

Probability Theory

Before studying statistics, it is necessary to learn some fundamentals of probability theory. We'll start with a few definitions in ordinary set theory, and use them as we move into probability.

If each element of a set A is also an element of a set B , the A is a **subset** of B , which we denote by $A \subset B$.

Suppose $A = \{1, 2, 3\}$ and $B = \{1, 2, 3, 4\}$. Then $A \subset B$.

If a set A has no elements, A is called the **null set**, which we denote by $A = \emptyset$.

The set of all elements which belong to at least one of A and B is called the **union** of A and B , and is denoted by $A \cup B$.

Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$. Then $A \cup B = \{1, 2, 3, 4, 5\}$.

The set of all elements that belong to each of the sets A and B is called the **intersection** of A and B , and is denoted by $A \cap B$.

Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$. Then $A \cap B = \{3\}$.

Let Ω denote the set comprised of the totality of all elements in our space of interest. If $A \subset \Omega$, then the set of all elements of Ω which do not belong to A is called the **complement** of A (with respect to Ω), and is denoted by \bar{A} .

Let $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, \dots\}$, the positive integers, and let $A = \{2, 4, 6, 8, \dots\}$.

Then $\bar{A} = \{1, 3, 5, 7, 9, \dots\}$.

Let $A_1, A_2, A_3, \dots, A_k$ be k subsets of Ω . We say that these sets are **mutually exclusive** if $A_i \cap A_j = \emptyset$ for all pairs (i, j) such that $i \neq j$.

Next we'll define some concepts in probability.

A **random experiment** is an experiment for which the outcome cannot be predicted with certainty, but all possible outcomes can be identified prior to its performance, and it may be repeated under the same conditions.

The set of all possible outcomes of a random experiment is called the **sample space**. The sample space is denoted by Ω .

Let A denote a subset of the sample space, $A \subset \Omega$. Then A is called an **event**. $\{\}$ is often used to denote an event. When obvious, notation will be dropped.

Examples

- Experiment is blood test to determine HIV status. Possible outcomes are $\{HIV+\}$ and $\{HIV-\}$.
 A_1 could be the event that a test comes out positive. A_2 could be the event that a test comes out negative.
- Experiment is blood test and further screening to determine HIV status (HIV+ or HIV-) and AIDS status (D+ or D-). Events are:
 $\{(HIV+, D+)\}, \{(HIV+, D-)\}, \{(HIV-, D+)\}, \{(HIV-, D-)\}$
- Experiment is to record the number of people that get tested for HIV in one week at a given clinic. Suppose 500 is the maximum possible number of tests given in a week. Then any non-negative integer less than or equal to 500 is a conceivable outcome. Events are:

$$\{0\}, \{1\}, \{2\}, \dots, \{500\}$$

Note that unions and intersections of events are events.

A_1 is the event that greater than 100 people get tested.

A_2 is the event that fewer than 220 people get tested.

A_3 is the event that greater than 100 people but fewer than 220 get tested.

Number of people tested for HIV at a given clinic.

Week	Number Tested	Week	Number Tested
1	121	8	312
2	38	9	201
3	208	10	221
4	415	11	389
5	99	12	305
6	109	13	43
7	296	14	483

Consider Experiment 3, which consists of counting the number of people that get tested for HIV in a given week. Suppose we would like to estimate $P(A_1)$, the probability that greater than 100 people get tested. We can repeat the experiment for many weeks (say 14) to estimate $P(A_1)$. Suppose results are as given in the table above.

We will adopt the **relative frequency** interpretation of probability, which says that the probability that an event A occurs is equal to the proportion of the time that A occurs if we repeat the random experiment again and again to infinity:

Let $I_j(A) = 1$ if A occurs on the j th experiment, and equal 0 otherwise.

$$P(A) = \lim_{n \rightarrow \infty} \frac{\sum_{j=1}^n I_j(A)}{n}$$

Clearly, $P(\cdot)$ is a function whose argument is a subset of Ω , and for every event A , $0 \leq P(A) \leq 1$. We refer to the function P as a **probability measure**.

Estimate of $P(A_1)$ is _____ .

If we did this "forever" (as the number of weeks $n \rightarrow \infty$), we would know the true $P(A_1)$.

$$P(A) = \lim_{n \rightarrow \infty} \frac{\sum_{j=1}^n I_j(A)}{n}$$

Laws of Probability: Let Ω be the sample space for a probability measure P .

i. $P(A) \geq 0$, for all events A .

ii. Let A_1, A_2, \dots, A_k be mutually exclusive events.

Then $P(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k)$

iii. $P(\Omega) = 1$

Several important theorems can be derived from these properties.

1. For each $A \subset \Omega$, $P(A) = 1 - P(\bar{A})$.
2. $P(\emptyset) = 0$
3. If $A_1 \subset A_2 \subset \Omega$, $P(A_1) \leq P(A_2)$
4. For each $A \subset \Omega$, $0 \leq P(A) \leq 1$
5. For any events A_1 and A_2 ,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

Example: Suppose that after intensive testing and screening, we can always accurately identify people that are *HIV+* and those that are *HIV-* (i.e., there are no misclassifications) and suppose we have done this for 100,000 people resulting in 89,000 *HIV-* (11,000 *HIV+*).

