Vectors:

A vector is a tuple of one or more values called scalars.

Vectors are built from components, which are ordinary numbers. You can think of a vector as a list of numbers, and vector algebra as operations performed on the numbers in the list.

Vectors are often represented using a lowercase character such as “v”; for example:

$$v = (v1, v2, v3)$$

Where v1, v2, v3 are scalar values, often real values.

Where v1, v2, v3 are scalar values, often real values.

Vectors are also shown using a vertical representation or a column; for example:

$$v = \begin{pmatrix} v1 \\ v2 \\ v3 \end{pmatrix}$$

It is common to introduce vectors using a geometric analogy, where a vector represents a

point or coordinate in an n-dimensional space, where n is the number of dimensions, such as 2.

The vector can also be thought of as a line from the origin of the vector space with a direction and a magnitude.

Defining a Vector:

We can represent a vector in Python as a NumPy array.

A NumPy array can be created from a list of numbers. For example, below we define a vector with the length of 3 and the integer values 1, 2 and 3.

```
from numpy import array
v = array([1, 2, 3])
print(v)
```

Vector Arithmetic

In this section will demonstrate simple vector-vector arithmetic, where all operations are performed element-wise between two vectors of equal length to result in a new vector with the same length

Vector Addition

Two vectors of equal length can be added together to create a new third vector.

$$c = a + b$$

The new vector has the same length as the other two vectors. Each element of the new vector is calculated as the addition of the elements of the other vectors at the same index; for example:

$$a + b = (a1 + b1, a2 + b2, a3 + b3)$$

Or, put another way:

$$c[0] = a[0] + b[0]$$
$$c[1] = a[1] + b[1$$
$$c[2] = a[2] + b[2]$$

We can add vectors directly in Python by adding NumPy arrays.

```
# add vectors
from numpy import array
a = array([1, 2, 3])
print(a)
b = array([1, 2, 3])
```

```
print(b)
c = a + b
print(c)
```

The example defines two vectors with three elements each, then adds them together.

Vector Subtraction

One vector can be subtracted from another vector of equal length to create a new third vector.

```
c = a – b
```

As with addition, the new vector has the same length as the parent vectors and each element of the new vector is calculated as the subtraction of the elements at the same indices.

```
1 a - b = (a1 - b1, a2 - b2, a3 - b3)
```

Or, put another way:

```
c[0] = a[0] - b[0]
c[1] = a[1] - b[1
c[2] = a[2] - b[2]
```

The NumPy arrays can be directly subtracted in Python.

```
# subtract vectors
from numpy import array
a = array([1, 2, 3])
print(a)
b = array([0.5, 0.5, 0.5])
print(b)
c = a - b
print(c)
```

Vector Multiplication

Two vectors of equal length can be multiplied together.

```
c = a * b
```

As with addition and subtraction, this operation is performed element-wise to result in a new vector of the same length.

```
a * b = (a1 * b1, a2 * b2, a3 * b3)
```

Or

```
ab = (a1b1, a2b2, a3b3)
```

Or, put another way:

```
1 c[0] = a[0] * b[0]
2 c[1] = a[1] * b[1]
3 c[2] = a[2] * b[2]
```

We can perform this operation directly in NumPy.

```
# multiply vectors
from numpy import array
a = array([1, 2, 3])
print(a)
b = array([1, 2, 3])
print(b)
c = a * b
print(c)
```

Vector Division

Two vectors of equal length can be divided.

```
c = a / b
```

As with other arithmetic operations, this operation is performed element-wise to result in a new vector of the same length.

```
a / b = (a1 / b1, a2 / b2, a3 / b3)
```

or

```
a / b = (a1b1, a2b2, a3b3)
```

Or, put another way:

```
c[0] = a[0] / b[0]
c[1] = a[1] / b[1]
c[2] = a[2] / b[2]
```

We can perform this operation directly in NumPy.

```
# divide vectors
from numpy import array
a = array([1, 2, 3])
print(a)
b = array([1, 2, 3])
print(b)
c = a / b
print(c)
```

Vector Dot Product

We can calculate the sum of the multiplied elements of two vectors of the same length to give a scalar.

This is called the dot product, named because of the dot operator used when describing the operation.

The dot product is the key tool for calculating vector projections, vector decompositions, and

determining orthogonality. The name dot product comes from the symbol used to denote it.

c = a . b

The operation can be used in machine learning to calculate the weighted sum of a vector.

The dot product is calculated as follows:

a . b = (a1 * b1 + a2 * b2 + a3 * b3)

Or

a . b = (a1b1 + a2b2 + a3b3)

We can calculate the dot product between two vectors in Python using the dot() function on a NumPy array.

```
# dot product vectors
from numpy import array
a = array([1, 2, 3])
print(a)
b = array([1, 2, 3])
print(b)
c = a.dot(b)
print(c)
```

Vector-Scalar Multiplication

A vector can be multiplied by a scalar, in effect scaling the magnitude of the vector.

To keep notation simple, we will use lowercase "s" to represent the scalar value.

```
c = s * v
```

The multiplication is performed on each element of the vector to result in a new scaled vector of the same length.

```
s * v = (s * v1, s * v2, s * v3)
```

Or, put another way:

```
c[0] = a[0] * s
c[1] = a[1] * s
c[2] = a[2] * s
```

We can perform this operation directly with the NumPy array.

```
# vector-scalar multiplication
from numpy import array
a = array([1, 2, 3])
print(a)
s = 0.5
print(s)
c = s * a
print(c)
```

Matrices play a central role in data science: they are probably the most common way of representing data to be analyzed and manipulated by virtually any machine learning or analytics algorithm. However, it is also important to understand that there really two uses to matrices within data science:

1. Matrices are the “obvious” way to store tabular data
2. Matrices are the foundation of linear algebra

matrices are essentially 2D arrays, this view is fundamentally a take on how to efficiently store data in the multi-dimensional arrays. But matrices are also the basic unit of linear algebra, which is a mathematical language for the expression and manipulation of linear systems of equations. There are naturally overlaps between the two, but the core operations of linear algebra, such as matrix multiplication and solving linear systems of equations, are largely orthogonal to the way in which matrices are stored as arrays in memory.

