Basic concepts from Probability Theory

MDT - Master Digital Transformation

May 29, 2020

Random experiments

The formal language of uncertainty which is the basis of statistical inference comes from probability theory. To introduce some basic facts about probability, we firstly need to introduce the notion of random experiment.

A **random experiment** is an experiment for which we know what outcomes could happen, but we don't know which particular outcome will happen.

Some basic examples are:

- rolling a die,
- tossing a coin,
- picking at random a ball from an urn containing balls of different colours.

Sample space

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

- In rolling a die
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
- In tossing a coin, letting H=head and T=tail,
$$\Omega = \{H, T\}$$
- In tossing a coin twice
$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

Events

The elements of Ω are called **sample points** or **elementary events**.

An **event** is a set of elementary events, that is a subset of the sample space.

The event A *occurs* if the outcome of the experiment is an element of A.

Example. A die is rolled once. Let A and B be the events of respectively obtaining an odd number, and a number greater than 4. Then

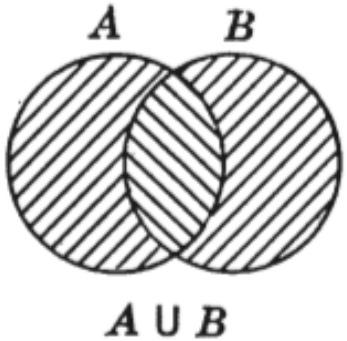
$$A = \{1, 3, 5\}, \quad \text{and} \quad B = \{5, 6\}.$$

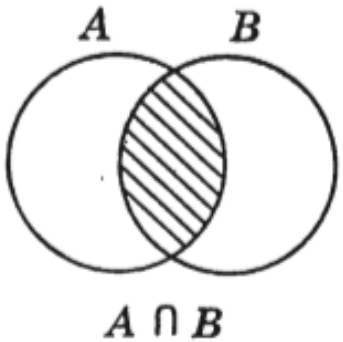
Example. A coin is tossed twice. Let A and B be the events of respectively obtaining a head in the first toss, and exactly one head. Then

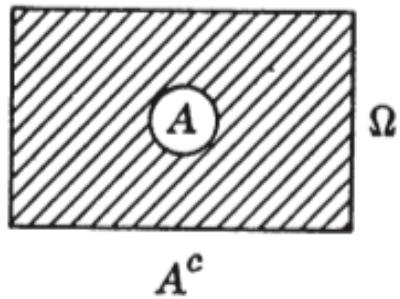
$$A = \{(H, T), (H, H)\} \quad \text{and} \quad B = \{(H, T), (T, H)\}.$$

Some set operations

Events are sets . . . they may be combined according to the usual **set operations**!!







Events and set operations I

- The **complement** of an event A is the event A^c which occurs when A does not occur.

  Example: Roll a die, and let A be the event of obtaining an even number. Then

  $$A = \{2, 4, 6\}, \quad \text{and} \quad A^c = \{1, 3, 5\}.$$

- The **union** between events A and B is the event $A \cup B$ which occurs when A or B (or possibly both) occur.

  Example. Roll a die, and let A and B be the events of respectively obtaining an odd number, and a number greater than 4. Then

  $$A = \{1, 3, 5\}, \quad B = \{5, 6\}, \quad \text{and} \quad A \cup B = \{1, 3, 5, 6\}.$$

Events and set operations II

- The **intersection** between events A and B is the event $A \cap B$ which occurs when both A and B occur.

  Example: Roll a die, and let A and B be the events of respectively obtaining a number smaller than 4, and an even number.Then

$$A = \{1, 2, 3\}, \quad B = \{2, 4, 6\} \text{ and } \quad A \cap B = \{2\}.$$

  A and B are **mutually exclusive**, if they have no outcomes in common.

  Example: Roll a die, and let A and B be the events of respectively obtaining an odd number, and an even number. Then

$$A = \{1, 3, 5\}, \quad B = \{2, 4, 6\} \text{ and } \quad A \cap B = \emptyset.$$

Assessing probability I

We assign a probability $P(A)$ to an event A to measure *how likely* the event is.

