IEDA 2540 Statistics for Engineers
Probability Distributions

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Spring, 2025

Introduction

Last time we saw descriptive statistics.

- Graphical: pie chart, bar chart, stem-and-leaf diagram, box plot, histogram.
- Numerical: sample range, sample mean, sample quartile, IQR, sample standard deviation, sample correlation.

Descriptive statistics provides an initial look into the dataset you collected.

- Next, we need a more sophisticated analysis of the data.
- Inferential statistics provides mathematical tools to infer the characteristics of, or make assertions about, the population from the sample.
- In this topic, we will explore how random variables are essential in building the framework for these inferential methods.

Introduction

If the population has no underlying structure, drawing scientifically valid conclusions becomes impossible. To resolve this, we assume that each observation is a **random variable**.

- **Question 1:** What is a random variable?
- **Question 2:** For a given experiment, what is a suitable random variable to model it? $\Rightarrow$ In this lecture, we review random variables commonly seen in statistics.
- **Question 3:** After we pinpoint a random variable, how do we use it to make statistical claims? $\Rightarrow$ This will be addressed later in the course.

Introduction to Random Experiments

Statistics regards experiments as random and focuses on their **outcomes**.

Random Experiment

A **random experiment** is an experiment in which the outcome is not known until the experiment is performed.

Example: Rolling two dice, where an outcome could be $(6, 6)$ or $(3, 1)$, etc.

Sample Space

The **sample space** is the set of all possible outcomes of a random experiment.

Example: For two dice, this includes all ordered pairs from $(1, 1), (1, 2), \ldots, (6, 6)$.

Outcomes and Events

Statistics regards experiments as random. Statistician cares about the outcomes that satisfies a certain description.

Event

An **event** is a specific subset of the sample space that satisfies a given condition. For instance, you win $1 if the event "the sum of the two dice is 6" occurs, which corresponds to the outcomes $\{(1,5),(2,4),(3,3),(4,2),(5,1)\}$.

The outcomes of a random experiment and the corresponding events may be difficult to handle mathematically:

Example: when rolling two dice, the outcome is an ordered pair of numbers, which does not lend itself directly to arithmetic operations.

Example: In clinical trials, the outcome might be a group of individuals with various treatment results, complicating direct analysis.

What is a Random Variable?

To perform statistical analysis, we would like to simplify these outcomes by assigning numerical values to it, so that mathematical calculation is possible. This process leads to the concept of a **random variable**.

Random variable

A random variable is a function that assigns a real number to each outcome of a random experiment.

- For random experiments, the outcome is not known in advance.
- Hence, the associated value (the variable!) to describe the even of interest is also unknown.

What is a Random Variable?

Example: Roll two dice, the outcome could be $(3, 1)$.

Let

$$X = \text{“sum of the numbers shown on the dice.”}$$

We have

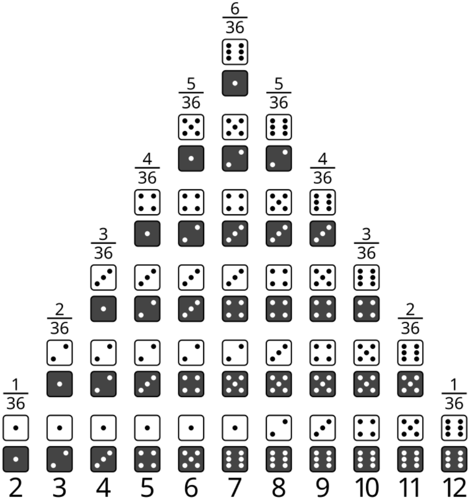
$$X(\{\underbrace{(3,1)}_{\text{an outcome}}\}) = \underbrace{4}_{\text{the value of the RV}},$$

$$X(\{\underbrace{(6,6)}_{\text{an outcome}}\}) = \underbrace{12}_{\text{the value of the RV}}.$$

- After the experiment, we now observe a deterministic/known number. This is called a **realization** of the random variable.

X maps an outcome (**not an event**) to a number.

Example: Rolling Two Dice



Random Variable

Notation convention

- A generic random variable is denoted by an uppercase letter such as X, Y, N.
- After the experiment is conducted, the observed value/realization (deterministic/known) is denoted by a lowercase letter such as x_i, y_i, n_i.

Types of Random Variables

Range of a random variable

The range of a random variable is all the possible values that it can take.

Example: What is the range of: $X =$ "sum of the numbers shown on the dice."

Discrete Random Variable

When a random variable is a discrete variable, we call it a **discrete random variable**. Equivalently, the range is finite (or countably infinite).

Continuous Random Variable

When a random variable is a continuous variable, we call it a **continuous random variable**. Equivalently, the range is an interval of real numbers.

We will review some of the commonly used probability distributions.

Probability Distributions: Why should we care?

Many commonly seen experiments can be characterized by simple distributions.

Example: consider the toss of a coin. The coin has two possible outcomes: $\{H, T\}$, where H appears with a probability p. The outcome is characterized by a simple random variable with two possible values.

In understanding random experiments, we need to analyze the behavior of the randomness of the associated random variables.

- The problem is: we do not know exactly the value of p!
- Statistics can help in *estimating this parameter*.
- The foundation is that we assume the coin indeed follows such a probability distribution – *the form of the distribution is assumed beforehand, with* **only the parameter p being unknown**.

Discrete Random Variables and Probability Mass Function (PMF)

In *describing the randomness* of an **event**, we introduced probability.

- The chance of an event E happening is denoted $\mathbb{P}(E)$.

To describe the randomness of a discrete random variable X, we consider *the probability of X taking certain values*.

