CGS698C, Module 1: Sets, probability, and random variables

Himanshu Yadav

2024-05-17

Contents

1 Set theory 2

1.1 Binary operations on sets 2

1.2 The algebra of sets 3

2 Probability theory 3

2.1 Foundations 3

2.2 Probability mass function and probability density function 5

3 Random variables 6

3.1 Discrete random variables 7

3.2 The expected value and variance of a random variable 8

3.3 Continuous random variables 9

3.4 Some important probability distributions 11

4 Conditional probability and Bayes' theorem 12

4.1 Conditional probability 12

4.2 Independent events 13

4.3 Total probability 13

4.4 Bayes' theorem 13

5 Using Bayes' theorem for statistical inference 14

1 Set theory

A set is a collection of objects or elements. Suppose that set A consists of two numbers 0 and 1. We can denote this set as follows:

$$A = \{0, 1\}$$

From the above, we can also say that 0 is a member of set A:

$$0 \in A$$

Similarly, $1 \in A$.
The number 2 is not a member:

$$2 \notin A$$

Let us say S is a set of natural numbers between 2 and 8. We can write:

$$S = \{2, 3, 4, 5, 6, 7, 8\}$$

Also, we can describe the set S as follows. The vertical bar, $|$, is read "such that."

$$S = \{x | x \in \mathbb{N}^+ \text{ and } 2 \leq x \leq 8\}$$

If A = {1,2,3}, B= {1,3,2} and C={3,1,2,1}, we can write $A = B = C$.

1.1 Binary operations on sets

1. The union of two sets A and B ia denoted by $A \cup B$

   $A \cup B$ is the set of all objects that are a member of A, or B, or both.

2. The intersection of two sets A and B is denoted by $A \cap B$

   $A \cap B$ is the set of all objects that are members of both A and B.

3. Set difference for the sets B and A is denoted by $B \backslash A$

   $B \backslash A$ is the set of all members of B that are not members of A.

   $B \backslash A = \{x | x \in B \text{ and } x \notin A\}$

4. The Cartesian product of A and B, denoted by $A \times B$

   $A \times B$ is the set whose members are all possible ordered pairs (a, b), such that $a \in A$, and $b \in B$.

5. A set A is a subset of another set B, denoted by $A \subseteq B$ if all members of A are also in B, i.e., for all $a \in A$, $a \in B$.

   A is a proper subset of B, denoted by $A \subset B$ if all members of A are also in B, but $A \neq B$.

6. The power set of a set A, denoted by $\mathcal{P}(A)$, is the set of all possible subsets of A.

7. Two sets A and B are called disjoint sets if A and B have no element in common.

   $A \cap B = \emptyset$ where $\emptyset$ represent an empty set (a set with no elements in it).

8. The complement of a set A is denoted by $\bar{A}$ or A^c

   $\bar{A}$ is the set of all those elements which belong to the universal set U but does not belong to A.

$$\bar{A} = \{x | x \notin A\}$$

1.2 *The algebra of sets*

1. $A \cup A = A$
   $A \cap A = A$
2. $A \cup B = B \cup A$
   $A \cap B = B \cap A$
3. $A \cup (B \cup C) = (A \cup B) \cup C$
   $A \cap (B \cap C) = (A \cap B) \cap C$
4. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
5. $A \cup \emptyset = A$
   $A \cap \emptyset = \emptyset$

2 *Probability theory*

Suppose you run an experiment where the participants have to decide whether a given sentence is grammatically correct or not. And, the participants are forced to select either yes or no. The recorded responses you have are "yes" and "no."

What is the set of all possible outcomes from the experiment?
Ω = {"yes", "no"}
The set of all possible outcomes from the experiment is called the **sample space** of the experiment.
What is the power set of the sample sapce Ω?
F = {$\emptyset$,{"yes"},{"no"},{"yes","no"}}
$\emptyset$: there is no outcome
{"yes"} : the outcome is "yes"
{"no"} : the outcome is "no"
{"yes","no"} : the outcome is either "yes" or "no"

The above are the all different collections of possible results. These collections are called **events**. For example, {"no"} is an event that the participant answers "no"; {"yes","no"} is the event that the participant answers either "yes" or "no". The proper set F is called the **event space**.