Matrices are 2D arrays of values, and we use the notation $A \in R^{m \times n}$ to denote a matrix with m rows and n columns. We can write out the elements explicitly as follows

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}$$

where Aij denotes the entry of A in the ith row and jth column. We will use the notation Ai: to refer to the ith row of A and the A:j to refer to the jth column of A

Matrices for tabular data and row/column ordering

Let's start with a simple example representing tabular data using matrices, one of the more natural ways to represent such data. Let's consider the "Grades" table that we previously discussed in our presentation of relational data:

Person ID	HW1 Grade	HW2 Grade
5	85	95
6	80	60

Person ID	HW1 Grade	HW2 Grade
100	100	100

$$A \in \mathbb{R}^{3\times 2} = \begin{bmatrix} 85 & 95 \\ 80 & 60 \\ 100 & 100 \end{bmatrix}$$

Since data in memory is laid out sequentially (at least logically as far as programs are concerned, if not physically on the chip) we can opt to store the data in *row major order*, that is, storing each row sequentially

(85,95,80,60,100,100)

or in *column major order*, storing each column sequentially

(85,80,100,95,60,100)

Basics of linear algebra

In addition to serving as a method for storing tabular data, vector and matrices also provide a method for studying sets of linear equations.

Consider the following two linear equations in two variables x1 and x2.

$$4x_1-5x_2=-13$$

$$-2x_1+3x_2=9$$

This can written compactly as the equation Ax=b, where

$$A \in \mathbb{R}^{2\times 2} = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \; b \in \mathbb{R}^2 = \begin{bmatrix} -13 \\ 9 \end{bmatrix}, \; x \in \mathbb{R}^2 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Basic operations and special matrices

Addition and substraction: Matrix addition and subtraction are applied elementwise over the matrices, and can only apply two two matrices of the same size. That is,
if $A \in R^{m\times n}$ and $B \in R^{m\times n}$ then their sum/difference $C=A+B$ is another matrix of the same size $C \in R^{m\times n}$ where

$$C_{ij}=A_{ij}+B_{ij}.$$

Transpose: Transposing a matrix flips its rows and columns. That is, if $A \in R^n$, then it's transpose, denoted $C=A^T$ is a matrix $C \in R^{n\times m}$ where

$$C_{ij}=A_{ji}.$$

Matrix multiplication: Matrix multiplication is a bit more involved. Unlike addition and subtraction, matrix multiplication does not perform elementwise multiplication of the two matrices. Instead, for a matrix $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$ (note these precise sizes, as they are important), their product $C = AB$ is a matrix $C \in \mathbb{R}^{m \times p}$ where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

- Matrix multiplication is associative: $(AB)C = A(BC)$ (i.e., it doesn't matter in what order you do the multiplications, though it *can* matter from a computational perspective, as some orderings will be more efficient to compute than others)
- Matrix multiplication is distributive: $A(B+C) = AB + AC$
- Matrix multiplication is *not* commutative: $AB \neq BA$. This is really true in two different ways. Under the above matrix sizes, the multiplication BA is not a valid expression if $m \neq p$ (since the number of columns in B would not match the number of rows in A). And even if the

dimensions *do* match (for instance if all the matrices were n×n) the products will still not be equal in general.

Identity matrix: The identity matrix $I \in \mathbb{R}^{n \times n}$ is a square matrix with ones on the diagonal an zeros everywhere else

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

It has the property that for any matrix $A \in \mathbb{R}^{m \times n}$

$$AI = IA = A$$

Matix inverse: For a square matrix $A \in \mathbb{R}^{n \times n}$, the matrix inverse $A^{-1} \in \mathbb{R}^{n \times n}$ is the unique matrix such that

$$A^{-1}A = AA^{-1} = I.$$

The matrix inverse need not exist for all square matrices (it will depend on the linear independence between rows/columns of A

Solving linear equations: The matrix inverse provides an immediate method to obtain the solution to systems of linear equations. Recall out

example above of a set of linear equations $Ax=b$. If we want to find the x that satisfies this equation, we multiply both sizes of the equation by A^{-1} on the left to get

$$A^{-1}Ax=A^{-1}b \Longrightarrow x=A^{-1}b.$$

The nice thing here is that as far as we are concerned in this course, the set of equations is now *solved*.

Transpose of matrix product: It follows immediately from the definition of matrix multiplication and the transpose that

$$(AB)^T=B^TA^T$$

i.e., the transpose of a matrix product is the product of the transposes, in reverse order.

Inverse of matrix: It also follows immediately from the definitions that for $A,B \in Rn\times n$ both square

$$(AB)^{-1}=B^{-1}A^{-1}$$

i.e. the inverse of a matrix product is the product of the inverses, in reverse order.

Inner products: One type of matrix multiplication is common enough that it deserves special mention. If x,y∈Rn are vectors of the same dimensions, then

$$x^T y = \sum_{i=1}^{n} x_i y_i$$

(the matrix product of x transposed, i.e., a row vector and y, a column vector) is a *scalar* quantity called the inner product of x and y; it is simply equal to the sum of the corresponding elements of x and y multiplied together.

Statistics

Central Tendency

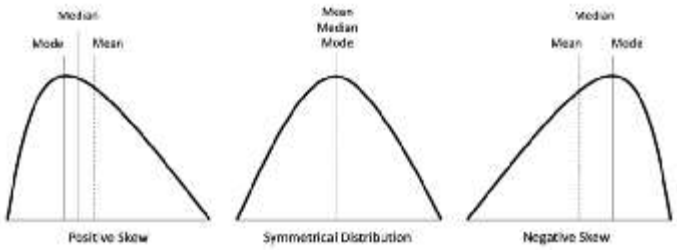
Mean: The average of the dataset.

Median: The middle value of an ordered dataset.

