CHAPTER 3

Probability The Basis of Statistical Inference

1. General Definitions and Concepts:

- Probability is a measure (or number) used to measure the chance of the occurrence of some event. This number is between 0 and 1.

- An experiment is some procedure (or process) that we do.

- The sample space of an experiment is the set of all possible outcomes of an experiment. Also, it is called the universal set, and is denoted by Ω.

- Any subset of the sample space Ω is called an event.

> - $\phi \subseteq \Omega$ is an event (impossible event)
> - $\Omega \subseteq \Omega$ is an event (sure event)

Example:

Selecting a ball from a box containing 6 balls numbered from 1 to 6 and observing the number on the selected ball. This experiment has 6 possible outcomes.

Solution:

The sample space Ω={1,2,3,4,5,6}.

Consider the following events:

$E_2 =$ getting a number less than 4 $= \{1, 2, 3\} \subseteq \Omega$

$E_3 =$ getting 1 or 3 $= \{1, 3\} \subseteq \Omega$

$E_4 =$ getting an odd number $= \{1, 3, 5\} \subseteq \Omega$

$E_5 =$ getting a negative number $= \{\ \} = \phi \subseteq \Omega$

$E_6 =$ getting a number less than 10 $= \{1, 2, 3, 4, 5, 6\} = \Omega \subseteq \Omega$

Notation:

$n(\Omega) =$ no. of outcomes (elements) in Ω

$n(E) =$ no. of outcomes (elements) in the event E

- The outcomes of an experiment are equally likely if the outcomes have the same chance of occurrence.

- If the experiment has $n(\Omega)$ equally likely outcomes, then the probability of the event E is denoted by P(E) and is defined by:

$$P(E)=\frac{n(E)}{n(\Omega)}=\frac{no.\ of\ outcomes\ \ in\ \ E}{no.\ of\ outcomes\ \ in\ \ \Omega}$$

Example:

In the ball experiment in the previous example, suppose the ball is selected at random. Determine the probabilities of the following events:

E_1 = getting an even number.

E_2 = getting a number less than 4.

E_3 = getting 1 or 3.

Solution:

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad ; \quad n(\Omega) = 6$$

$$E_1 = \{2, 4, 6\} \quad ; \quad n(E_1) = 3$$

$$E_2 = \{1, 2, 3\} \quad ; \quad n(E_2) = 3$$

$$E_3 = \{1, 3\} \quad ; \quad n(E_3) = 2$$

The outcomes are equally likely.

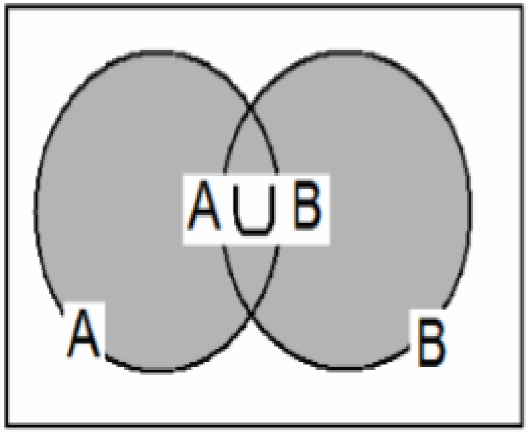
$$\therefore P(E_1) = \frac{3}{6}, \qquad P(E_2) = \frac{3}{6}, \qquad P(E_3) = \frac{2}{6},$$

2. Some Operations on Events:

Let A and B be two events defined on the sample space Ω.

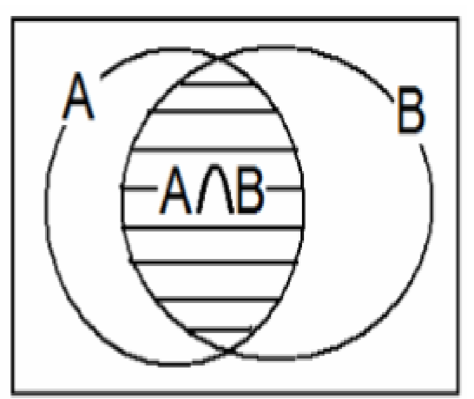
- Union of Two events: (A ∪ B) or (A + B)

The event A ∪ B consists of all outcomes in A or in B or in both A and B.