Probability is a way of assigning every event a real value between 0 and 1 based on some requirements. What are those requirements? How do we assign a probability value to every event?

2.1 *Foundations*

We can assign a probability value $P(E)$ to an event E based on the following three axioms.

1. **First axiom:**

   The probability of an event E is a non-negative real number

   $P(E) \in \mathbb{R},\ P(E) \geq 0$ where $E \in F$

   It follows that $P(E)$ is always finite.

2. **Second axiom:**

   The probability that at least one of the elementary events in the entire sample space will occur is 1.

   $P(\Omega) = 1$

3. **Third axiom:**

   Any countable sequence of disjoint sets (also called mutually exclusive events) $E_1, E_2, \ldots E_n$ satisfies the following

   $P(E_1 \cup E_2 \cup E_3 \cup \ldots) = P(E_1) + P(E_2) + P(E_3) + \ldots$

   $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

Let's see what we can deduce from the above three axioms about our grammaticality judgment example.

Suppose that E_1 and E_2 are two mututally exclusive events in the sample space Ω, and an empty set $\emptyset$ is also an event in the same sample space. According to the third axiom,

$$P(E_1 \cup E_2 \emptyset \cup \emptyset \cup \emptyset \ldots) = P(E_1) + P(E_2) + P(\emptyset) + P(\emptyset) + P(\emptyset) + \ldots \tag{1}$$

From set theory you know that $E_1 \cup \emptyset \cup \emptyset = E_1$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) + \sum_{i=3}^{\infty} P(\emptyset) \tag{2}$$

From the first axiom we know that $P(\emptyset) \geq 0$, $P(E_1) \geq 0$, $P(E_2) \geq 0$ and $P(E_1 \cup E_2)$ is finite. Hence,

$$P(\emptyset) = 0 \tag{3}$$

Let us go back to our grammaticality judgment experiment.
Sample space: Ω = {"yes", "no"}
Event space: F = {$\emptyset$,{"yes"},{"no"},{"yes","no"}}
We just verified that

$$P(\emptyset) = 0 \tag{4}$$

Now, the second axiom implies that:

$$P(\{"yes"\} \cup \{"no"\}) = 1 \tag{5}$$

Finally, the third axiom implies that,

$$P(\emptyset \cup \{"yes"\} \cup \{"no"\}) = P(\emptyset) + P(\{"yes"\}) + P(\{"no"\}) \tag{6}$$

$$P(\{"yes"\} \cup \{"no"\}) = P(\emptyset) + P(\{"yes"\}) + P(\{"no"\}) \tag{7}$$

$$1 = 0 + P(\{"yes"\}) + P(\{"no"\}) \tag{8}$$

So,

$$P(\{"yes"\}) + P(\{"no"\}) = 1 \tag{9}$$

Above equation implies that the sum of probabilities of all elementary events in the sample space Ω is equal to 1. More generally, if $x \in \Omega$:

$$\sum_{x \in \Omega} f(x) = 1 \tag{10}$$

where $f(x) \in [0, 1]$.

f is a function that assigns a probability value to each elementary event x.

Also, for any event $E \in F$:

$$P(E) = \sum_{x \in E} f(x) \tag{11}$$

2.2 *Probability mass function and probability density function*

The function $f(x)$ in Equation~11 maps a discrete outcome x in the sample space Ω to a probability value; it is called a **probability mass function**.

Now, consider another experiment. You record the reading times for each participant: how much time (in milliseconds) does it take to read a sentence?

What is the sample space now?

$\Omega = \mathbb{R}^+$

This sample space is not a finite or countable set now. It is a continuous sample space.

How do we assign probabilities to an outcome x such that $x \in \Omega$?

$$\int_0^\infty f(x)\, dx = 1 \tag{12}$$

Suppose an event E exists such that $E \subset \mathbb{R}^+$

$$P(X \in E) = \int_{x \in E} f(x)\, dx \tag{13}$$

The function $f(x)$ maps the values (outcomes) in the continuous sample space Ω to a continuous probability space, such that

$$P(X \leq x) = \int_0^x f(x)\, dx \tag{14}$$

The function $f(x)$ is called a **probability density function**.

In the above equation, there is a variable we have not defined yet, the variable X. The value of X depend on the outcomes in the continuous sample space; it is called a continuous random variable. More generally, for any experiment, you can define a random variable X whose values depend on the outcome of the experiment. We will talk about random variables in the next section.