- Consider a special event
$$E_x = \text{``}X = x\text{''}.$$
- Here, E_x collects all outcomes ω such that $X(\omega) = x$, so E_x is indeed an event.
- With the definition of probability for events, we can say that the probability of X taking a value of x is
$$\mathbb{P}(X = x) = \mathbb{P}(E_x).$$

PMF

Such a function $p(x) = \mathbb{P}(X = x)$ is called a **probability mass function (PMF)** of the random variable X.

Example: Flip two coins

# of heads	0	1	2	Total
Probability	0.25	0.5	0.25	1

Uniform Random Variable

Uniform Random Variable

A random variable with equal probability for all outcomes are called a uniform random variable.

Example: Tossing a fair coin (2 possible outcomes); color of the card picked randomly from a deck (2 possible outcomes).

Bernoulli Distribution

In statistical studies, we usually have experiments that gives binary outcomes.

Example: A product passes/fails quality control.

Example: A coin toss gives head/tail.

These experiments are usually called Bernoulli trials or Bernoulli experiments.

Bernoulli experiment/trial

A **Bernoulli experiment** has

- One trial that can take two mutually exclusive results: "1" as success and "0" as failure.
- The probability of success is p.

Bernoulli Distribution

Bernoulli

A random variable X is said to be a **Bernoulli random variable** if

$$X = \begin{cases} 1 & \text{if success} \\ 0 & \text{if fail.} \end{cases}$$

A random variable X is said to follow a **Bernoulli distribution** with success probability p, if

$$\mathbb{P}\{X = x\} = \begin{cases} p & x = 1, \\ 1 - p & x = 0. \end{cases}$$

Expectation

Expectation for discrete random variable

The *expectation* (expected value, mean) of a random variable X is denoted by $\mathbb{E}[X]$. In the discrete case, the expectation is the average of all possible values, weighted by the probability. Mathematically, it is defined as

$$\mathbb{E}[X] = \sum_i x_i \, \mathbb{P}(X = x_i).$$

Example: let X be a $\mathrm{Bernoulli}(p)$[1] random variable. Then,

$$\mathbb{E}[X] = 1 \times \mathbb{P}(X = 1) + 0 \times \mathbb{P}(X = 0) = p.$$

[1] Abbreviation for "Bernoulli distribution with success probability p".

Linearity of Expectation

- For any constant a and b,

$$\mathbb{E}[aX + b] = aE[X] + b.$$

- For any constants $a_1, \ldots, a_n$ and b,

$$\mathbb{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n + b] = a_1\mathbb{E}[X_1] + a_2\mathbb{E}[X_2] + \cdots + a_n\mathbb{E}[X_n] + b.$$

Cf. Linearity of sample mean $\bar{x}$:

- $\overline{(ax + b)} = a\bar{x} + b.$
- $\overline{(a_1x_1 + a_2x_2 + \cdots + a_nx_n + b)} = a_1\bar{x}_1 + a_2\bar{x}_2 + \cdots + a_n\bar{x}_n + b.$

Variance

Variance

The *variance* of a random variable X, denoted by $\mathrm{Var}(X)$, is the expected value of the squared deviation from the mean of X, that is,

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2, \quad \text{where } \mu = \mathbb{E}[X].$$

Property

For any constants a and b,

$$\mathrm{Var}(aX+b) = a^2 \mathrm{Var}(X).$$

Cf. property of sample variance:

- If $y_i = a\,x_i + b$, $i = 1, 2, \ldots, n$, then $s_y^2 = a^2 s_x^2$.

Bernoulli Distribution

Example: Let X be a $\mathrm{Bernoulli}(p)$ random variable. Then

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = (1-p)^2 \cdot \mathbb{P}(X = 1) + (0-p)^2 \cdot \mathbb{P}(X = 0) = p(1-p).$$

Binomial Trial

Binomial Trial

A binomial experiment has the following characteristics:

- The experiment consists of a fixed number of observations n.
- Each trial is a Bernoulli trial with success probability p.
- The trials are independent, i.e the outcome of one trial does not impact the outcome on other trials.

- Suppose we have a Binomial trial with parameters n and p.
- How many successes are there in total?

Independence of Random Variables

Independence

Two random variables X and Y (not necessarily discrete) are **independent** if for *any two sets of real numbers* A *and* B

$$\mathbb{P}(\{X \in A\} \cap \{Y \in B\}) = \mathbb{P}(\{X \in A\})\,\mathbb{P}(\{Y \in B\}).$$

For discrete random variables, X and Y are independent if

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\,\mathbb{P}(Y = y_j).$$

$X_1, \dots, X_n$ are independent if

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \times \cdots \times \mathbb{P}(X_n = x_n).$$

Independence: Expectation and Variance Properties

Product of Independent Random Variables

If X and Y (not necessarily discrete) are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

Variance of a Sum of Independent Random Variables

For independent random variables $X_1, X_2, \ldots, X_n$, the variance of their sum is equal to the sum of their variances:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y),$$

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i).$$

Proofs for reference

$$
\begin{aligned}
\mathbb{E}[XY] &= \sum_i \sum_j x_i\, y_j\, \mathbb{P}(X = x_i, Y = y_j) \\
&= \sum_i \sum_j x_i\, y_j\, p_X(x_i)\, p_Y(y_j) \\
&= \left(\sum_i x_i\, p_X(x_i) \right) \left(\sum_j y_j\, p_Y(y_j) \right) \\
&= \mathbb{E}[X]\, \mathbb{E}[Y].
\end{aligned}
$$

Proofs for reference

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - \Big((\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2\Big) \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \Big((\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2\Big) \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2] - \Big((\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2\Big) \\
&= \Big(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\Big) + \Big(\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2\Big) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y).
\end{aligned}
$$

Binomial Distribution

Binomial distribution

A **Binomial random variable** X is the *total number of success* from n independent Bernoulli trials, each with success probability p.