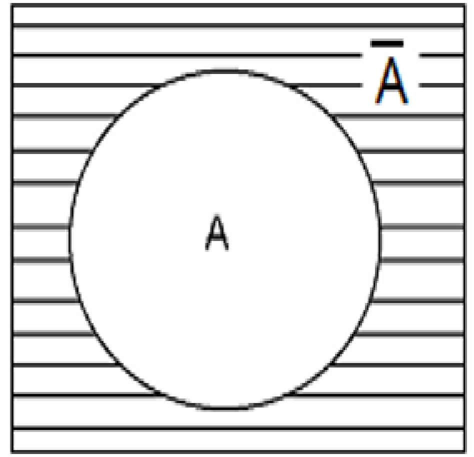
- Intersection of Two Events: (A ∩ B)

The event A ∩ B consists of all outcomes in both A and B.



- Complement of an Event: $\bar{A}$ or A^c or $\grave{A}$

The even A consists of all outcomes of Ω but are not in A.



Example:

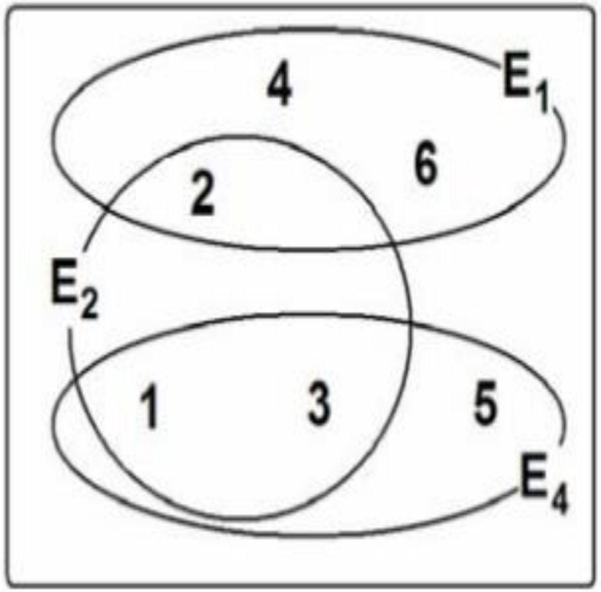
Experiment: Selecting a ball from a box containing 6 balls numbered 1, 2, 3, 4, 5, and 6 randomly.

Define the following events:

$E_1 = \{2,4,6\}$ = getting an even number.

$E_2 = \{1,2,3\}$ = getting a number less than 4.

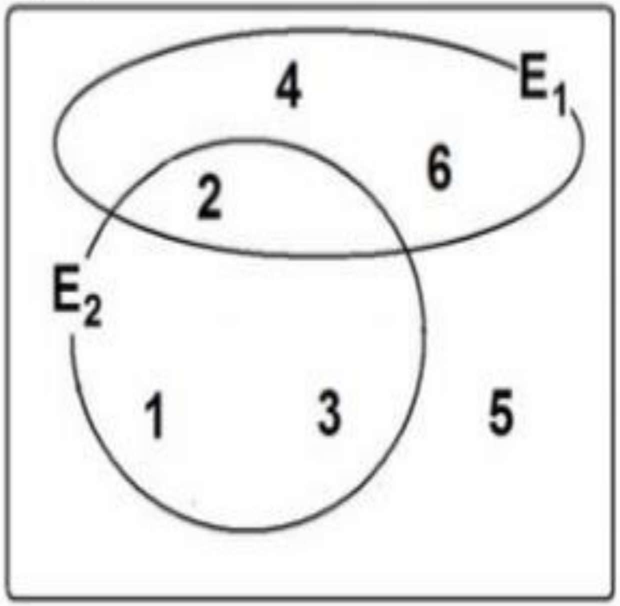
$E_4 = \{1,3,5\}$ = getting an odd number.



(1) $E_1 \cup E_2 = \{1, 2, 3, 4, 6\}$

= getting an even number **or** a number less than 4.