3 *Random variables*

A random variable X is a function that maps the outcomes in a sample space Ω (say {"yes","no"}) to another (real-valued) space $\otimes_{\S}$ (e.g., $\{0,1\}$ where 1 corresponds "yes" and 0 corresponds to "no"). We can write a random variable X as

$X : \Omega \to \Omega_x$

such that

$X(\omega) \in \Omega_x$ where $\omega \in \Omega$

For example, in a single-coin toss experiment, the sample space is $\Omega = \{H, T\}$. We can define a random variable X which is a function that counts the number of heads in an outcome ω that belongs to Ω.

X: No. of heads in ω where $\omega \in \Omega$

$$X(\omega) = \begin{Bmatrix} 0 & \text{if } \omega \neq H \\ 1 & \text{if } \omega = H \end{Bmatrix} \text{ where } w \in \Omega$$

Similarly, we can also define a random variable Y that counts the number of tails in an outcome ω.

The probabilities are always assigned to the values of the random variable. We will see in the next part that why the random variables are so useful for assigning probability values.

In the case of continuous sample space, the measurable space is often the same as the sample space of the experiment. Suppose an experiment (or any generative process) produces outcomes that belongs to a sample space $\Omega = \{x | x \in \mathbb{R}^+ \text{ and } 2 \leq x \leq 5\}$. These outcomes are values that are coming from what is known as a (continuous) random variable. Suppose that in an experimental trial the outcome is 2.5; we will write $X = 2.5$, where X is a random variable associated with the experiment. A random variable is written with capital letters (X or Y, etc.), and the outcomes are written in lower case (x or y, etc.). In Bayesian statistics, where parameters are also random variables, it is common to use Greek letters like α, β, etc., to represent random variables (i.e., the capital letter convention generally only applies to letters of the English alphabet).

Let us see how it is more convenient to assign probabilites to the values of the random variable rather than assigning probabilities to the sample space Ω.

Consider an experiment where a coin is tossed three times. What will be the sample space of the experiment?

There are a total of eight possible outcomes.

$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

It is difficult to directly assign probabilities to this sample space, because we will need a probability mass function that has probabilities defined for all 8 outcomes.

Now consider a different idea. What if we ask: how many heads appear (in a trial / experiment) when a coin is tossed three times? We can define a random variable X.

X: No. of heads in the outcome ω where $\omega \in \Omega$.

If we represent outcome of each toss as ω_i, we can write

$\omega = (\omega_1, \omega_2, \omega_3) \in \Omega$. The random variable X is given by

$$X(\omega) = \sum_{i=1}^{3} \phi(\omega_i) \quad \text{where } \phi(\omega_i) = \begin{Bmatrix} 0 & \text{if } \omega_i \neq H \\ 1 & \text{if } \omega_i = H \end{Bmatrix}$$

The above random variables yields the following values:

X(HHH) = 3, X(THH)=2, and so on.

Hence, $X(w) \in \{0, 1, 2, 3\}$

So, the random variable X takes "number of heads in three coin-tosses", i.e., $\{0, 1, 2, 3\}$, as its values.

An experiment can be associated with more than one random variable. For example, consider another idea: how many tails appear when a coin is tossed three times?

You can define another random variable $Y : \Omega \rightarrow \Omega_y$ which takes "the number of tails in three coin-tosses'' as its values. It would map the sample space Ω to another space Ω_y, such that $\otimes_y = \{0, 1, 2, 3\}$

- Random variables can be discrete. For example, in our coin tossing example, the random variable X takes a countable list of values (i.e., 0, 1, 2 and 3).
- Random variables can be continuous: they can take any numerical value in an interval or collection of intervals. For example, in an experiment where we record response or reading time, a random variable X associated with the experiment can take any positive real number value.
- A random variable is associated with a function, called probability mass function (PMF) for discrete random variables, and probability density function (PDF) for continuous random variables.
- The PMF assigns probabilities to the values of a discrete random variable. The PDF assigns probabilities to particular **intervals** (ranges) of values of a continuous random variable. The PDF does not assign a probability to a point value, but rather a density.

