$$

1.3.4 Mean and variances

The mean of a random variable is the average value it would take in the long run. The mean is also called the *expectation value* of X.

- Discrete variable:

$$\langle X \rangle = \sum_x x p(x). \tag{5}$$

- Continuous variable:

$$\langle X \rangle = \int_{-\infty}^{\infty} x p(x)\ \mathrm{d}x \tag{6}$$

The variance measures the spread (uncertainty) of a random variable, which is the average distance square from its mean value. Common symbols are $\sigma^2(X)$ and $\mathrm{Var}(X)$.

- Discrete:

$$\mathrm{Var}(X) = \sum_x (x - \langle X \rangle)^2 \, p(x) \tag{7}$$

- Continuous:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \langle X \rangle)^2 \, p(x) \, \mathrm{d}x \tag{8}$$

1.4 Common Probability Distributions

Since in ML, the continuous distribution is more used. We will only consider some common continuous distributions that are used in ML tasks.

- **Uniform distribution**
  A uniform distribution is defined in a finite sample space, say $X \in [a, b]$, and

$$\text{PDF: } p(x) = \frac{1}{b-a}, \quad \text{Mean: } \langle X \rangle = \frac{a+b}{2}, \quad \text{Variance: } \mathrm{Var}(X) = \frac{(b-a)^2}{12}. \tag{9}$$

- **Normal (Gaussian) distribution** Most common for measurement noise and naturally arising distributions, $X \in (-\infty, \infty)$.

$$\text{PDF: } p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad \text{Mean: } \langle X \rangle = \mu, \quad \text{Variance: } \mathrm{Var}(X) = \sigma^2. \tag{10}$$

- **Exponential distribution** The exponential distribution models the time between events in a Poisson process, or other positive continuous variables, $X \in [0, \infty)$.

$$\text{PDF: } p(x) = \lambda e^{-\lambda x}, \quad \text{Mean: } \langle X \rangle = \frac{1}{\lambda}, \quad \text{Variance: } \mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{11}$$

- **Log-normal distribution** X is log-normally distributed corresponds to $\ln X$ is normally distributed. It is useful for positive variables spanning *multiple orders of magnitude*.

$X \in (0, \infty)$.

$$\begin{aligned} &\text{PDF: } p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \\ &\text{Mean: } \langle X \rangle = e^{\mu+\sigma^2/2}, \quad \text{Variance: } \mathrm{Var}(X) = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}. \end{aligned} \tag{12}$$

- **Beta distribution** The beta distribution is defined on a finite interval $X \in [0, 1]$ and is often used to model probabilities or fractions.

$$\begin{aligned} &\text{PDF: } p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \\ &\text{Mean: } \langle X \rangle = \frac{\alpha}{\alpha+\beta}, \quad \text{Variance: } \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \end{aligned} \tag{13}$$

  where $B(\alpha, \beta)$ is the Beta function.

1.5 Joint Probability Distribution

In many experiments, we measure more than one type of random variable. For example, for a solution, we may simultaneously measure its concentration and its conductivity. We can represent these outcomes with two random variables, X and Y.

The **joint PDF** describes the probability density of observing $X = x$ and $Y = y$ simultaneously:

$$p(x, y). \tag{14}$$

If X and Y are independent (i.e., uncorrelated), then the joint PDF factorizes:

$$p(x, y) = p(x)\, p(y).$$

Expectation values

The expectation (mean) of X is given by

$$\langle X \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, p(x, y)\, \mathrm{d}x\, \mathrm{d}y, \tag{15}$$

which is analogous to the single-variable case, except that we now integrate over both x and y, and replace $p(x)$ with the joint distribution $p(x, y)$.

Variance

Similarly, the variance of X can be calculated as

$$\mathrm{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2, \tag{16}$$

with

$$\langle X^2 \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 \, p(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Joint distributions allow us to describe correlations between variables. If X and Y are correlated, $p(x, y) \neq p(x)p(y)$, and covariance becomes relevant (explained later). The concepts generalize to more than two variables by extending the integrals to multiple dimensions.