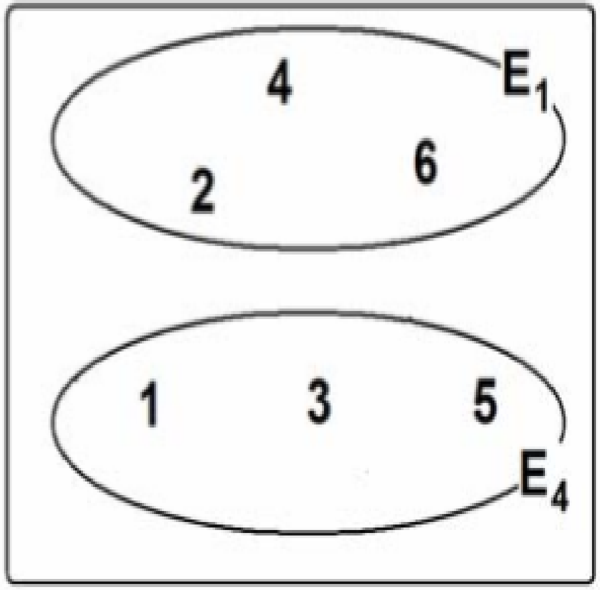
$$P(E_1 \cup E_2) = \frac{n(E_1 \cup E_2)}{n(\Omega)} = \frac{5}{6}$$



(2) $E_1 \cup E_4 = \{1, 2, 3, 4, 5, 6\} = \Omega$

= getting an even number **or** an odd number.

$$P(E_1 \cup E_4) = \frac{n(E_1 \cup E_4)}{n(\Omega)} = \frac{6}{6} = 1$$

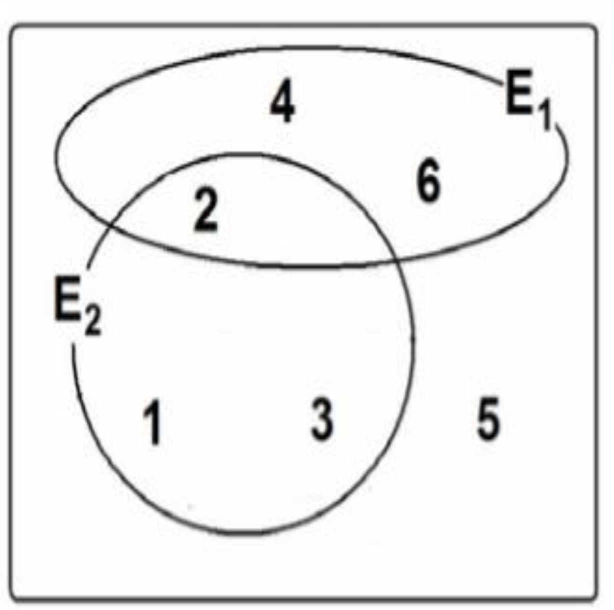


Note:

E_1 and E_4 are exhaustive events. The union of these events gives the whole sample space.

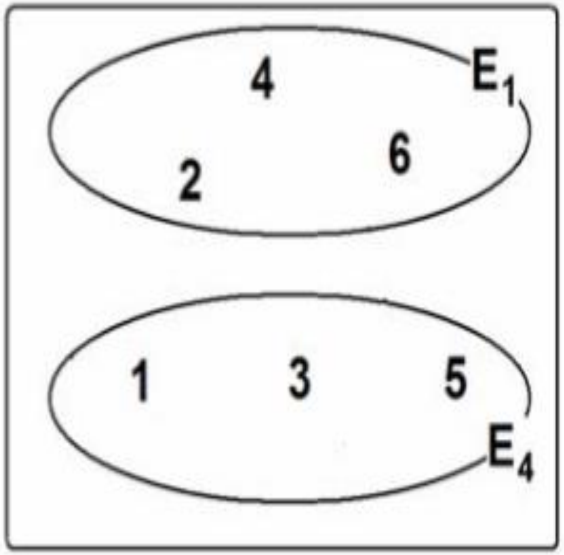
(3) $E_1 \cap E_2 = \{ 2 \}$ = getting an even number **and** a number less than 4.

$$P(E_1 \cap E_2) = \frac{n(E_1 \cap E_2)}{n(\Omega)} = \frac{1}{6}$$



(4) $E_1 \cap E_4 = \phi$ = getting an even number **and** an odd number.

$$P(E_1 \cap E_4) = \frac{n(E_1 \cap E_4)}{n(\Omega)} = \frac{n(\phi)}{6} = \frac{0}{6} = 0$$



Note:

E1 and E4 are called disjoint (or mutually exclusive) events.

(5) The complement of E_1

$\bar{E}_1$ = <u>not</u> getting an even number = $\overline{\{2,4,6\}} = \{1, 3, 5\}$

= getting an odd number.