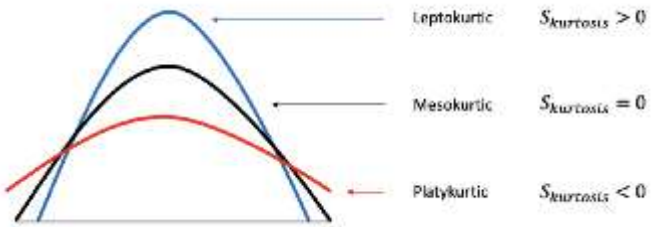
Mode: The most frequent value in the dataset. If the data have multiple values that occurred the most frequently, we have a multimodal distribution.

Skewness: A measure of symmetry.

Kurtosis: A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution



Skewness.



Kurtosis.

Variability

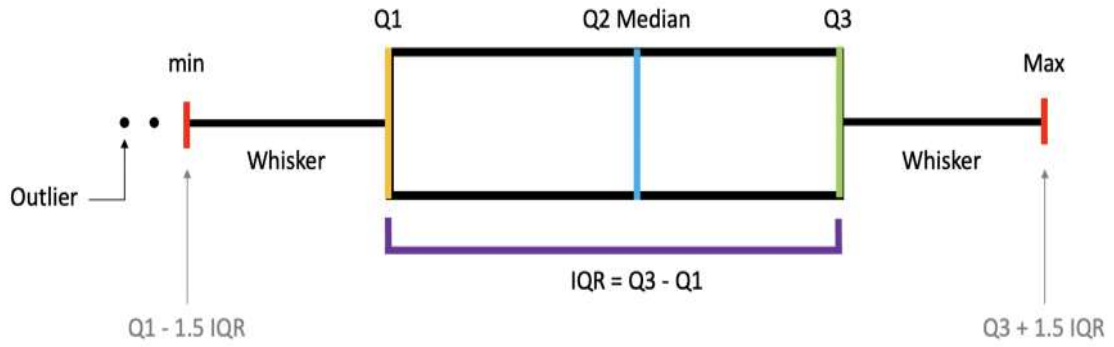
Range: The difference between the highest and lowest value in the dataset.

Percentiles, Quartiles and Interquartile Range (IQR)

- **Percentiles** — A measure that indicates the value below which a given percentage of observations in a group of observations falls.

- **Quantiles**— Values that divide the number of data points into four more or less equal parts, or quarters.

- **Interquartile Range (IQR)**— A measure of statistical dispersion and variability based on dividing a data set into quartiles. IQR = Q3 − Q1



Percentiles, Quartiles and Interquartile Range (IQR).

Variance: The average squared difference of the values from the mean to measure how spread out a set of data is relative to mean.

Standard Deviation: The standard difference between each data point and the mean and the square root of variance.

	Variance	Standard Deviation
Population	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
Sample	$s^2 = \frac{\sum(x_i - \bar{x})^2}{N - 1}$	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}}$

Population and Sample Variance and Standard Deviation.

Standard Error (**SE**): An estimate of the standard deviation of the sampling distribution.

	Standard Error	Estimate
Population	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$
Sample	$s_{\bar{x}} = \frac{s}{\sqrt{n}}$	

Population and Sample Standard Error.

Relationship Between Variables (Very Important Topic)

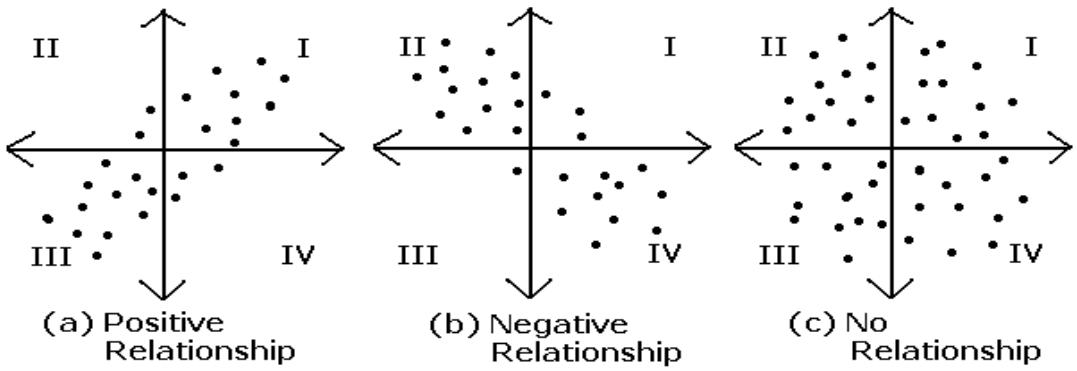
Causality: Relationship between two events where one event is affected by the other.

Covariance: A quantitative measure of the joint variability between two or more variables.

Correlation: Measure the relationship between two variables and ranges from *-1 to 1*, the normalized version of covariance.

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$cor(x,y) = \frac{cov(x,y)}{\sqrt{var(x)\,var(y)}}$$



Covariance and Correlation

Simpson's Paradox: (Important Topic)

Simpson's paradox, is a paradox in probability and statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined.

One of the bestknown examples of Simpson's paradox is a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

But when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women. In fact, the pooled and corrected data showed a "small but statistically significant bias in favor of women." The data from the six largest departments is listed below.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Probability

Data and probability are inseperable. Data is computational side of story whereas probability is the theoretical side of the story

Probability needs to address the practical problems. Probability address the prior knownledge of data and how to compute the confidence interval of an estimate.

Probability is a measure of size of a set. Probability telling us about the relative frequency of the data.

Example:

Probability of getting even number when rolling a die.

probability space:

- **Sample Space** Ω: The set of all possible outcomes from an experiment.
- **Event Space** $\mathcal{F}$: The collection of all possible events. An event E is a subset in Ω that defines an outcome or a combination of outcomes.
- **Probability Law** $\mathbb{P}$: A mapping from an event E to a number $\mathbb{P}[E]$ which, ideally, should measure the size of the event.