Classical probability: if Ω is finite, assuming that all outcomes in Ω are *equally likely*, the probability of an event A is

$$P(A) = \frac{\text{number of outcomes favourable to } A}{\text{number of outcomes in } \Omega}$$

Example. If two fair coins are tossed, the probability of getting exactly one head is $1/2$.

Example. If a card is drawn from a deck of 52 playing cards, the probability of obtaining a spade is $1/4$.

Assessing probability II

Frequentist probability: the probability of A is the relative frequency of occurrence of A, in a large number of *repetitions* of the experiment under the *same conditions*.

Example. In tossing a fair coin n times, with n *large*, the relative frequency of heads approaches to $1/2$.

Example. If a card is drawn from a perfectly shuffled deck of 52 cards, then the card is replaced, the deck reshuffled, and the experiment is repeated over and over again, there is *convergence* of relative frequency of spade to $1/4$.

Subjective probability: the probability of an event is an individual's personal judgement about whether the event is likely to occur.

Example. You think you have an 80% chance of your best friend calling today, because her car broke down yesterday and she'll probably need a ride!

Assessing probability III

The **axiomatic perspective** is a unifying one which says that probability is a function P which satisfies the following *axioms*:

- for any event A, $P(A) \geq 0$,
- $P(\Omega) = 1$,
- if A and B are mutually exclusive events, then

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

From the above axioms, one can derive several *properties* of probability. For example, for any event A, it holds that $P(A^c) = 1 - P(A)$, and $P(\emptyset) = 0$.

Conditional probability I

As you obtain additional information, how should you update probabilities of events?

Example. Suppose that in a certain city, 20% of the days are rainy. Thus, if you pick a random a day, the probability that it rains is $P(R) = 0.2$, with R being the event that it rains on the chosen day.

Now if I tell you that it is cloudy on the chosen day, how do you update this probability? If C is the event that it is cloudy, then we write the probability that it rains given that it is cloudy as $P(R \mid C)$, that is the *conditional probability* of R given that C has occurred.

Conditional probability II

If A and B are events in Ω, and $P(B) > 0$, then the **conditional probability** of A given B is

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

If we know that B has occurred, every outcome that is outside B should be *discarded*. Thus, the sample space reduces to B.

Example. Toss a die. Let A and B be the events of respectively getting a number less than or equal to 3, and an odd number. Then $P(B \mid A) = 2/3$.

Example. Throw two dice. Let A be the events that the sum of the faces is 8, and let B be the event that the faces are equal. Then $P(B \mid A) = 1/5$.

Independent events I

Example. Let A be the event that it rains tomorrow. Now, assume that I toss a coin, and let B be the event that the result is a tail. What is $P(A \mid B)$? Tomorrow's weather is not influenced by whether or not B occurred!! Thus, no matter if B, the probability of A should not change, i.e. $P(A \mid B) = P(A)$. These are *independent* events.

Formally, two events A and B are said to be **independent** if

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

Recalling conditional probability definition, this is equivalent to

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

(or $P(B \mid A) = P(B)$). Independence is a *symmetric relation*.

Independent events II

Example. Let two dice be tossed. Let A be the event of obtaining a number less than or equal to 3 at the first die, let B be the event that the sum of the two faces is 9, and let C be the event that the sum of the two faces is 7. Then

$$P(A \cap B) = \frac{1}{36} \neq P(A)P(B) = \frac{1}{2}\frac{4}{36} = \frac{1}{18},$$

and

$$P(A \cap C) = \frac{1}{12} = P(A)P(C) = \frac{1}{2}\frac{1}{6}.$$

Remark. Suppose that A and B are mutually exclusive, each with positive probability. Can they be independent? NO! This follows since $P(A)P(B) > 0$ yet $P(A \cap B) = 0$. Except in this special case, there is no way to judge independence by looking at the sets in a Venn diagram.