1.5.1 Covariance

The covariance of two random variables is very important in feature analysis and feature selection, such as principal component analysis (PCA). It measures how two variables vary together.

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [(x - \langle X \rangle)(y - \langle Y \rangle) p(x, y)] \; \mathrm{d}x\mathrm{d}y = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle, \tag{17}$$

which is the expectation value of $(X - \langle X \rangle)(Y - \langle Y \rangle)$ under the joint distribution $p(x, y)$. We can prove that

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [xyp(x, y)] - \langle X \rangle \langle Y \rangle \, \mathrm{d}x\mathrm{d}y \\ &= \langle XY \rangle - \langle X \rangle \langle Y \rangle \end{aligned} \tag{18}$$

The above can also be called the **correlation function** between X and Y.

- $\mathrm{Cov}(X, Y) > 0$: X and Y tend to increase together.
- $\mathrm{Cov}(X, Y) < 0$: X increases when Y decreases.
- $\mathrm{Cov}(X, Y) = 0$: no linear relationship (they may still be nonlinearly dependent, and we need to examine higher order relations.)

We can normalize $\mathrm{Cov}(X, Y)$ into a **correlation coefficient**.

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1], \tag{19}$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$, $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$.

1.5.2 Marginal distributions

Suppose we only know the joint distribution $p(x, y)$, and we would like to recover the distribution of one variable, we could simply integrate the other variable out

$$p_X(x) = \int_{-\infty}^{\infty} p(x, y)\, \mathrm{d}y, \quad p_Y(y) = \int_{-\infty}^{\infty} p(x, y)\, \mathrm{d}x \tag{20}$$

the single distributions recovered from a joint distribution is called **marginal distributions**.

Example Discrete marginal probability.

Let's look at a simple example of discrete probability. Suppose both X and Y can be two numbers 0 or 1, i.e., $X \in \{0, 1\}$ and $Y \in \{0, 1\}$, and we have the following probabilities of $P(X, Y)$:

$$P(0, 0) = \frac{1}{8}, \quad P(0, 1) = \frac{1}{2}, \quad P(1, 0) = \frac{1}{4}, \quad P(1, 1) = \frac{1}{8}.$$

What is the probability of $X = 0$ regardless of the Y value? We simply sum up all outcomes with $X = 0$:

$$P_X(0) = P(0, 0) + P(0, 1) = \sum_{y=0}^{1} P(0, y) = \frac{5}{8}.$$

The integration over y in the continuous case is equivalent to the summation in the discrete case, so we recover the marginal distribution in Eq. (20).

1.5.3 Conditional Probability Distribution

The conditional probability distribution is very useful for Bayesian modeling and ML, it describes the distribution of a variable *given* another variable.

$p(x|y)$ means the distribution of x given the distribution of y, so it is different from $p(x)$, we say it is *conditioned*. The math definition is

$$p(x|y) = \frac{p(x, y)}{p_Y(y)}, \quad p(y|x) = \frac{p(x, y)}{p_X(x)}. \tag{21}$$

Example Discrete conditional probability.

Now let's still look at the previous example to understand this definition. Given

$$P(0, 0) = \frac{1}{8}, \quad P(0, 1) = \frac{1}{2}, \quad P(1, 0) = \frac{1}{4}, \quad P(1, 1) = \frac{1}{8},$$

what is the probability of $Y = 1$ given $X = 0$? We only look at

$$P(0,0) = \frac{1}{8}, \quad P(0,1) = \frac{1}{2}$$

and in the above set, the probability of $Y = 1$ is given by

$$P(Y = 1|X = 0) = \frac{P(0,1)}{P(0,0) + P(0,1)} = \frac{P(0,1)}{P_X(0)}.$$

We can also rewrite Eq. (21) using a **chain rule**:

$$p(x,y) = p(x|y)p_Y(y) = p(y|x)p_X(x). \tag{22}$$

2 Statistical Inference

So far we have learned the fundamentals of probability theory. Life would be much simpler if we knew the probability distribution of every variable! In reality, we only get a set of observations. We suspect that there is some hidden pattern underlying the data, but we have no idea what it is.