It is the collection of all possible states that can be drawn from an experiment. The probability law is the interface with the data analysis

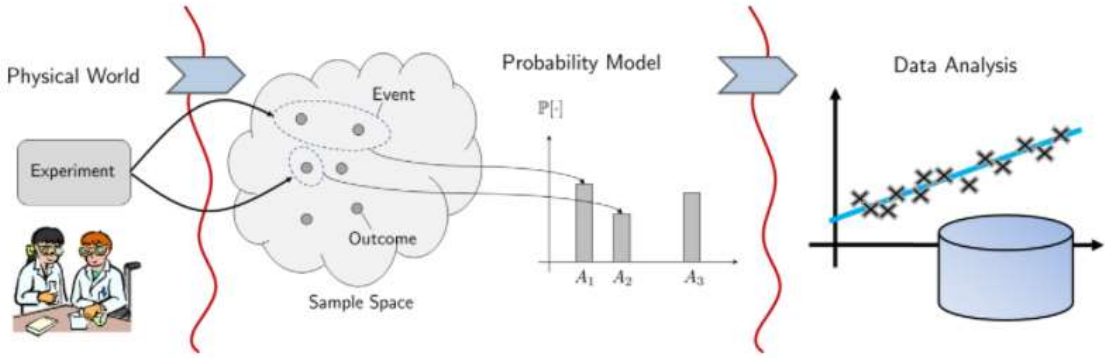


Figure 2.12: Given an experiment, we define the collection of all outcomes as the sample space. A subset in the sample space is called an event. The probability law is a mapping that maps an event to a number. This number denotes the size of the event.

Sample Space: Set of all possible outcomes from an experiment is a Sample space. A Sample space can contain discrete or continuous outcomes.

Example 1: (Discrete Outcomes)

- Coin flip: $\Omega = \{H,\ T\}$.
- Throw a dice: $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.
- Paper / scissor / stone: $\Omega = \{\text{paper}, \text{scissor}, \text{stone}\}$.
- Draw an even integer: $\Omega = \{2, 4, 6, 8, \ldots\}$.

In the last example, we see that a sample space can be infinite.

Example 2: (Continuous Outcomes)

- Waiting time for a bus in West Lafayette: $\Omega = \{t \mid 0 \leq t \leq 30 \text{ minutes}\}$.
- Phase angle of a voltage: $\Omega = \{\theta \mid 0 \leq \theta \leq 2\pi\}$.
- Frequency of a pitch: $\Omega = \{f \mid 0 \leq f \leq f_{\max}\}$.

Practice Exercise. There are 8 processors on a computer. A computer job scheduler chooses one processor randomly. What is the sample space? If the computer job schedular can choose two processors at once, what is the sample space then?

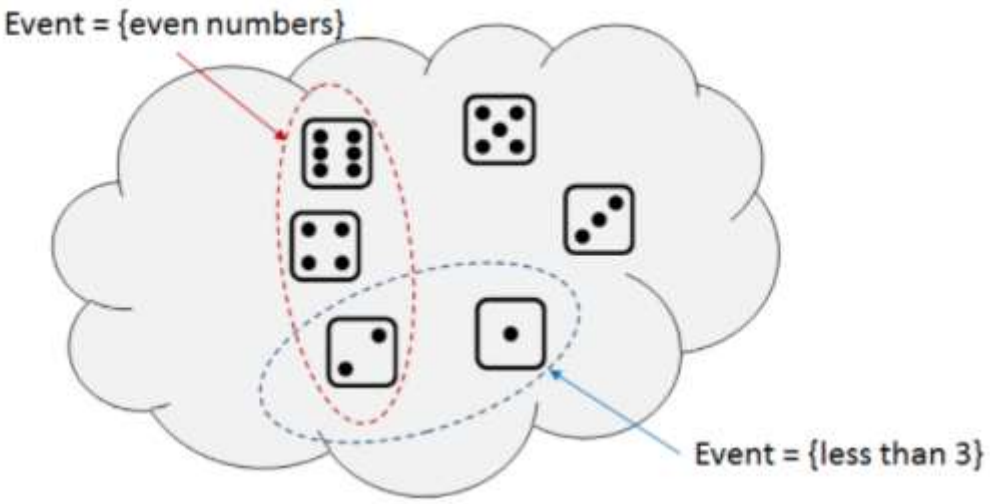
Solution. The sample space of the first case is $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The sample space of the second case is $\Omega = \{(1, 2), (1, 3), (1, 4), \ldots, (7, 8)\}$.

All the outcomes are Exhaustive and equally

Definition 2.14. *An **event** E is a subset in the sample space Ω. The set of all possible events is denoted as $\mathcal{F}$.*

Example 1. Throw a dice. Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. The followings are two possible events, as illustrated in Figure 2.14.

- $E_1 = \{\text{even numbers}\} = \{\boxdot, \boxdot, \boxdot\}$.
- $E_2 = \{\text{less than } 3\} = \{\boxdot, \boxdot\}$.



Definition 2.18. *A **probability law** is a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *that maps an event* E *to a real number in* $[0, 1]$.

Example 1. Consider flipping a coin. The event space $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. We can define the probability law as

$$\mathbb{P}[\emptyset] = 0, \;\; \mathbb{P}[\{H\}] = \frac{1}{2}, \;\; \mathbb{P}[\{T\}] = \frac{1}{2}, \;\; \mathbb{P}[\Omega] = 1,$$

as shown in Figure 2.16. This $\mathbb{P}$ is clearly consistent for all the events in $\mathcal{F}$.