Test for a disease

A medical test for a disease D has outcomes $+$ and $-$. The probabilities are:

	D	D^c
$+$	0.009	0.099
$-$	0 .001	0.891

From the definition of conditional probability,

$$P(+ \mid D) = \frac{P(D \cap +)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

and

$$P(- \mid D^c) = \frac{P(D^c \cap -)}{P(D^c)} = \frac{0.891}{0.891 + 0.099} \simeq 0.9$$

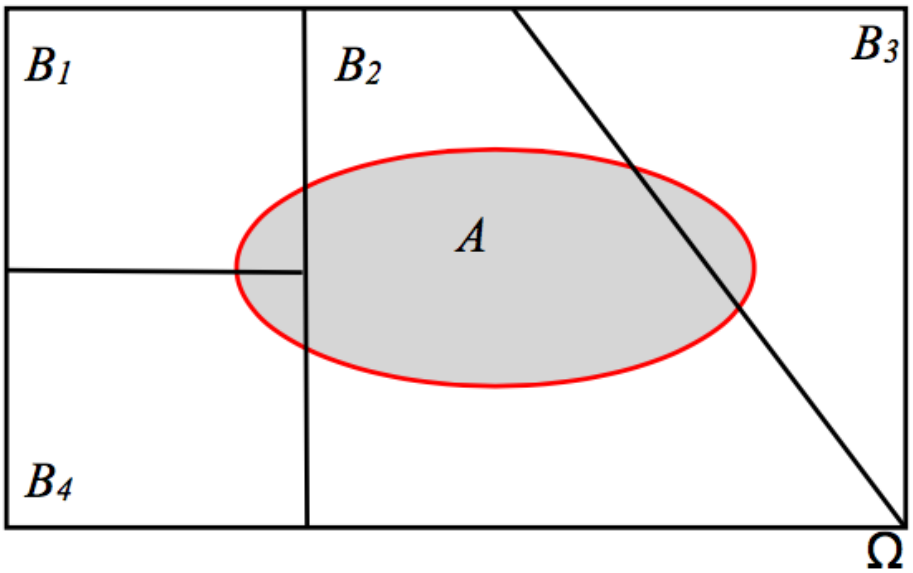
Apparently, the test is fairly accurate. Sick people yield a positive 90 percent of the time and healthy people yield a negative about 90 percent of the time. Suppose you go for a test and get a positive. What is the probability you have the disease? Most people answer 0.9. The correct answer is

$$P(D \mid +) = \frac{P(D \cap +)}{P(+)} = \frac{0.009}{0.009 + 0.099} \simeq 0.08.$$

Law of total probability

If events $B_1, B_2, \ldots, B_k$ form a *partition* of Ω, then

$$P(A) = \sum_{j=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A \mid B_j) P(B_j).$$



Example. Consider 3 bags: bag 1 has 75 red and 25 blue marbles, bag 2 has 60 red and 40 blue marbles, bag 3 has 45 red and 55 blue marbles. You choose one of the bags at random and then pick at random a marble from the chosen bag. What is the probability that the chosen marble is red?

Bayes' rule

Bayes' rule describes the probability of an event, based on prior knowledge of conditions that might be related to the event. If $B_1, B_2, \ldots, B_k$ form a partition of Ω and A is any event with $P(A) > 0$

$$P(B_j \mid A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(B_j)P(A \mid B_j)}{\sum_{j=1}^{k} P(B_j)P(A \mid B_j)}.$$