$= E_4$

3. Mutually exclusive (disjoint) Events:

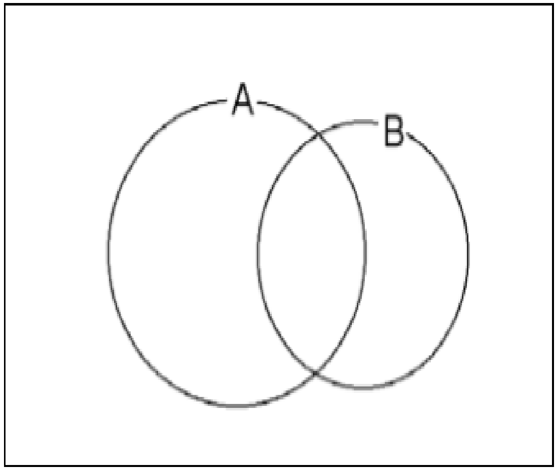
The events A and B are disjoint (or mutually exclusive) if:

$$A \cap B = \phi$$

In this case:

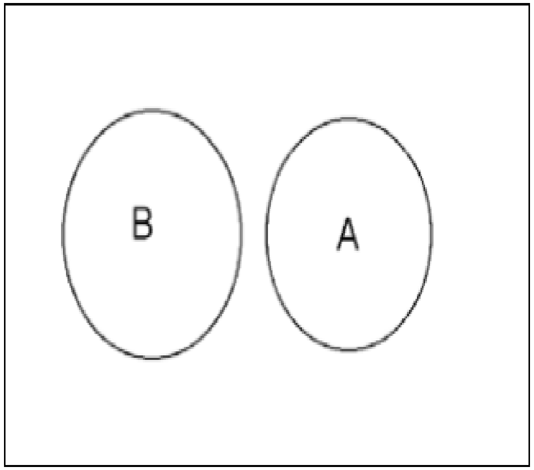
(i) $P(A \cap B) = 0$

(ii) $P(A \cup B) = P(A) + P(B)$



$A \cap B \neq \phi$

A and *B* are not mutually exclusive



$A \cap B = \phi$

A and *B* are mutually exclusive (disjoint)

4. Exhaustive Events:

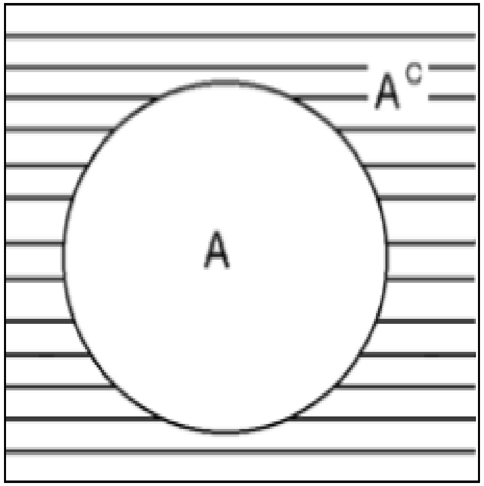
The events $A_1, A_2, \ldots, A_n$ are exhaustive events if:

$$A_1 \cup A_2 \cup \ldots \cup A_n = \Omega.$$

For this case $P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(\Omega) = 1.$

Notes:

1. $A \cup \overline{A} = \Omega$ (A and $\overline{A}$ are exhaustive events)
2. $A \cap \overline{A} = \phi$ (A and $\overline{A}$ are mutually exclusive (disjoint) events)
3. $n(\overline{A}) = n(\Omega) - n(A)$
4. $P(\overline{A}) = 1 - P(A)$



5. General Probability Rules:

1. $0 \le P(A) \le 1$
2. $P(\Omega) = 1$
3. $P(\phi) = 0$
4. $P(\overline{A}) = 1 - P(A)$

6. The addition rule:

- For any two events A and B:

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

- For mutually exclusive (disjoint) events A and B

$$P(A \cup B) = P(A) + P(B)$$

- If the events $A1, A2, \ldots, An$ are exhaustive and mutually exclusive events, then:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n)$$
$$= P(A_1) + P(A_2) + \ldots + P(A_n)$$
$$= P(\Omega) = 1$$

7. Marginal Probability:

The marginal probability of A_i, $P(A_i)$, is equal to the sum of the joint probabilities A_i with all categories of B. That is,

$$P(A_i) = P(A_i \cap B_1) + P(A_i \cap B_2) + \ldots + P(A_i \cap B_n)$$
$$= \sum_{j=1}^{n} P(A_i \cap B_j)$$