So, for any experiment or any generative process, you can define a sample space Ω, a random variable X that maps its sample space to another space $\otimes_x$, an event space F which is a power set of $\otimes_x$, and a function P that maps the event space F to a set of probability values.

The sample space Ω, the event space F, and the function P from the event space F to a set of probabilities together make the formal model of an underlying generative process, denoted by (Ω, F, P).

3.1 Discrete random variables

Suppose a discrete random variable X takes the values $x_1, x_2, x_3, \ldots, x_n$. What is the probability that the random variable X takes the value x_i, where $i = 1, \ldots, n$? The probabilities can be assigned by the probability mass functon f, such that

$$P(X = x_i) = f(x_i)$$

under the requirement

$$\sum_{i=1}^{n} f(x_i) = 1$$

An important example: The binomial random variable

Suppose that in an experiment, the trials are independent. And, in each trial, one of the two possible outcomes can occur with probabilities p and $1 - p$. If p remains constant throughout the experiment, each one of these trials is called a Bernoulli trial. Bernoulli trials can represent the generative processes where each outcome is strictly binary, such as heads/tails, on/off, up/down, etc. The pair

of possible outcomes is usually represented by success/failure where p is the probability of success and $1 - p$ is the probability of failure.

The sample space is $S = \{\text{success}, \text{failure}\}$. For a single Bernoulli trial, let us define a random variable X such that success is assigned a real number value 1 and failure is assigned 0.

$x_i =$	0	1
$f(x_i) =$	$1 - p$	p

Table 1: A random variable in which two outcomes are possible: success or failure. The outcome success is assigned the number 1, and failure the number 0, and a probability is assigned to each number.

A further distribution can arise from Bernoulli trials. Consider an experiment containing n independent Bernoulli trials. Suppose there were k successes in n trials. We can define a new random variable X which takes number of successes (out of total number of trials) as its values.

The probability distribution of the random variable X that represents the number of successes in n Bernoulli trials is given by

$$P(X = k) = f(k, n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{15}$$

The expression $\frac{n!}{k!(n-k)!}$ is written as $\binom{n}{k}$ in mathematics, leading to the above PMF being commonly written as:

$$P(X = k) = f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{16}$$

The above distribution is called the binomial distribution, and the random variable is called the binomial random variable.

$k =$	0	1	2	...	n
$f(k, n, p) =$	$p^k(1-p)^{n-k}$	$np^k(1-p)^{n-k}$	$\frac{n(n-1)}{2} p^k(1-p)^{n-k}$	...	$p^k(1-p)^{n-k}$

Table 2: The probability mass function when we carry out n independent Bernoulli trials.

3.2 *The expected value and variance of a random variable*

The expected value (also called expectation, mean, first moment) of a random variable X is the weighted average of the possible values. Suppose a discrete random variable X can take values $x_1, x_2, x_3, \ldots$, with probabilities $f(x_1), f(x_2), f(x_3), \ldots$, where f represent the probability mass function. The expected value of X is given by

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

The expected value of X is the arithmetic mean of large number of independently drawn values for the variable X.

The expected value satisfies the following relationships

1. $E(cX) = cE(X)$
2. $E(X + Y) = E(X) + E(Y)$
3. $E(XY) = E(X)E(Y)$ (if X and Y are independent)

The variance of a random variable X is given by:

$$Var(X) = E[(X - E(X))^2]$$

This can be rewritten as:

$$Var(X) = E[X^2 + E(X)^2 - 2XE(X)]$$

Equivalently:

$$Var(X) = E(X^2) + E(E(X)^2) - E(2XE(X))$$

$$Var(X) = E(X^2) + E(X)^2 - 2E(X)E(X)$$

$$Var(X) = E(X^2) - E(X)^2$$

$E(X)$ is often written as μ.

The standard deviation of a random variable X is given by

$$\sigma_X = \sqrt{Var(X)}$$

The variance satisfies the following relationships

1. $Var(cX) = c^2 Var(X)$
2. $Var(X + c) = Var(X)$
3. $Var(X + Y) = Var(X) + Var(Y)$ (if X and Y are independent)

3.3 Continuous random variables

Consider another experiment where you record the decision times on a grammaticality judgment task: how much time (in milliseconds) does it take to decide whether the sentence is grammatical or not?