$$X \sim \text{Binomial}(n, p)$$

The probability distribution is given by

$$\mathbb{P}\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, 2, \ldots, n$$

$$\binom{n}{i} = \frac{n!}{i! \times (n-i)!}$$

Binomial Distribution

Named after binomial expansion

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}.$$

Does the probabilities sum to 1? We have that

$$\sum_{i=0}^{n} \mathbb{P}\{X=i\} = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = [p + (1-p)]^n = 1.$$

Connection between Binomial and Bernoulli

Let $X \sim \text{Binomial}(n, p)$ and let $I_i \sim \text{Bernoulli}(p)$ be *independent* Bernoulli random variables with success rate p. Then

$$X = \sum_i I_i.$$

- **Ecpectation:**

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} I_i\right] = \sum_{i=1}^{n} \mathbb{E}[I_i] = \sum_{i=1}^{n} p = np.$$

- **Variance:**

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} I_i\right) = \sum_{i=1}^{n} \text{Var}(I_i) = np(1-p).$$

Binomial Distribution

Example: Suppose the probability that an item produced by a certain machine will be defective is 0.1, independent of other items. Find the prbability that a sample of 10 items will contain at most one defective item.

Solution:

Let X be the number of defects, then $X \sim \mathrm{Binomial}(10, 0.1)$.

So the probability is

$$\mathbb{P}\{X = 0\} + \mathbb{P}\{X = 1\} = \binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 = 0.7361$$

Application of Binomial Trial

When n is large and p is small, binomial trial can be use to model

- The number of misprints on a page (or a group of pages) of a book.
- The number of customers entering a post office on a given day.
- The number of patients in an emergency department in a day.
- The number of calls to a call center over a week.

Sounds good! But do we have problems in the calculation?

$$\mathbb{P}\{X = 0\} + \mathbb{P}\{X = 1\} = \binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 = 0.7361$$

Hard to calculate, numerically unstable. Approximations?

Example: Wire Flaws

Flaws occur at random along the length of a thin copper wire.

- X number of flaws in a unit length wire.
- λ rate at which flaw occur, so λ flaws per unit length.
- Expectation $\mathbb{E}[X] = \lambda \times 1$.

How to calculate the distribution of X?

- Partition the wire in to n sections, each with length $\Delta t = 1/n$
- Assume that at most one flaw may occur in each section.
- Assume that the flaws occur at random in each section, with probability p.
- Look familiar?

Example: Wire Flaws

The number of flaws can be approximated by a binomial random variable

$$X \approx \text{Binomial}(n, p),$$

where n is the number of sections and p is <u>chosen to match</u> the expectation:

$$\mathbb{E}[X] = \lambda = np.$$

Hence $p = \lambda/n$.

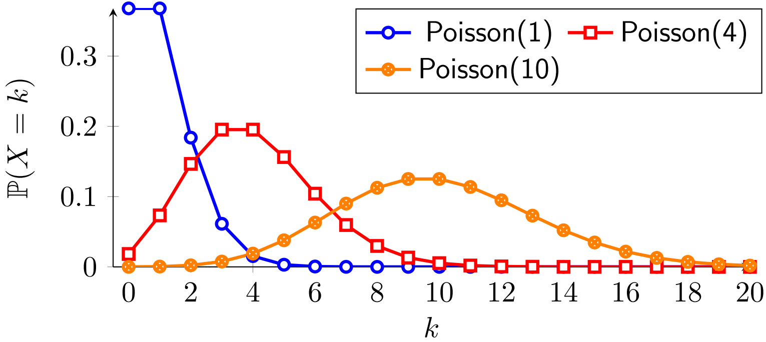
$$\begin{aligned}\mathbb{P}(X = i) \approx \binom{n}{i} p^i (1-p)^{n-i} &= \binom{n}{i} \left(\frac{\lambda}{n}\right)^i (1 - \lambda/n)^{n-i} \\ &= \frac{n!}{n^i (n-i)!} \frac{\lambda^i}{i!} (1 - \lambda/n)^{n-i} \\ &\to 1 \times \frac{\lambda^i}{i!} e^{-\lambda}\end{aligned}$$

Poisson distribution

Poisson distribution

A random variable X is $\text{Poisson}(\lambda)$ with $\lambda > 0$ if the PMF is

$$\mathbb{P}\{X = i\} = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \dots$$



Properties of Poisson RV

- Sum 1: (Taylor expansion of e^x)

$$e^\lambda = \sum_{i=0} \frac{\lambda^i}{i!}$$

- Mean:

$$\mathbb{E}[X] = \sum_{i=0} ie^{-\lambda}\frac{\lambda^i}{i!} = \sum_{i=1} e^{-\lambda}\frac{\lambda^i}{(i-1)!} = \lambda \sum_{i=0} e^{-\lambda}\frac{\lambda^i}{i!} = \lambda.$$

- Variance

$$\mathbb{E}[X^2] = \sum_{i=0} i^2 e^{-\lambda}\frac{\lambda^i}{i!} = \sum_{i=1} e^{-\lambda}(i-1+1)\frac{\lambda^i}{(i-1)!} = \lambda + \lambda^2.$$

Hence $\mathrm{Var}(X) = \lambda$.