Is it possible to construct an invalid $\mathbb{P}$? Certainly. Consider the following probability law:

$$\mathbb{P}[\emptyset] = 0, \;\; \mathbb{P}[\{H\}] = \frac{1}{3}, \;\; \mathbb{P}[\{T\}] = \frac{1}{3}, \;\; \mathbb{P}[\Omega] = 1.$$

This law is invalid because the individual events $\mathbb{P}[\{H\}] = \frac{1}{3}$ and $\mathbb{P}[\{T\}] = \frac{1}{3}$ but the union $\mathbb{P}[\Omega] = 1$. To fix this problem, one possible solution is to define the probability law as

$$\mathbb{P}[\emptyset] = 0, \;\; \mathbb{P}[\{H\}] = \frac{1}{3}, \;\; \mathbb{P}[\{T\}] = \frac{2}{3}, \;\; \mathbb{P}[\Omega] = 1.$$

Then, the probabilities for all the events are well defined and consistent.

What is a probability law $\mathbb{P}$?

- A probability law $\mathbb{P}$ is a **function**.
- It takes a subset (an element in $\mathcal{F}$) and maps it to a number between 0 and 1.
- $\mathbb{P}$ is a **measure**. It measures the size of a set.
- For $\mathbb{P}$ to make sense, it must satisfy the **axioms of probability**.

Why these three axioms?

- Axiom I (Non-negativity) ensures that probability is never negative.
- Axiom II (Normalization) ensures that probability is never bigger than 1.
- Axiom III (Additivity) allows us to add probabilities when two events do not overlap.

In words, if A and B are disjoint, then the probability of observing either A or B is the sum of the two individual probabilities. Figure 2.22 illustrates the idea.

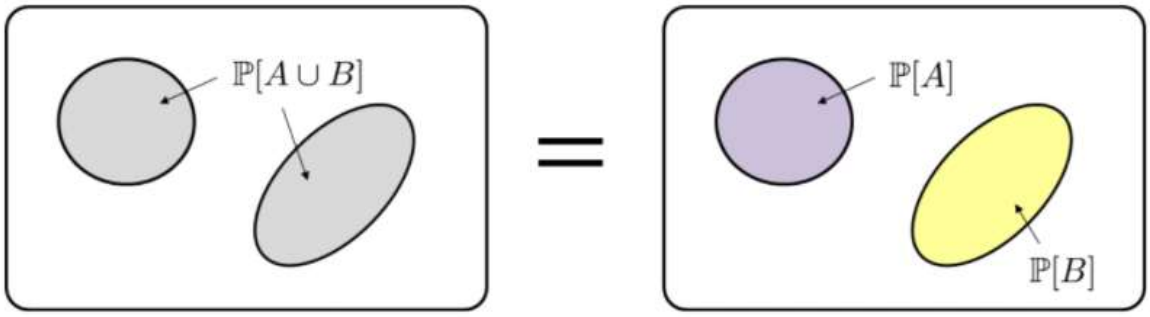


Figure 2.22: Axiom III says $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ if $A \cap B = \emptyset$.

Example 1. Consider a sample space with $\Omega = \{\clubsuit, \heartsuit, \maltese\}$. The probability for each outcome is

$$\mathbb{P}[\{\clubsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{\maltese\}] = \frac{3}{6}.$$

Suppose we construct two disjoint events $E_1 = \{\clubsuit, \heartsuit\}$ and $E_2 = \{\maltese\}$. Then, Axiom 3 says

$$\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2] = \left(\frac{2}{6} + \frac{1}{6}\right) + \frac{3}{6} = 1.$$

Corollary 2.1. *Let $A \in \mathcal{F}$ be an event. Then,*

(a) $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$.

(b) $\mathbb{P}[A] \leq 1$.

(c) $\mathbb{P}[\emptyset] = 0$.

Corollary 2.2 (Unions of two Non-Disjoint Sets)**.** *For any A and B in $\mathcal{F}$,*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

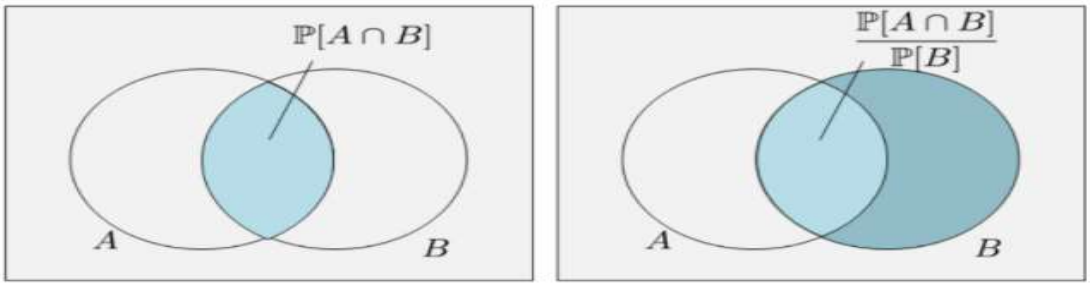
Definition 2.22. *Consider two events A and B. Assume $\mathbb{P}[B] \neq 0$. The* ***conditional probability*** *of A given B is*

$$\mathbb{P}[A \mid B] \stackrel{def}{=} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}. \tag{2.15}$$

- Section 2.4.1: **Conditional probability**. Conditional probability of A given B is $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$.
- Section 2.4.2: **Independence**. Two events are *independent* if the occurrence of one does not influence the occurrence of the other: $\mathbb{P}[A|B] = \mathbb{P}[A]$.
- Section 2.4.3: **Bayes theorem and law of total probability**. Bayes theorem allows us to switch the order of the conditioning: $\mathbb{P}[A|B]$ vs $\mathbb{P}[B|A]$, whereas the law of total probability allows us to decompose an event into smaller events.

Independent and Dependent Events (Important Topic)

Definition . Let E1 and E2 be any two events of a sample space. If the occurrence of E1 does not depend on the occurrence of E2 and the occurrence of E2 does not depend on the occurrence of E1 or in other words the occurrence of any one does not depend on the occurrence of other then E1 and E2 are called independent events otherwise they are called dependent events



e 2.26: Illustration of conditional probability and its comparison with $\mathbb{P}[A \cap B]$.

Example 2. Consider throwing a dice. Let

$$A = \{\text{Getting a 3}\} \quad \text{and} \quad B = \{\text{getting an odd number}\}.$$

Find $\mathbb{P}[A \mid B]$ and $\mathbb{P}[B \mid A]$.