Example:

Table of number of elements in each event

	B_1	B_2	B_3	Total
A_1	50	30	70	150
A_2	20	70	10	100
A_3	30	100	120	250
Total	100	200	200	500

Table of probabilities of each event:

	B_1	B_2	B_3	Marginal Probability
A_1	0.1	0.06	0.14	0.3
A_2	0.04	0.14	0.02	0.2
A_3	0.06	0.2	0.24	0.5
Marginal Probability	0.2	0.4	0.4	1

For example,

$$P(A_2) = P(A_2 \cap B_1) + P(A_2 \cap B_2) + P(A_2 \cap B_n)$$
$$= 0.04 + 0.14 + 0.02$$
$$= 0.2$$

Example:

Smoking Habit

Age		Daily (B_1)	Occasionally (B_2)	Not at all (B_3)	Total
	20 - 29 (A_1)	31	9	7	47
	30 - 39 (A_2)	110	30	49	189
	40 - 49 (A_3)	29	21	29	79
	50+ (A_4)	6	0	18	24
	Total	176	60	103	339

A_3 = the selected physician is aged 40 – 49.

B_2 = the selected physician smokes occasionally.

$A_3 \cap B_2$ = the selected physician is aged 40-49 **and** smokes occasionally.

$A_3 \cup B_2$ = the selected physician is aged 40-49 **or** smokes occasionally (or both).

$\overline{A_4}$ = the selected physician is **not** 50 years or older.

$A_2 \cup A_3$ = the selected physician is aged 30-39 **or** is aged 40-49.

1. Conditional probability:

The conditional probability of the event A when we know that the event B has already occurred is defined by:

$$P(A|B)=\frac{P(A\cap B)}{P(B)} \qquad ; \; P(B)\neq 0$$

2. Multiplication Rules of Probability:

$$P(A\cap B)=P(B)P(A|B)$$
$$P(A\cap B)=P(A)P(B|A)$$

Example:

		Smoking Habit			
		Daily (B_1)	Occasionally (B_2)	Not at all (B_3)	Total
Age	20-29 (A_1)	31	9	7	47
	30-39 (A_2)	110	30	49	189
	40-49 (A_3)	29	21	29	79
	50+ (A_4)	6	0	18	24
	Total	176	60	103	339

$(B_1|A_2)$ = the selected physician smokes daily given that his age is between 30 and 39

Solution:

- $P(B_1) = \frac{n(B_1)}{n(\Omega)} = \frac{176}{339} = 0.519$

- $P(B_1 \mid A_2) = \frac{P(B_1 \cap A_2)}{P(A_2)}$

$$= \frac{0.324484}{0.557522} = 0.5820$$

$$\begin{cases} P(B_1 \cap A_2) = \frac{n(B_1 \cap A_2)}{n(\Omega)} = \frac{110}{339} = 0.324484 \\ \\ P(A_2) = \frac{n(A_2)}{n(\Omega)} = \frac{189}{339} = 0.557522 \end{cases}$$

another solution:

$$P(B_1 \mid A_2) = \frac{n(B_1 \cap A_2)}{n(A_2)} = \frac{110}{189} = 0.5820$$

Example:

A training health program consists of two consecutive parts. To pass this program, the trainee must pass both parts of the program. From the past experience, it is known that 90% of the trainees pass the first part, and 80% of those who pass the first part pass the second part. If you are admitted to this program, what is the probability that you will pass the program? What is the percentage of trainees who pass the program?

Solution:

A = the event of passing the first part
B = the event of passing the second part
A∩B = the event of passing the first part and the second Part = the event of passing both parts = the event of passing the program

Then, the probability of passing the program is P(A∩B).

$$P(A \cap B) = P(A)\,P(B|A) = (0.9)(0.8) = 0.72$$

3. Independent events:

- $P(A|B) > P(A)$
  (knowing B increases the probability of occurrence of A)
- $P(A|B) < P(A)$
  (knowing B decreases the probability of occurrence of A)
- $P(A|B) = P(A)$
  (knowing B has no effect on the probability of occurrence of A). In this case A is independent of B.