A new, quicker, less expensive, test can also be used to detect HIV. Suppose we want to evaluate this test and so we run it on the 100,000 people of known HIV status. The new test gave 15,000 positive (10,000 were truly positive and 5,000 were truly negative) and 85,000 negative (84,000 were truly negative and 1,000 were positive).

Probabilities from the hypothetical experiment are given below.

test result	HIV status		total
	(<i>HIV</i> ⁺)	(<i>HIV</i> ⁻)	
+	0.10	0.05	0.15
-	0.01	0.84	0.85
total	0.11	0.89	1.00

test result	HIV status		total
	(<i>HIV</i> ⁺)	(<i>HIV</i> ⁻)	
+	0.10	0.05	0.15
-	0.01	0.84	0.85
total	0.11	0.89	1.00

What is the chance of either having HIV or getting a positive result?

$$\begin{aligned} P[\text{HIV}^+ \cup \text{test}^+] &= 0.11 + 0.15 - 0.10 \\ &\neq 0.11 + 0.15 \\ &= P[\text{HIV}^+] + P[\text{test}^+]. \end{aligned}$$

This is because *HIV*⁺ and *test*⁺ are NOT mutually exclusive.

Consider the following questions:

- What is the probability that the test is positive given the individual has HIV? (____)
- What is the probability that the test is positive given the individual does not have HIV? (____)
- What is the probability that an individual has HIV given the test is positive? (____)

These questions have to do with **conditional probability**. The conditional probability of *C* given *D* is denoted $P[C|D]$.

The conditional probability $P[C|D]$ can be determined from the **multiplication rule**. The **multiplication rule** of probability for any events C and D is given by

$$P[C \cap D] = P[D]P[C|D],$$

or equivalently

$$P[C|D] = P[C \cap D]/P[D].$$

This is how we computed answers to the previous three questions. Another consequence of the multiplication rule has to do with independence:

- If C and D are independent, the $P[C|D] = P[C]$.
- As a result, if C and D are independent, $P[C \cap D] = P[C]P[D]$.
- In other words, if C and D are independent, the event D is not influenced by the event C . (What about HIV test results for two unrelated individuals?)

Positive Predictive and Negative Predictive Values

Consider the table of values used to assess the sensitivity and specificity of a test

test result	disease category	
	diseased (+)	nondiseased (-)
+	A	B
-	C	D

Positive Predictive Value: is the proportion of people who tested positive that truly are positive. (estimated by $PPV = A/(A + B)$).

Negative Predictive Value: is the proportion of people who tested negative that truly are negative. (estimated by $NPV = D/(C + D)$).

False Negative: The probability of a false negative is the probability of testing negative given a truly positive condition.

False Positive: The probability of a false positive is the probability of testing positive given a truly negative condition.

Sensitivity and Specificity

Data for assessing the sensitivity and specificity of a test are usually of the form

test result	disease category	
	diseased (+)	nondiseased (-)
+	A	B
-	C	D

Sensitivity: is the proportion of diseased people who would be correctly classified (estimated by $Sens = A/(A + C)$).

Specificity: is the proportion of nondiseased people who would be correctly classified (estimated by $Spec = D/(B + D)$).

The **prevalence** of a disease is the percent of the population with the disease (estimated by $R = (A + C)/(A + B + C + D)$). Note that a random sample is required to estimate prevalence.

For example, consider again the following data from that study to evaluate the new test for HIV:

test result	disease category		total
	HIV +	HIV -	
+	10,000	5,000	15,000
-	1,000	84,000	85,000
total	11,000	89,000	100,000

- The estimated sensitivity is $Sens = A/(A + C) = 10,000/11,000 = 0.909$ (90.9%).
- The estimated specificity is $Spec = D/(B + D) = 84,000/89,000 = 0.944$ (94.4%).
- The estimated ppv is $ppv = A/(A + B) = 10,000/15,000 = 0.667$ (66.7%).
- The estimated npv is $npv = D/(C + D) = 84,000/85,000 = 0.988$ (98.8%).
- The estimated prevalence is $R = (A + C)/(A + B + C + D) = 11,000/100,000 = 0.11$ (11.0%). This is a reasonable estimator only if the sample is random.

Relationships between Sens, Spec, PPV, NPV, FN and FP.

Since 9.1 % of the truly positive people tested negative (1,000 out of 11,000), then $(100 - 9.1 = 90.9)$ % of these truly positive people tested positive. Note that $10,000/11,000 = 0.909$.

$\text{Prob}(\text{false neg}) = 0.091$ and Sens is $1 - 0.091 = 0.909$.

Convince yourself that $\text{Spec} = 1 - \text{PFP}$ where PFP is the probability of a false positive.