Poisson approximates Binomial

When n is large and p is small,

$$\text{Poisson}(np) \approx \text{Binomial}(n, p).$$

Example: Recall the defective machine example.

$$\mathbb{P}\{X = 0\} + \mathbb{P}\{X = 1\} = 0.7361$$

If we want to use a Poisson random variable to approximate, we match the expectation:

$$\lambda = np = 10 \times 0.1 = 1.$$

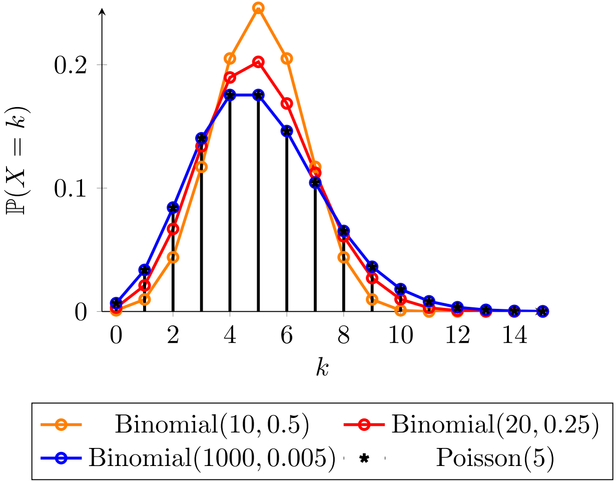
Try $Y \sim$ Poisson(1) to approximate it

$$\mathbb{P}\{Y = 0\} + \mathbb{P}\{Y = 1\} = e^{-1}\frac{1^0}{0!} + e^{-1}\frac{1^1}{1!} = 2e^{-1} = 0.7358$$

Poisson approximates Binomial

Poisson approximating Binomial

- Decent accuracy.
- Asymptotically the same.
- Much less computation needed.



Poisson: Statistical View

Poisson random variable can be use to model the number of patients visiting the Emergency Department (ED) every day.

- Knowing λ can help with the staffing decision.
- If we observe $X_1, \ldots, X_n$ patients in an ED over a period of n days, how can we estimate λ?

Intuitively (we will see), sample average and standard deviation are close to the expectation and the square root of variance.

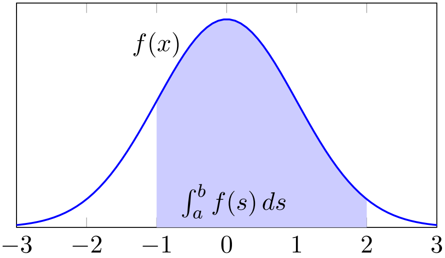
- Because $\mathbb{E}[X] = \lambda$, we can use the average $\hat{\lambda}_1 = \sum_{i=1}^{n} X_i/n$.
- Because $\mathrm{Var}[X] = \lambda$, we may also use $\hat{\lambda}_2 = S^2$, where S^2 is the sample variance.

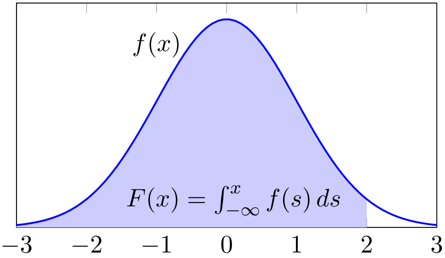
Which is better? $\Rightarrow$ Estimation theory.

Continuous Random Variables

Continuous Random Variable

When a random variable is a continuous variable, we call it a **continuous random variable**. Equivalently, the range is an interval of real numbers.





$$f(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} f(s)\,ds.$$

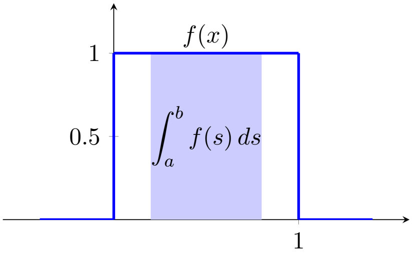
Uniform Distribution on [0,1]: CDF and PDF

Uniform Distribution on $[0, 1]$

A random variable X is said to follow a **Uniform** distribution on the interval $[0, 1]$ if it has an equal chance to take any value in $[0, 1]$.

The probability density function (PDF):

$$f(x) = F'(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$



The cumulative distribution function (CDF):

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & x > 1. \end{cases}$$

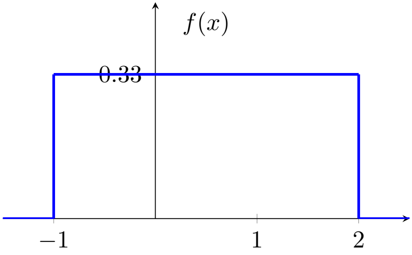
Uniform Distribution on $[a, b]$: CDF and PDF

Uniform Distribution on $[a, b]$

A random variable X is said to follow a **Uniform** distribution on the interval $[a, b]$ if it has an equal chance to take any value in $[a, b]$.

The probability density function (PDF):

$$f(x) = F'(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$



The cumulative distribution function (CDF):

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

Expectation and Variance (Continuous Case)

Expectation

The expectation of a continuous random variable X is given by:

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

If we know the distribution of X, then the expectation of a function $g(X)$ is

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx.$$

Variance

The variance is defined similarly as in the discrete case:

$$\sigma^2 = \mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2.$$

Example: $\text{Uniform}[a, b]$

Expectation

$$\mu = \mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \left.\frac{x^2}{2(b-a)}\right|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

Second moment

$$\mathbb{E}[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\, dx = \left.\frac{x^3}{3(b-a)}\right|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}.$$

Variance

$$\text{Var}(X) = \mathbb{E}[X^2] - \mu^2 = \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Generating Uniform R.V.

We assume that a $\mathrm{Uniform}[0,1]$ random variable can always be generated.

- Python:

```
import numpy as np
np.random()        # generate one Uniform[0,1]
np.random((3,2))   # a matrix of 3-by-2 Uniform[0,1]
```

- R: `runif(1)`
- Matlab: `rand()`

Why is Uniform R.V. Important?

The inverse of a distribution function $F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}$.