This is where statistical inference comes in. It gives us a way to make educated guesses about the data-generating process. Even without knowing the full sample space or the exact distribution, we can still estimate key properties, quantify uncertainty, and make predictions. In other words, statistical inference lets us move from abstract probability to practical understanding of real data.

2.1 Sample Statistics

Just like event is a subset of the whole sample space, in real life, we cannot enumerate all possible outcomes of an experiment. In statistics, we use **population** to represent the entire set of possible outcomes, while a **sample** is the finite subset we actually observe.

Suppose we measure a property X for n molecules. Then our sample is:

$$\{x_1, x_2, \ldots, x_n\}.$$

Each x_i is a realization of the random variable X. The population would be the property of all molecules, which is impossible to have!

When drawing samples, we adopt the **i.i.d. assumption:**[1], which means each x_i is drawn from the same underlying distribution and does not influence others.

We can use the sample to approximate the real probability distribution of the population.

[1] i.i.d.: independent and identically distributed

- **Sample mean:**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This estimates the expectation $\langle X \rangle$ of the underlying random variable.

- **Sample variance:**

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

This estimates the population variance $\mathrm{Var}(X)$. Notice we divide by $n-1$ to correct for bias, known as *Bessel's correction*.

- **Sample covariance (for two variables X and Y):**

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

This estimates how two variables vary together. Normalizing by the standard deviations gives the correlation coefficient:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

Why do we have $n-1$ instead of n here?

Note that we are doing two approximations here instead of one.

- Using a sample $\{x_i\}_{i=1}^n$ to approximate the whole population.
- Using the mean of the sample $\bar{X}$ to approximate the true mean $\langle X \rangle$.

The true mean $\langle X \rangle$ does not depend on the sample, but the sample mean $\bar{X}$ is derived from the sample! Therefore, there are actually $n-1$ degrees of freedom in the sample variance in Eq. (2.1) given the value of $\bar{X}$, instead of n degrees in the formula $\sum_{i=1}^n (x_i - \langle X \rangle)^2$. That is where the $n-1$ come from.

2.1.1 Law of Large Numbers

Although a sample is just a subset of the population, the sample statistics is still meaningful due to the *law of large numbers* (LLN).

The LLN states that as the number of independent and identically distributed (i.i.d.) trials (sample size) increases, their average $\bar{X}$ gets closer and closer to the expected value $\langle X \rangle$ of the population.

2.2 Estimating the parameters of probability

2.2.1 Estimator

An *estimator* is a function of the sample data used to estimate an unknown population parameter. When we change our sample, the estimation also changes. Therefore, we can view the estimator itself as a random variable, which has all the statistical properties such as the mean and variance.

If you took the quantum chemistry course, you know that the true wavefunction of the quantum state is too complicated to be known, but we can make some approximated function formed, called *ansatz*. You can see ansatz as some sort of estimators of the wavefunction.

Suppose we are using an estimator $\tilde{\theta}$ to estimate the population parameter θ. A good estimator should satisfy the following criteria

- **Unbiasedness**. The mean of the estimator equal to the true parameter: $\langle\tilde{\theta}\rangle = \theta$.
  The bias of $\tilde{\theta}$ is thus defined as
  $$\text{Bias}(\tilde{\theta}) = \langle\tilde{\theta}\rangle - \theta$$
- **Consistency**. $\tilde{\theta} \to \theta$ as the sample size $n \to \infty$.
- **Efficiency**. If you have two unbiased estimators, the one with the lower variance is more efficient.

Mean Squared Error (MSE) is the ultimate "quality scorecard" for an estimator.

$$\text{MSE}(\tilde{\theta}) = \langle(\tilde{\theta} - \theta)^2\rangle = \text{Var}(\hat{\theta}) + (\text{Bias}[\hat{\theta}])^2. \tag{23}$$

The bias talks about the *accuracy* of the estimator, while the variance about the *precision*. Therefore, the MSE is sourced from both accuracy and precision.

MSE has been served as the loss function of models such as linear regression.

2.2.2 Likelihood

When you have eliminated the impossible, whatever remains, however improbable, must be the truth.