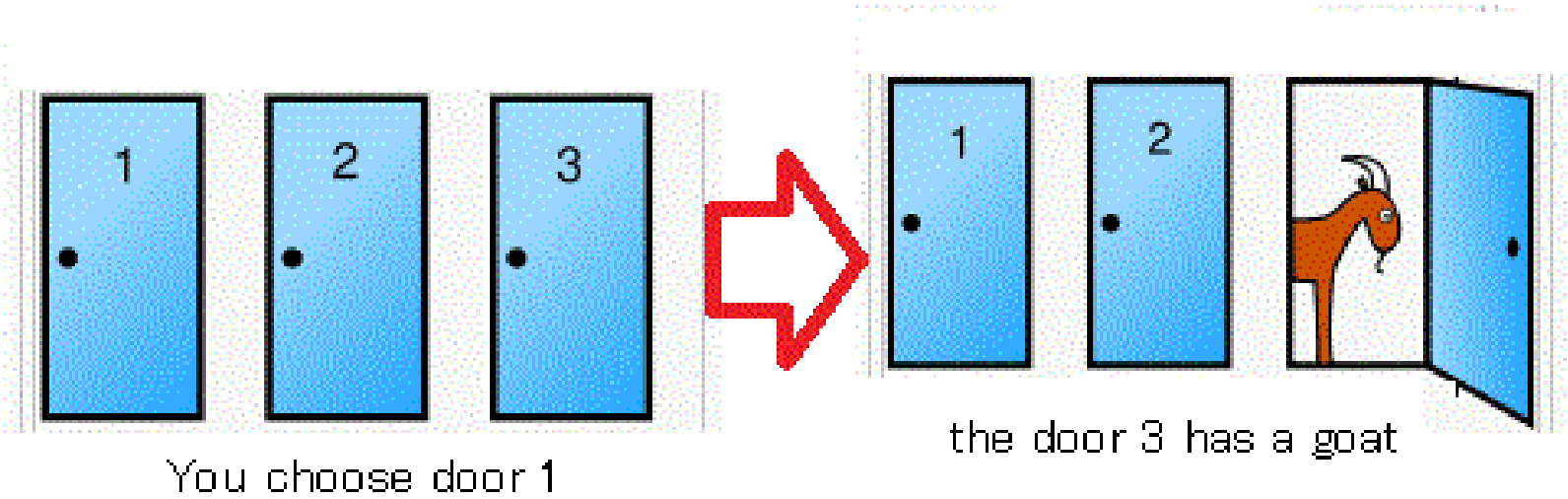
$P(B_j)$ is the **prior probability** of B_j, $P(B_j \mid A)$ is the **posterior probability** of B_j.

The Monty Hall Problem I

The Monty Hall problem relies on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall.

Suppose you're on the game show, and you're given the choice of 3 doors: behind one door is a car; behind the others, goats. You pick a door, say 1, and Monty Hall, who knows what is behind the doors, opens another door which has a goat. He then gives you the opportunity to keep your door or switch to the other unopened door.

Should you stay or switch?



The Monty Hall Problem II

The correct answer is that you should switch!!

There are three possible arrangements of one car and two goats behind three doors, and different results of staying or switching in each case.

Door 1	Door 2	Door 3	If you stay	If you switch
Goat	Goat	Car	You win goat	You win car
Goat	Car	Goat	You win goat	You win car
Car	Goat	Goat	You win car	You win goat

Monty Hall and Bayes' rule

Letting B_j be the event that the car is behind door j, and letting A_j be the event that Monty shows the door j. If you choose door 1, then

$$\begin{aligned} P(A_2) &= P(B_1 \cap A_2) + P(B_2 \cap A_2) + P(B_3 \cap A_2) \\ &= P(B_1)P(A_2 \mid B_1) + P(B_2)P(A_2 \mid B_2) + P(B_3)P(A_2 \mid B_3) \\ &= \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 + \frac{1}{3} \times 1 = \frac{1}{2}, \end{aligned}$$

and

$$P(B_1 \mid A_2) = \frac{P(B_1)P(A_2 \mid B_1)}{P(A_2)} = \frac{1/3 \times 1/2}{1/2} = \frac{1}{3}.$$

Random variables I

Statistics and data mining are concerned with data. How do we link sample spaces and events to data? The link is provided by the concept of a random variable.

Example. In an opinion pool, we decide to ask 50 people if they agree or not with a certain issue. We record 1 for agree and 0 for disagree, then the sample space has 2^{50} elements. We are interested in the number X of people which agree out of 50, then X counts the number of 1s and the corresponding sample space is $\{0, 1, \ldots, 50\}$. X is an example of a *random variable*.

Formally, a **random variable** maps any outcome in Ω to a real number. In this way the description of an experiment can be made in terms of values of random variables.

For each value or a set of values of the random variable, there are underlying collections of events, and through these events one connects the values of random variables with probability measures.