Inverse of CDF

- For any random variable X with continuous CDF $F_X(\cdot)$, $Y = F_X(X)$ is $\text{Uniform}[0,1]$.
- If Y is $\text{Uniform}[0,1]$, then for any CDF $F_X(\cdot)$, $X = F_X^{-1}(Y)$ has CDF $F_X(\cdot)$.

Proof: for continuous CDF,

$$\mathbb{P}(F_X(X) \leq x) = \mathbb{P}(X \leq F_X^{-1}(x)) = F_X(F_X^{-1}(x)) = x$$
$$\mathbb{P}(F_X^{-1}(Y) \leq x) = \mathbb{P}(Y \leq F_X(x)) = F(x).$$

To generate R.V. from any distribution, all you need is the inverse of the CDF and $\text{Uniform}[0,1]$.

The Galton Board

The **Galton board** is a vertical board with interleaved rows of pegs. Beads are dropped from the top and, when the board is level, each bead bounces either left or right as it hits a peg. Eventually, the beads are collected into bins at the bottom, and the height of the bead columns in the bins represents the frequency of outcomes.

- It serves as a physical demonstration of random processes, where each left/right bounce mimics an *independent Bernoulli trial*.
- As the number of rows increases, the distribution of beads in the bins tends to approximate a normal distribution, illustrating the *Central Limit Theorem*.
- The Galton board visually demonstrates **how randomness at the micro-level can lead to predictable, statistical behavior at the macro-level**.

`https://www.lexaloffle.com/bbs/?pid=114907`

Galton Board: Modeling a Bead's Path

Consider a bead on its way down the Galton board. After each row of pegs, the bead takes a step:

- It goes left one step, or
- It goes right one step.

Let X be the random variable representing the bead's move at a given row:

$$X = \begin{cases} -1, & \text{if the bead goes left,} \\ 1, & \text{if the bead goes right,} \end{cases}$$

with $\mathbb{P}(X = -1) = 0.5$ and $\mathbb{P}(X = 1) = 0.5$.

Galton Board: Modeling a Bead's Path

Denote by Y_i the position of the bead in the i-th row. The initial position is defined as $Y_0 = 0$. Then the position in the first row is $Y_1 = Y_0 + X_1 = X_1$, where X_1 is *an independent and identically distributed (i.i.d.) copy of* X.

As we continue, the position at row k is given by

$$Y_k = Y_{k-1} + X_k = \sum_{i=1}^{k} X_i.$$

By mathematical induction, the position at any row n is:

$$Y_n = \sum_{i=1}^{n} X_i.$$

The Galton board suggests that as the number of rows increases, the distribution of the bead's position Y_n tends to form a **bell curve** due to the Central Limit Theorem.

Central Tendency and Dispersion of Y_n

Recall that the bead's position after n rows is given by $Y_n = \sum_{i=1}^{n} X_i$.

Expectation

$$\mathbb{E}[Y_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\,\mathbb{E}[X] = n \times 0 = 0.$$

Variance

$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\,\mathrm{Var}(X) = n \times 1 = n.$$

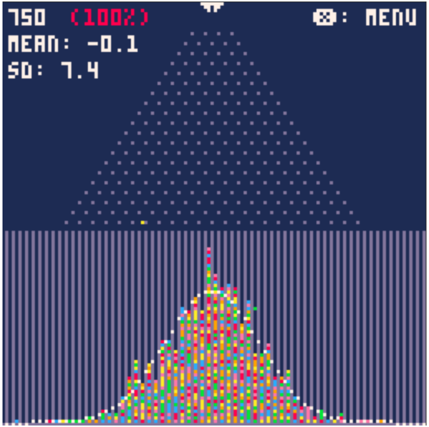
Galton Board (12 Rows): Central Tendency and Dispersion

There are 25 rows in this Galton board.

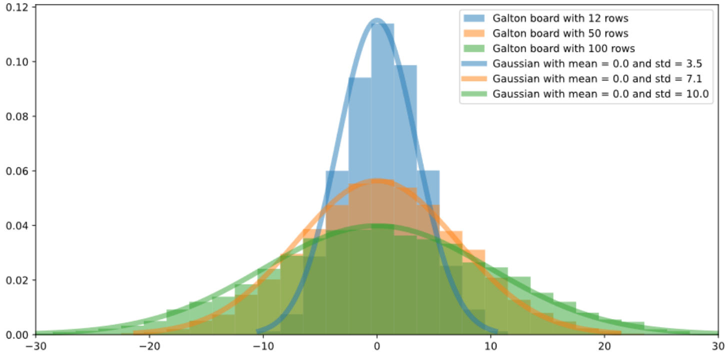
- $\mu = \mathbb{E}[Y_{25}] = 0$.
- $\sigma = \sqrt{\operatorname{Var}(Y_{25})} = \sqrt{25} = 5$.

The Galton board suggests that:

- Roughly 68% of the beads end up in the bins in positions $\mu \pm \sigma$.
- Roughly 95% of the beads end up in the bins in positions $\mu \pm 2\sigma$.
- Almost all beads end up in the bins in positions $\mu \pm 3\sigma$.

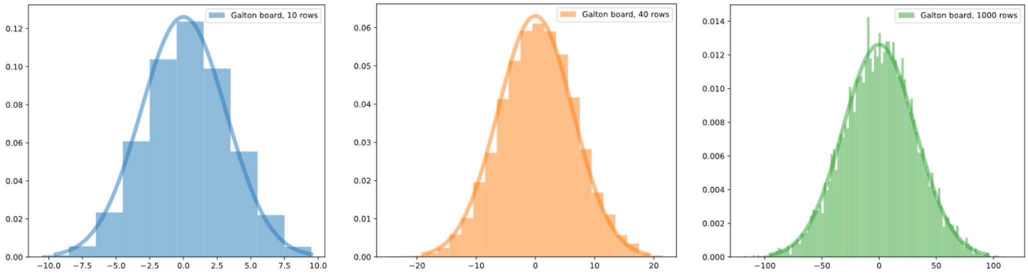


Increasing the rows in the Galton board



- As the number of rows n increases, the distribution of location of the beads spreads wider and wider: $\sigma = \sqrt{\mathrm{Var}(Y_n)} = \sqrt{n}$.
- The shape of the distribution becomes more and more like a bell curve.