Suppose we define a random variable X which takes decision times as its values.

The variable X cannot take its values from a countable list; X is a continuous random variable and can take any value in the continuous space $X \geq 0$.

It is impossible to determine the probability of a specific value of X, so we cannot assign probabilities like $P(X = x_i)$.

We can however assign probability to an interval of values of X. For example, we can determine probability of obtaining a value between x_1 and x_2, i.e., $P(x_1 \leq X \leq x_2)$ or, equivalently, $P(x_1 < X < x_2)$.

A continuous random variable X is associated with a probability density function $f(x)$, which assigns probabilities over an interval of values of X in the following way:

(a) $\int_{-\infty}^{\infty} f(x)\, dx = 1$ where $f(x) \geq 0$, and $-\infty \leq x \leq \infty$

(b) For any x_1, x_2 such that $-\infty < x_1 < x_2 < \infty$

$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\,dx$

We can also define a cumulative distribution function F(x) such that

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

Expected value and variance of a Continuous random variable

The expected value or the mean of a continuous random variable X is given by

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

The variance is given by

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

The normal distribution

Think about the distribution of heights in a population.

Suppose the average height of the population is 6 feet, and the number of people with height > 6 and < 6 is almost same.

Consider an experiment where you randomly pick an individual from the population and record their height. Let us say we define a random variable X that takes the recorded height as its value.

The variable X is a continuous random variable with specific properties. For example, it is symmetrically distributed around its mean, i.e., $P(X < E(X)) \approx P(X > E(X))$.

The distribution of variable X in this example can be characterized by a normal distribution with the following probability density function:

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

such that:

- $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- $\int_{-\infty}^{\infty} xf(x)\,dx = \mu$
- $\int_{-\infty}^{\infty} x^2 f(x)\,dx = \sigma^2$

3.4 *Some important probability distributions*

	Type of Random variable	**Name of the distribution**	**Probability density function (PDF)** or **Probability mass function (PMF)**
1	Discrete	Binomial	PMF: $f(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^k$
2	Discrete	Poisson	PMF: $f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ where $\lambda > 0$
3	Continuous	Normal	PDF: $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
4	Continuous	Beta	PDF: $f(x; \alpha, \beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$ (where $\alpha, \beta > 0$)
5	Continuous	Gamma	PDF: $f(x; \alpha, \beta) = \frac{\beta^\alpha}{(\alpha-1)!} x^{\alpha-1} e^{-\beta x}$ (where $\alpha, \beta > 0$)

4 Conditional probability and Bayes' theorem

Let us look at some useful results and properties that emerge from the three axioms of probability.

4.1 Conditional probability

The probability of occurrence of an event A given that another event B has already occurred is called the **conditional probability** of A given B, and it is denoted by $P(A|B)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ given that } P(B) \neq 0$$

Let us verify the above relationship using an example.

Suppose you toss two fair coins simultaneously. The sample space would be $\Omega = \{HH, HT, TH, TT\}$. Consider two events A and B.

A : both the coins show heads

B : at least one coin show heads

What is the probability of occurrence of B given that A has occurred?

It will be equal to probability of A such that A is an event in the sample space B, $A \subseteq B$, where

$B = \{HH, HT, TH\}$

$A = \{HH\}$

Given that the coins were fair.

$P(HH) = P(HT) = P(TH)$

Consider B as the sample space, from the second and the third axoim we can deduce that,

$P(HH) + P(HT) + P(TH) = 1$

So,

$P(HH) = P(HT) = P(TH) = \frac{1}{3}$

Hence,

$$P(\{HH\}|B) = \frac{1}{3}$$

$$P(A|B) = \frac{1}{3}$$

What is the probability of an event $A \cap B$ in the sample space Ω?

$A \cap B = \{HH\}$

For the sample space Ω, $P(HH) + P(HT) + P(TH) + P(TT) = 1$

so, $P(HH) = P(HT) = P(TH) = P(TT) = 1/4$, which implies that $P(A \cap B) = \frac{1}{4}$

and,

$P(\{HH, HT, TH\}) = P(HH) + P(HT) + P(TT) = \frac{3}{4}$

hence, $P(B) = \frac{3}{4}$

Finally,

$$\frac{P(A \cap B)}{P(B)} = \frac{1}{3} = P(A|B)$$

4.2 *Independent events*

Two events A and B are said to be independent if the occurrence of one does not affect the (probability or odds of) occurrence of the other.