Solution. The following probabilities are easy to calculate:

$$\mathbb{P}[A] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}, \qquad \text{and} \qquad \mathbb{P}[B] = \mathbb{P}[\{\boxdot, \boxdot, \boxtimes\}] = \frac{3}{6}.$$

Also, the intersection is

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}.$$

Given these values, the conditional probability of A given B can be calculated as

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

In words, if we know that we have an odd number, then the probability of obtaining a 3 has to be computed over $\{\boxdot, \boxdot, \boxtimes\}$, which give us a probability $\frac{1}{3}$. If we do not know that we have an odd number, then the probability of obtaining a 3 has to be computed from the sample space $\{\boxdot, \boxdot, \boxdot, \boxdot\!\boxdot, \boxtimes, \boxplus\}$ which will give us $\frac{1}{6}$.

The other conditional probability is

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = 1.$$

Therefore, if we know that the dice is 3, then the probability for this number being an odd number is 1.

Definition 2.23. *Two events A and B are statistically* ***independent*** *if*

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

Why define independence in this way? Recall that $\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$. If A and B are independent, then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$ and so

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\,\mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A]. \tag{2.18}$$

Definition 2.24. *Let A and B be two events such that $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$. Then A and B are independent if*

$$\mathbb{P}[A \mid B] = \mathbb{P}[A] \quad \textit{or} \quad \mathbb{P}[B \mid A] = \mathbb{P}[B]. \tag{2.19}$$

The two statements are equivalent as long as $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$. This is because $\mathbb{P}[A|B] = \mathbb{P}[A \cap B]/\mathbb{P}[B]$. If $\mathbb{P}[A|B] = \mathbb{P}[A]$ then $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$, which implies that $\mathbb{P}[B|A] = \mathbb{P}[A \cap B]/\mathbb{P}[A] = \mathbb{P}[B]$.

If A and B are disjoint, then $A \cap B = \emptyset$. This only implies that $\mathbb{P}[A \cap B] = 0$. However, it says nothing about if $\mathbb{P}[A \cap B]$ can be factorized into $\mathbb{P}[A]\,\mathbb{P}[B]$. If A and B are independent, then we have $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. But this does not imply that $\mathbb{P}[A \cap B] = 0$. The only possibility that Disjoint $\Leftrightarrow$ Independence is when $\mathbb{P}[A] = 0$ or $\mathbb{P}[B] = 0$.

Practice Exercise 2. Throw a dice twice. Are A and B independent?

$$A = \{\text{1st dice is 3}\} \quad \text{and} \quad B = \{\text{sum is 7}\}.$$

Solution. Note that

$$\mathbb{P}[A \cap B] = \mathbb{P}[(3,6)] = \frac{1}{36} \qquad \mathbb{P}[A] = \frac{3}{6}$$

$$\mathbb{P}[B] = \mathbb{P}[(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)] = \frac{1}{6}.$$

So $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. Thus, A and B are independent.

Theorem 2.7. ***(Bayes Theorem)*** *For any two events A and B such that $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$, it holds that*

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\,\mathbb{P}[A]}{\mathbb{P}[B]}.$$

Theorem 2.8. *(**Law of Total Probability**) Let $\{A_1, \ldots, A_n\}$ be a partition of Ω, i.e., $A_1, \ldots, A_n$ are disjoint and $\Omega = A_1 \cup \ldots \cup A_n$. Then, for any $B \subseteq \Omega$,*

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\, \mathbb{P}[A_i].$$

Corollary 2.4. *Let $\{A_1, A_2, \ldots, A_n\}$ be a partition of Ω, i.e., $A_1, \ldots, A_n$ are disjoint and $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$. Then, for any $B \subseteq \Omega$,*

$$\mathbb{P}[A_j \mid B] = \frac{\mathbb{P}[B \mid A_j]\, \mathbb{P}[A_j]}{\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\, \mathbb{P}[A_i]}. \tag{2.22}$$

Proof. We just need to apply Bayes Theorem and Law of Total Probability:

$$\begin{aligned}\mathbb{P}[A_j \mid B] &= \frac{\mathbb{P}[B \mid A_j]\, \mathbb{P}[A_j]}{\mathbb{P}[B]} \\ &= \frac{\mathbb{P}[B \mid A_j]\, \mathbb{P}[A_j]}{\sum_{i=1}^{n} \mathbb{P}[B \mid A_i]\, \mathbb{P}[A_i]}.\end{aligned}$$

Example 1. Consider a tennis tournament. There are three types of players A, B, and C. The percentage of these players are: A 50%, B 25%, and C 25%. Your chance of winning these players are different.

0.3 against player A.
0.4 against player B.
0.5 against player C.

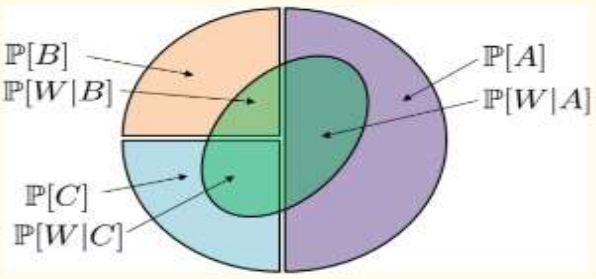
So now, if you enter the game, what is the probability of winning the game? Suppose that you have won the game, what is the probability that you played against player A?