Increasing the rows in the Galton board



Observation: When plotting a histogram (e.g., using Python libraries such as Matplotlib), the plotter automatically scales the canvas so that the histogram **fills most of the available space**. But why is this done?

- **Maximizing Visual Information:** Scaling ensures that the details of the data distribution (such as peaks, valleys, and spread) are clearly visible.

Question: What is the "natural scale" for the plot as function of n?

$$3 * \sqrt{10} \approx 10, 3 * \sqrt{40} \approx 20, 3 * \sqrt{1000} \approx 100.$$

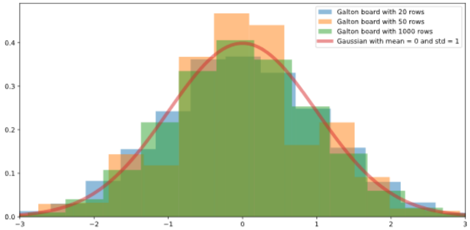
Standardizing the Beads' Positions

To implement the scaling logic of the plotter, we can **standardize** the positions of the beads:

$$Z_n = \frac{Y_n - \mathbb{E}[Y_n]}{\sqrt{\mathrm{Var}(Y_n)}} = \frac{Y_n - 0}{\sqrt{n}}.$$

The standardized positions Z_n will be centered around 0 and have a "standard" standard deviation of 1.

It turns out that regardless of the number of rows, the standardized positions follow the same **bell curve**.



Central Limit Theorem and the Galton Board

The math behind the Galton board can be rigorously justified, yielding the famous

Central Limit Theorem (CLT) for summation

Recall that $Y_n = \sum_{i=1}^n X_i$, we have:

$$Z_n = \frac{Y_n - \mathbb{E}[Y_n]}{\sqrt{\mathrm{Var}(Y_n)}} \approx Z, \quad \text{as } n \text{ becomes large.}$$

Here, Z is a continuous random variable following the "bell curve" distribution, called the **standard normal distribution**, denoted by $\mathcal{N}(0,1)$.

- The CLT explains why the distribution of bead positions tends to become bell-shaped as the number of rows increases.
- More generally, the CLT states that sums of many independent and identically distributed (i.i.d.) random variables, when standardized, converge to $\mathcal{N}(0,1)$.

Sample Mean and Standardized Position

Starting with

$$Z_n = \frac{Y_n - \mathbb{E}[Y_n]}{\sqrt{\mathrm{Var}(Y_n)}} \quad \text{where} \quad Y_n = \sum_{i=1}^{n} X_i,$$

we note that $\mathbb{E}[Y_n] = n\,\mathbb{E}[X_1]$ and $\mathrm{Var}(Y_n) = n\,\mathrm{Var}(X_1)$. Hence,

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\,\mathbb{E}[X_1]}{\sqrt{n\,\mathrm{Var}(X_1)}} = \frac{\overline{X} - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)/n}}, \quad \text{where} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

CLT for sample mean

$$\overline{X} = \mathbb{E}[X_1] + \sqrt{\frac{\mathrm{Var}(X_1)}{n}}\, Z_n \;\approx\; \mathbb{E}[X_1] + \sqrt{\frac{\mathrm{Var}(X_1)}{n}}\, Z, \qquad \text{for large } n.$$

- The sample mean $\overline{X}$ is *centered* around the population mean $\mathbb{E}[X_1]$.
- The *scale* of its fluctuations is about $\sqrt{\mathrm{Var}(X_1)/n}$ – *decreases* as n increases.

Central Limit Theorem (CLT) Works for Any Distribution

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables **with finite variance**. Then

$$Z_n = \frac{\bar{X} - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)/n}} \approx Z \sim \mathcal{N}(0,1), \quad \text{where} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This result implies that

$$\bar{X} \approx \mathbb{E}[X_1] + \sqrt{\frac{\mathrm{Var}(X_1)}{n}}\, Z,$$

$$\sum_{i=1}^{n} X_i \approx n\,\mathbb{E}[X_1] + \sqrt{n\,\mathrm{Var}(X_1)}\, Z.$$

Normal Approximation for a Binomial Random Variable

Suppose Y_n is a binomial random variable with parameters n and p, then Y_n can be approximated by a normal distribution **when n is large** (works well when $np \geq 5$ and $n(1-p) \geq 5$).

Solution:

$$Y_n = \sum_{i=1}^{n} I_i, \quad \text{where each } I_i \sim \mathrm{Bernoulli}(p).$$