The above statement implies that $P(A|B) = P(A)$ (and also, $P(B|A) = P(B)$). The following relationship is satisfied

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

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

The above result implies that two events A and B are independent if and only if the the probability of joint occurrence of A and B is equal to the product of their probabilites.

The term $P(A \cap B)$ gives the probability that both events A and B occur, it is called the joint probability and also represented by $P(A, B)$.

Generally, n events E_1, E_2, ..., E_n are independent if and only if $P(E_1, E_2, E_3, \ldots E_n) = P(E_1)P(E_2)P(E_3)\ldots P(E_n)$.

4.3 *Total probability*

Suppose n mutually exclusive events $A_1, A_2, A_3, \ldots, A_n$ occur in an event space F, such that
$\cap_{i=1}^{n} A_i = \emptyset$ and $\cup_{i=1}^{n} A_i = S$
For another event B in F, ($B \subseteq F$),
we can say that $B \cap A_1$ and $B \cap A_2$ are mutually exclusive. So,

$$P((B \cap A_1) \cup (B \cap A_2) \cup \ldots) = P(B \cap A_1) + P(B \cap A_2) + \ldots$$

From set theory we know that, $(B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup \ldots = B \cup (A_1 \cap A_2 \cap A_3 \cap \ldots)$.

$$P(B \cup (\cap_{i=1}^{n} A_i)) = \sum_{i=1}^{n} P(B \cap A_i)$$

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

We know that $P(B \cap A_i) = P(B|A_i)P(A_i)$. Hence,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{17}$$

The above relationship is called the law of total probability.

4.4 *Bayes' theorem*

Suppose two mututally exclusive and exhaustive events A_1 and A_2 occur in an event space F such that
$A_1 \cup A_2 = S$ and $A_1 \cap A_2 = \emptyset$
For an event B in F we can say that:

$$P(B \cap A_1) = P(B|A_1)P(A_1) = P(A_1|B)P(B)$$

Similarly:

$$P(B \cap A_2) = P(B|A_2)P(A_2) = P(A_2|B)P(B)$$

From the above equations we can derive the following:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)}$$

And, from the law of total probability we know that,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$$

Hence,

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}$$

The above equation is Bayes' rule.

Let us talk about the variables that assign values to the outcomes of an underlying generative process (random event).

5 *Using Bayes' theorem for statistical inference*

Suppose that an outcome x observed in an experiment is assumed to come from a normal distribution, such that

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

where $f(x)$ is the probability density function; $f(x)$ assigns the probability density value to the outcome x conditional on the parameters mean μ and variance σ^2 of the normal distribution. The probability density of x conditional on μ and σ^2 can be written as,

$p(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

The goal of statistical inference is figure out what value(s) of μ and σ^2 have generated the observed outcome x.

We know the probability density of obtaining x given μ and σ^2, can we calculate the probability density of (a range of) values μ and σ^2 conditional on the observed outcome x?

$p(\mu, \sigma^2|x) =$?

Using Bayes' theorem,

$p(\mu, \sigma^2|x) = \frac{p(x|\mu,\sigma^2) \cdot p(\mu,\sigma^2)}{\int\int p(x|\mu,\sigma^2) \cdot p(\mu,\sigma^2)\, d\mu\, d\sigma^2}$

More generally, suppose the observed outcome x is assumed to be a value of the random variable X whose probability density function is $f(x;\theta)$; $f(x;\theta)$ assigns a probability density value to x conditional on a parameter θ. The probability density of x given the parameter θ is given by $p(x|\theta)$.

Our goal is to infer what value(s) of the parameter θ has generated the given (observed) datapoint x.

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x|\theta) \cdot p(\theta)\, d\theta} \tag{18}$$

The term $p(x|\theta)$ is called the **likelihood function**, $p(\theta)$ is called the **prior distribution** of θ, and $p(\theta|x)$ is called the **posterior distribution** of θ.

Note: When $f(x;\theta)$ is seen as a function of x, it is called a probability density function; and when $f(x;\theta)$ is seen as a function of θ, it is called a likelihood function, also denoted by $\mathcal{L}(\theta|x)$.