Solution. The first thing to do in this problem is to list out all the available probabilities. We know from the percentage of players that

$$\mathbb{P}[A] = 0.5, \quad \mathbb{P}[B] = 0.25, \quad \mathbb{P}[C] = 0.25.$$

Now, let W be the event that you win the game. Then, the conditional probabilities are defined as follows:

$$\mathbb{P}[W|A] = 0.3, \quad \mathbb{P}[W|B] = 0.4, \quad \mathbb{P}[W|C] = 0.5.$$



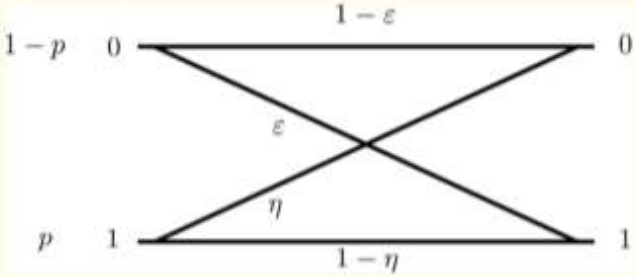
Therefore, by law of total probability, we can show that the probability of winning the game is

$$\begin{aligned}\mathbb{P}[W] &= \mathbb{P}[W \mid A]\, \mathbb{P}[A] + \mathbb{P}[W \mid B]\, \mathbb{P}[B] + \mathbb{P}[W \mid C]\, \mathbb{P}[C] \\ &= (0.3)(0.5) + (0.4)(0.25) + (0.5)(0.25) \\ &= 0.375.\end{aligned}$$

Suppose you have won the game, then the probability of A given W is

$$\begin{aligned}\mathbb{P}[A|W] &= \frac{\mathbb{P}[W|A]\mathbb{P}[A]}{\mathbb{P}[W]} \\ &= \frac{(0.3)(0.5)}{0.375} = 0.4.\end{aligned}$$

Example 2. Consider a communication channel shown below. The probability of sending a 1 is p and the probability of sending a 0 is $1-p$. Given that 1 is sent, the probability of receiving 1 is $1-\eta$. Given that 0 is sent, the probability of receiving 0 is $1-\varepsilon$. Find the probability that a 1 has been correctly received.



Solution. Define the events

$$\begin{aligned} S_0 &= \text{“0 is sent”}, \quad \text{and} \quad R_0 = \text{“0 is received”}. \\ S_1 &= \text{“1 is sent”}, \quad \text{and} \quad R_1 = \text{“1 is received”}. \end{aligned}$$

Then, the probability that 1 is received is $\mathbb{P}[R_1]$. However, $\mathbb{P}[R_1] \neq 1-\eta$ because $1-\eta$ is the conditional probability that 1 is received given 1 is sent. It is possible that we receive 1 as a result of an error when 0 is sent. Therefore, we need to consider the probabilities of having S_0 and S_1. Using Law of total probability we have

$$\begin{aligned}\mathbb{P}[R_1] &= \mathbb{P}[R_1 \mid S_1]\ \mathbb{P}[S_1] + \mathbb{P}[R_1 \mid S_0]\ \mathbb{P}[S_0] \\ &= (1-\eta)p + \varepsilon(1-p).\end{aligned}$$

Now, suppose that we have received 1. What is the probability that 1 was originally sent? This is asking the posterior probability $\mathbb{P}[S_1 \mid R_1]$, which can be found using Bayes Theorem

$$\mathbb{P}[S_1 \mid R_1] = \frac{\mathbb{P}[R_1 \mid S_1]\ \mathbb{P}[S_1]}{\mathbb{P}[R_1]} = \frac{(1-\eta)p}{(1-\eta)p + \varepsilon(1-p)}.$$

Bayes Theorem: (Very Important)

An event B can be explained by a set of exhaustive and mutually exclusive hypothesis A1,A2,….,An. Given ‘a priori‘ probabilities P(A1),P(A2),…,P(An) corresponding to a total absence of knowledge regarding the occurrence of B and conditional probabilities

P(B/A1),P(B/A2),…,P(B/An)

the ‘a posterior‘ probability P(A / B) j of some event Aj is given by

$$P(A_j / B) = \frac{P(A_j).P(B / A_j)}{\sum_{i=1}^{n} P(A_i)P(B / A_i)}$$

Proof: since the event B can occur when either A1 occurs, or A2 occurs, or…. An occurs i.e , B can occur i $B = BA_1 \cup BA_2 \cup BA_3 \cup .....BA_n$ 2 conseq

$$P(B) = P(BA_1 \cup BA_2 \cup BA_3 \cup .....BA_n)$$

Since A_1,A_2…An are mutually exclusive, hence BA_1,BA_2,…BA_n are mutually exclusive forms, therefore by total probability theorem , we have

$$P(B) = P(BA_1) + P(BA_2) + P(BA_3) + \ldots\ldots\ldots + P(BA_n)$$

$$= \sum_{i=1}^{n} P(BA_i) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

Where $P(B/A_i)$ is the conditional probability of B when A_i has already occurred .

Now from the theorem of compound probability , we have

$$P(A_jB) = P(A_j)P(A_j/B)$$

$$P(A_j/B) = \frac{P(A_jB)}{P(B)} = \frac{P(A_j)P(A_j/B)}{P(B)} \qquad \ldots(2)$$

From(1)and (2) we get

$$P(A_j/B) = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)} \qquad \ldots(3)$$

Random Variables:

Random Variables : A random variable is a rule that assigns a numerical value to each possible outcome of a probabilistic experiment.

We denote a random variable by a capital letter (such as “X”)

Examples of random variables:

X: the age of a randomly selected student here today.

Y: the number of planes completed in the past week

Discrete Variables or Continuous Variables: (Important Topic)

A probability distribution for a A probability distribution for a discrete discrete r.v. X consists of: consists of: –

Possible values Possible values $x_1, x_2, \ldots, x_n$

Corresponding probabilities $p_1, p_2, \ldots, p_n$

with the interpretation that $p(X = x_1) = p_1$, $p(X = x_2) = p_2, \ldots, p(X = x_n) = p_n$

Note the following: – Variable names are capital letters (e.g., X)

Values of variables are lower case letters

Each $p_i \geq 0$ and $p_1 + p_2 + \ldots + p_n = 1.0$

- ***Mean or Expected Value:***

Represents "average" outcome; a measure of "central tendency"

$$E(X) = \mu_X = \sum_{i=1}^{n} P(X = x_i)x_i = \sum_{i=1}^{n} p_i x_i$$

- ***Variance:***

Squared deviation around the mean; a measure of "spread"

$$Var(X) = \sigma^2_X = \sum_{i=1}^{n} P(X = x_i)(x_i - \mu_X)^2 = \sum_{i=1}^{n} p_i (x_i - \mu_X)^2$$

- ***Standard Deviation :***

Square root of the variance. A measure of spread in the same units as the random variable X.

$$SD(X) = \sigma_X = \sqrt{\sigma_X^2}$$

Continuous Random Variable:

A continuous random variable can take any real value in some interval

Example: X = time a customer spends waiting in line at the store

• “Infinite” number of possible values for the random variable.