Definition:

Two events A and B are independent if one of the following conditions is satisfied:

$$(i)\quad P(A \mid B) = P(A)$$

$$(ii)\quad P(B \mid A) = P(B)$$

$$(iii)\quad P(B \cap A) = P(A)P(B)$$

Example:

Suppose that A and B are two events such that:

$$P(A) = 0.5, \quad P(B)=0.6, \quad P(A \cap B)=0.2.$$

Theses two events are not independent (they are dependent) because:

$$P(A)\,P(B) = 0.5 \times 0.6 = 0.3$$

$$P(A \cap B)=0.2.$$

$$P(A \cap B) \neq P(A)\,P(B)$$

Also, $P(A)= 0.5 \neq P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.2}{0.6} = 0.3333$.

Also, $P(B) = 0.6 \neq P(B|A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.2}{0.5} = 0.4$.

Example:

Complete the following two-way table:

	B	$\overline{B}$	Total
A	**0.2**	?	**0.5**
$\overline{A}$	?	?	?
Total	**0.6**	?	**1.00**

Solution:

	B	$\overline{B}$	Total
A	**0.2**	0.3	**0.5**
$\overline{A}$	0.4	0.1	0.5
Total	**0.6**	0.4	**1.00**

Given the previous table, we can easily calculate:

$$P(\bar{A}) = 0.5$$

$$P(\bar{B}) = 0.4$$

$$P(A \cap \bar{B}) = 0.3$$

$$P(\bar{A} \cap B) = 0.4$$

$$P(\bar{A} \cap \bar{B}) = 0.1$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.6 - 0.2 = 0.9$$

$$P(A \cup \bar{B}) = P(A) + P(\bar{B}) - P(A \cap \bar{B}) = 0.5 + 0.4 - 0.3 = 0.6$$

$$P(\bar{A} \cup B) = exercise$$

$$P(\bar{A} \cup \bar{B}) = exercise$$

Sensitivity, Specificity, and Bayes' Theorem.

Introduction

Consider the following definitions:

D : the individual has the disease (presence of the disease)

$\bar{D}$: the individual does not have the disease (absence of The disease)

T : the individual has a positive screening test result

$\bar{T}$: the individual has a negative screening test result

We will have 4 possible situations:

		True status of the disease	
		+ve (D: Present)	−ve ($\bar{D}$:Absent)
Result of the test	+ve (T)	Correct diagnosing	false positive result
	−ve ($\bar{T}$)	false negative result	Correct diagnosing

1. Sensitivity

The sensitivity of a test is the probability of a **positive** test result given the **presence** of the disease.

$P(T|D)$ = P(positive result of the test | presence of the disease)

2. Specificity

The specificity of a test is the probability of a negative test result given the absence of the disease.

$P(\bar{T}|\bar{D})$ = P(negative result of the test | absence of the disease)

Example:

Suppose we have a sample of (n) subjects who are cross-classified to Disease Status and Screening Test Result as follows:

Test Result	Disease: Present (D)	Disease: Absent ($\bar{D}$)	Total
Positive (T)	a	b	$a + b = n(T)$
Negative ($\bar{T}$)	c	d	$c + d = n(\bar{T})$
Total	$a + c = n(D)$	$b + d = n(\bar{D})$	n

Solution:

We can compute the following conditional probabilities:

1. The probability of false positive result:

$$P(T \mid \overline{D}) = \frac{n(T \cap \overline{D})}{n(\overline{D})} = \frac{b}{b+d}$$

2. The probability of false negative result:

$$P(\overline{T} \mid D) = \frac{n(\overline{T} \cap D)}{n(D)} = \frac{c}{a+c}$$

3. The sensitivity of the screening test:

$$P(T \mid D) = \frac{n(T \cap D)}{n(D)} = \frac{a}{a+c}$$

4. The specificity of the screening test:

$$P(\overline{T} \mid \overline{D}) = \frac{n(\overline{T} \cap \overline{D})}{n(\overline{D})} = \frac{d}{b+d}$$

Example:

A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease. The two samples were drawn from populations of subjects who were 65 years of age or older. The results are as follows:

Test Result	Alzheimer Disease		
	Present (D)	Absent ($\bar{D}$)	Total
Positive (T)	436	5	441
Negative ($\bar{T}$)	14	495	509
Total	450	500	950

Solution:

Using these data we estimate the following quantities

1. The sensitivity of the test:

$$P(T \mid D) = \frac{n(T \cap D)}{n(D)} = \frac{436}{450} = 0.9689$$

2. The specificity of the test:

$$P(\overline{T} \mid \overline{D}) = \frac{n(\overline{T} \cap \overline{D})}{n(\overline{D})} = \frac{495}{500} = 0.99$$