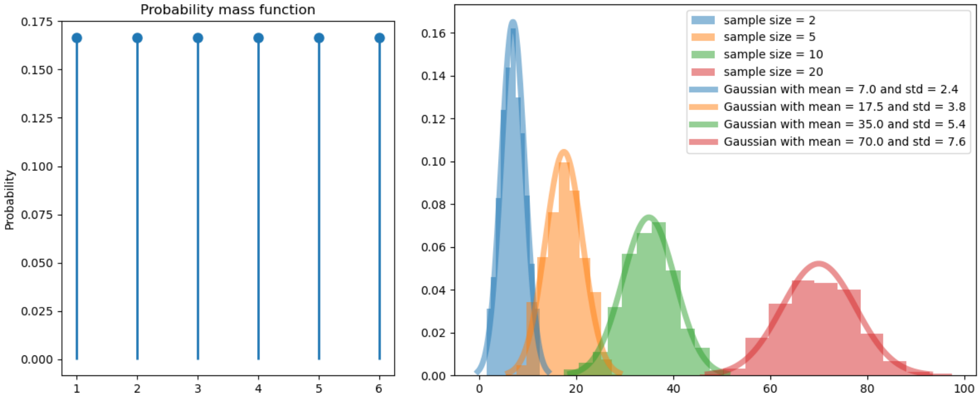
Note that

$$\mathbb{E}[I_i] = p \quad \text{and} \quad \mathrm{Var}(I_i) = p(1-p).$$

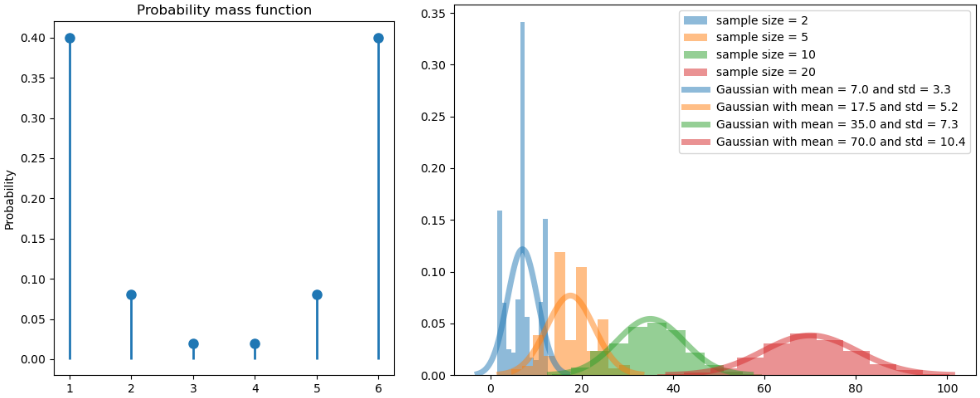
By the Central Limit Theorem,

$$Y_n \approx n\,\mathbb{E}[I_1] + \sqrt{n\,\mathrm{Var}(I_1)}\,Z = np + \sqrt{np(1-p)}\,Z, \quad \text{where } Z \sim \mathcal{N}(0,1).$$

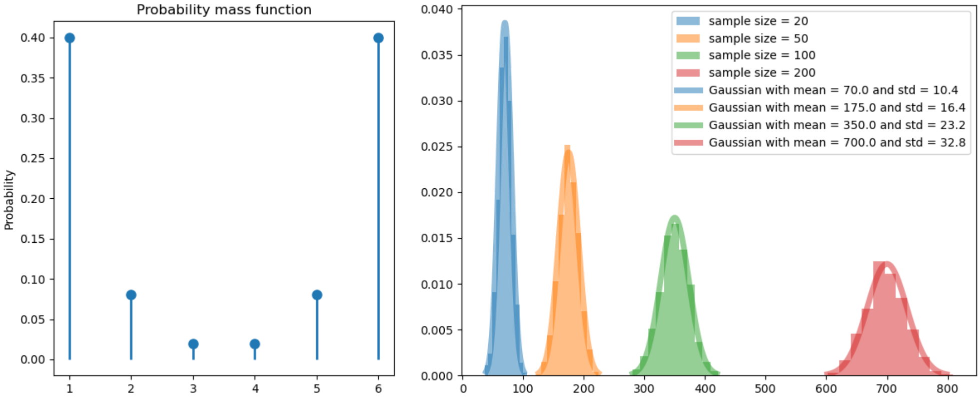
Quality of the Normal Approximation



Quality of the Normal Approximation



Quality of the Normal Approximation



Quality of the Normal Approximation

The quality of the normal approximation via the Central Limit Theorem (CLT) varies:

- If the underlying distribution is normal, the approximation is exact.
- If the underlying distribution is *skewed or have large variance*, the approximation may be poor for small sample sizes.
- The quality of the approximation improves as the sample size increases.
- As a rule of thumb, if the distribution is not too skewed and the variance is moderate, a sample size of $n \geq 30$ should provide a reasonably accurate approximation.

Further Reading: 3Blue1Brown Videos

For a deeper understanding of the Central Limit Theorem and related topics, consider watching these insightful videos by `@3Blue1Brown`:

- "But what is the Central Limit Theorem?"
- "A pretty reason why Gaussian + Gaussian = Gaussian"

Normal

Having seen the universality of the standard normal distribution, ensured by the central limit theorem, we can now focus more on "The Bell Curve" itself.

Normal Distribution

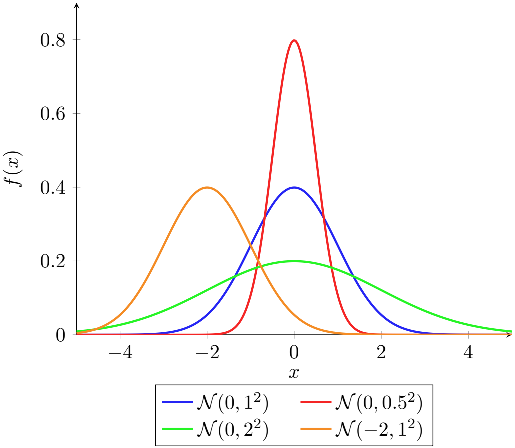
A random variable is said to be **normally distributed** with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if the PDF is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

$$\mathbb{E}[X] = \mu, \quad \mathbb{E}[X^2] = \sigma^2 + \mu^2, \quad \mathrm{Var}(X) = \sigma^2.$$

Density of Normal Distribution

- The normal distribution is **symmetric** around its mean μ.
- The density function is **unimodal**, with the peak at μ.
- The spread of the distribution is controlled by the standard deviation σ – most density (99.7%) lies within $\mu \pm 3\sigma$.