• For a continuous random variable, questions are phrased in terms of a range of values.

Example: We might talk about the event that a customer waits between 5.0 and 10.0 minutes, and not about the event that a customer waits exactly 5.25 minutes!

Normal Distribution (important topic)

The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.

Characteristics of the Normal distribution

• Symmetric, bell shaped

• Continuous for all values of X between $-\infty$ and ∞ so that each conceivable interval of real numbers has a probability other than zero.

• $-\infty \leq X \leq \infty$

• Two parameters, μ and σ. Note that the normal distribution is actually a family of distributions, since μ and σ determine the shape of the distribution.

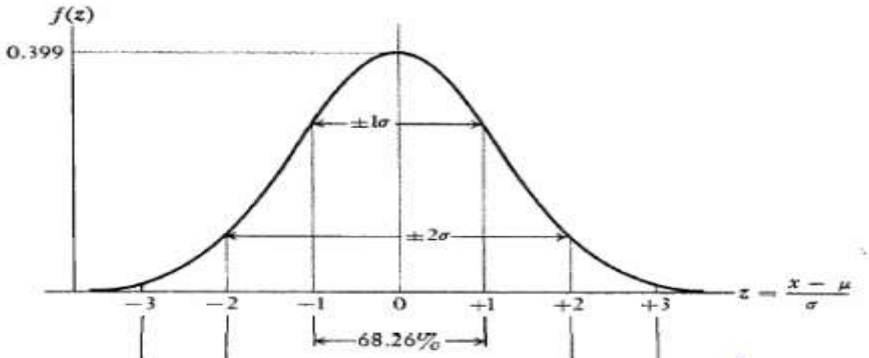
• The rule for a normal density function is

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

• The notation $N(\mu, \sigma^2)$ means normally distributed with mean μ and variance σ^2. If we say $X \sim N(\mu, \sigma^2)$ we mean that X is distributed $N(\mu, \sigma^2)$.

• About 2/3 of all cases fall within one standard deviation of the mean, that is $P(\mu - \sigma \leq X \leq \mu + \sigma) = .6826$.

• About 95% of cases lie within 2 standard deviations of the mean, that is $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = .9544$



Why is the normal distribution useful?

• Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution

• The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.

• There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

Central Limit Theorem:

Regardless of the population distribution model, as the sample size increases, the sample mean tends to be normally distributed around the population

mean, and its standard deviation shrinks as n increases.

Certain conditions must be met to use the CLT.

- The samples must be independent
- The sample size must be “big enough”

CLT Conditions Independent Samples Test

- “Randomization”: Each sample should represent a random sample from the population, or at least follow the population distribution.
- “10% Rule”: The sample size must not be bigger than 10% of the entire population. Large Enough Sample Size
- Sample size n should be large enough so that $np \geq 10$ and $nq \geq 10$

$X_1, X_2, ..., X_n$ are n random variables that are independent and identically distributed with mean μ and standard deviation σ.

- $\bar{X} = (X_1 + X_2 + ... + X_n)/n$ is the sample mean
- We can show $E(\bar{X}) = \mu$ and $SD(\bar{X}) = \sigma/\sqrt{n}$

● CLT states: as $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$

Hypothesis and Inference

Estimation: Use sample statistics to estimate the unknown population parameter.

Point Estimate: the best single value to describe the unknown parameter.

Standard Error (SE): standard deviation of the sample statistic. Indicates how precise is the point estimate.

Confidence Interval (CI): the range with the most probable values for the unknown parameter with a (1-α)% level of confidence.

Hypothesis Testing:

• Suppose X continuous from a population with mean μ and standard deviation σ.

• What is the value of μ?

• We select a random sample from that population and try to make inference about μ.

Null hypothesis (H0): – An explicit statement about an unknown parameter the validity of which you wish to test, e.g., μ=μ0 •

Alternative hypothesis (H1): – An alternative statement about the unknown parameter used to compare your null with, e.g., • μ ≠ μ0 (two-sided test)

• μ < μ0 (one-sided test)

• μ > μ0 (one-sided test)

• Errors: – Type I : reject H0 | H0 is true (crucial)

Type II: do not reject H0 | H1 is true (moderate)

(Follow PPT for t test z test, F test, ANOVA and Chi square test.)

Bayesian Statistical Inference (Important Topic)

Bayes' rule is an equation from probability theory. The various terms in Bayes' rule are all probabilities, but notice that there are conditional probabilities in there. For example, the left hand side of the equation is P(A|B) and that means the probability of A given B. That is, it's the probability of A after taking into account the information B. In other words, P(A|B) is a posterior probability, and Bayes' rule tells us how to calculate it from other probabilities. Bayes' rule is true for any statements A and B.

In Bayesian statistics, most of the terms in Bayes' rule have special names. Some of them even have more than one name, with different scientific communities preferring different terminology. Here is a list of the various terms and the names we will use for them:

• P(A|B) is the posterior probability. It describes how certain or confident we are that hypothesis A

is true, given that we have observed data B. Calculating posterior probabilities is the main goal of Bayesian statistics!

• P(A) is the prior probability, which describes how sure we were that A was true, before we observed the data D.

• P(B|A) is the likelihood. If you were to assume that A is true, this is the probability that you would have observed data B.

• P(B) is the marginal likelihood. This is the probability that you would have observed data B, whether A is true or not

$$P(A|B) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$