– Sherlock Holmes

Likelihood provides an elegant example of how we can approach the same problem from different angles. If the estimator is *forward-thinking*, then likelihood is *reverse engineering* (like the quote from *Sherlock Holms*).

- The Estimator follows this path: "I assume there is a true, hidden parameter θ. Given that, what is the best mathematical rule—the estimator $\tilde{\theta}$—to extract a guess from a finite

sample?"

- Likelihood reverses that thought process: "I have already made my observations. Now, what is the most plausible value of θ that would make these specific observations reasonable?"

This type of reverse engineering is exactly how we find our best guess for the truth. While the goal is still to find a reliable estimator, the "search engine" we use to find it is **maximum likelihood estimation** (MLE).

If we have a model with parameters θ and we observe data x, the **likelihood function** $L(\theta; x)$ is numerically equal to the conditional probability of *seeing that data given those parameters*.

$$L(\theta; x) = p(x|\theta). \tag{24}$$

In probability $p(x|\theta)$, θ is assumed to be a known value (fixed), while x is a variable, while in the likelihood function $L(\theta; x)$, x is given, and we want to optimize θ so that the observations have the largest chance to occur.

Now if we have many observed data $\{x_1, x_2, \cdots x_n\}$, the likelihood function is

$$L(\theta; \{x_i\}) = \prod_{i=1}^{n} p(x_i|\theta). \tag{25}$$

Example: If we know the probability distribution is Gaussian,$X \sim \mathcal{N}(\mu, \sigma^2)$, and σ is known, we want to find out the parameter μ, the likelihood is:

$$L(\mu; \{x_i\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right].$$

Maximizing this likelihood gives the most plausible value of μ given the observations $\{x_i\}$.

2.2.3 Maximum Likelihood Estimator (MLE)

MLE is one of the most widely used methods for parameter estimation. Line fitting can be seen as a specific case of MLE, where given a set of discrete dots on the x-y plane, you want to find the best smooth curve to estimate the function describing those dots.

Let's use $\tilde{\theta}_{\text{MLE}}$ to denote this estimator, then

$$\tilde{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta; \{x_i\}) \tag{26}$$

Example: Let's still look at the above example of estimating the μ for a Gaussian distribution, we would get

$$\tilde{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}.$$

In practice, it is easier to maximize the logorithm of $L(\theta; \{x_i\})$, i.e., $\ln L(\theta; \{x_i\})$, which turns the product $\prod_i$ into a sum $\sum_i$, where the latter is usually easier to optimize.

2.2.4 Uncertainty Quantification

Estimating parameters is only half the story; we also need to know how uncertain these estimates are. There are two key concepts to evaluate our estimations: **standard error** and **confidence intervals**.

- **Standard error (SE):** Standard deviation of an estimator,

$$\text{SE}(\bar{X}) = \frac{S}{\sqrt{n}}, \tag{27}$$

  where n is the sample size, and S is the sample standard deviation.

- **Confidence intervals (CI):** The range where you can confidently say the true parameter resides.
  E.g., for the Gaussian distribution example, the CI of μ is

$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

  where $z_{\alpha/2}$ is a value related to the confidence level $(1 - \alpha)$.

2.3 Connection to Real Tasks

In the next lecture, we will learn some statistical analysis techniques, such as regression and QSAR (Quantitative Structure–Activity Relationship). This lecture's content is fundamental for use to better understand the handy statistical tools.

- Linear regression can be interpreted as MLE under Gaussian noise.
- Covariance and correlation inform feature selection.
- QSAR modeling fits chemical descriptors to activity using statistical inference as the foundation.
- Understanding uncertainty in estimates prepares us for probabilistic ML, Bayesian regression, and Gaussian processes.

3 Lab

We will have a short Lab session in the lecture to illustrate the maximum likelihood estimator (MLE). We will practice with the following packages

- `numpy`.
- `scipy`.
- `matplotlib`.

`https://colab.research.google.com/drive/1sfhe9uHdCxMLvWa1Kowzvyk1WKdIsH3D?usp=sharing`