Scalability of the Normal Distribution

Scaling Property

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, then

$$Y \sim \mathcal{N}(a\mu + b,\, a^2\sigma^2).$$

Now, if we set

$$a = \frac{1}{\sigma} \quad \text{and} \quad b = -\frac{\mu}{\sigma},$$

we obtain

$$Y = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

This transformation is known as the **standardization** of a normal distribution.

We can always write $X = \sigma Z + \mu$, where $Z \sim \mathcal{N}(0, 1)$.

We call Z the **standard normal distribution**.

Standard Normal Distribution

The **standard normal distribution**, denoted by $Z \sim \mathcal{N}(0, 1)$, satisfies

Probability Density Function (PDF) of standard normal

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in (-\infty, \infty).$$

Cumulative Distribution Function (CDF) of standard normal

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy, \quad x \in (-\infty, \infty).$$

Note: Although $\Phi(x)$ has no closed-form expression, numerical values are widely available.

Properties of the Standard Normal Distribution

Complement Rule

$$\mathbb{P}(Z > x) = 1 - \mathbb{P}(Z \leq x) = 1 - \Phi(x).$$

Symmetry Property

$$\mathbb{P}(Z < -x) = \Phi(-x).$$

Since $\mathbb{P}(Z > x) = \mathbb{P}(Z < -x)$, it follows that:

$$\Phi(-x) = 1 - \Phi(x).$$

Example: Evaluating Probabilities for a Normal Distribution

For any $X \sim \mathcal{N}(\mu, \sigma^2)$, the standardization

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

implies

$$\mathbb{P}(X < b) = \mathbb{P}\Big(Z < \frac{b - \mu}{\sigma}\Big) = \Phi\Big(\frac{b - \mu}{\sigma}\Big)$$

and

$$\mathbb{P}(a < X < b) = \Phi\Big(\frac{b - \mu}{\sigma}\Big) - \Phi\Big(\frac{a - \mu}{\sigma}\Big).$$

Evaluating Probabilities for a Normal Distribution

Example: Suppose $X \sim \mathcal{N}(3, 16)$

(a) Find $P\{X < 11\}$:

$$P(X < 11) = \Phi\left(\frac{11-3}{4}\right) = \Phi(2).$$

Using standard normal tables, $\Phi(2) \approx 0.9772$.

(b) Find $P\{X > -1\}$:

$$P(X > -1) = 1 - P(X \leq -1) = 1 - \Phi\left(\frac{-1-3}{4}\right) = 1 - \Phi(-1).$$

Since $\Phi(-1) = 1 - \Phi(1)$ and $\Phi(1) \approx 0.8413$, it follows that

$$P(X > -1) = \Phi(1) \approx 0.8413.$$

(c) Show that $P\{2 < X < 7\} \approx 0.44$.

Approximating Probability for the Sample Mean

By the scalability of the normal distribution,

$$\bar{X} = \mu + \sqrt{\sigma^2/n} \times Z, \quad \text{for } Z \sim \mathcal{N}(0, 1).$$

For general distributions of the population, the CLT suggests that

$$\bar{X} \approx \mu + \sqrt{\sigma^2/n} \times Z, \quad \text{for } Z \sim \mathcal{N}(0, 1).$$

We may use the probability density function of the standard normal to approximate probabilities regarding the sample mean.

Approximating Probability for the Sample Mean

Example: The weights of a population of workers have mean 167 and standard deviation 27. Suppose we want to approximate the probability that the sample mean of their weights lies between 163 and 171.

Case A: Sample Size $n_1 = 36$

Note that

$$\sqrt{\sigma^2/n_1} = \sqrt{27^2/36} = 4.5.$$

Hence,

$$\begin{aligned} P(163 < \bar{X} < 171) &= P\left(\frac{163-167}{4.5} < \frac{\bar{X}-167}{4.5} < \frac{171-167}{4.5}\right) \\ &= P(-0.89 < Z < 0.89) \\ &= 2P(Z < 0.89) - 1 \approx 0.626. \end{aligned}$$

Approximating Probability for the Sample Mean

Case B: Sample Size $n_2 = 144$

Note that

$$\sqrt{\sigma^2/n_2} = \sqrt{27^2/144} = 2.25.$$

Thus,

$$\begin{aligned} P(163 < \bar{X} < 171) &= P\Big(\frac{163-167}{2.25} < \frac{\bar{X}-167}{2.25} < \frac{171-167}{2.25}\Big) \\ &= P(-1.78 < Z < 1.78) \\ &= 2P(Z < 1.78) - 1 \approx 0.925. \end{aligned}$$

Sum of Independent Normal is Normal

Sum of Independent Normal Random Variables

Let $X_1, X_2, \ldots, X_n$ be independent randon variables distributed as $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of these random variables. Then,

$$S_n \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Key Points:

- The sum of independent normal random variables is also normal.
- The mean of the sum is the sum of the individual means.
- The variance of the sum is the sum of the individual variances.

Sum of i.i.d. Normal is Normal

Sum of i.i.d. Normal Random Variables

Let $X_1, X_2, \ldots, X_n$ be independent and **identically** distributed as $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of these random variables. Then,

$$S_n \sim N\Big(n\mu,\ n\sigma^2\Big).$$

Extended Reading and Exercises

- Sections 3.1-3.5, 3.8, 4.1-4.6, 7.2 of **Douglas C. Montgomery and George C. Runger, Applied Statistics and Probability for Engineers, 7th Ed.**
- Videos by `@3Blue1Brown`:
    - "But what is the Central Limit Theorem?"
    - "A pretty reason why Gaussian + Gaussian = Gaussian"