According to the set on which a random variable takes values, one can distinguish discrete and continuous random variables.

Discete random variables

A random variable X is said to be **discrete** if it can assume only a finite or countably infinite number of distinct values.

Example. Suppose an Internet business firm had 1000 hits on a particular day. Let the random variable X be defined as the number of sales resulted on that day. Then, X can take values $0, 1, \ldots, 1000$.

Example. If we are to define a random variable as the number of phone calls made from a big city in the next 24 hours, this take values $0, 1, \ldots$.

If X is a discrete random variable, the **probability mass function** associates a probability to each possible value of X.

Example. Flip a fair coin twice and let X be the number of heads. Then $P(X = 0) = P(X = 2) = 1/4$ and $P(X = 1) = 1/2$.

Bernoulli distribution I

Bernoulli distribution models situations where there are two possible outcomes, a *success* and a *failure*. According to this model a variable X takes value 1 if the success occurs, with probability p and 0 otherwise, with probability $1 - p$.

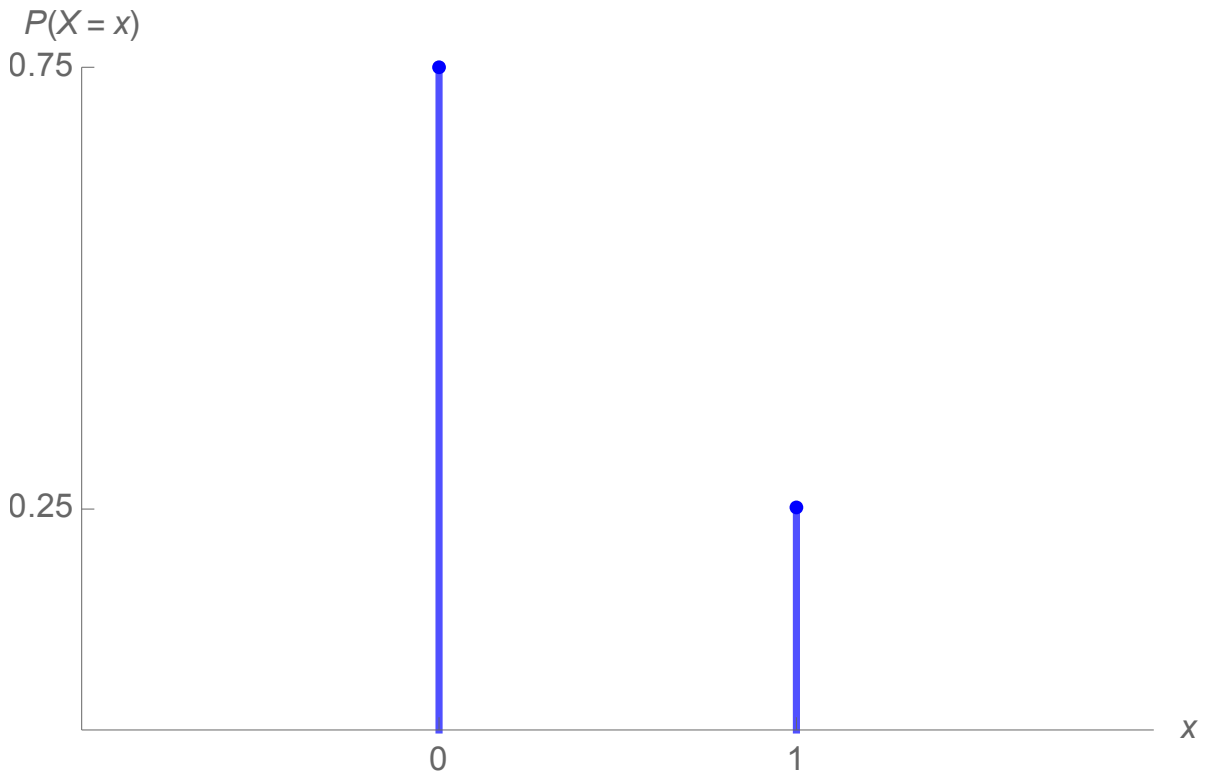
p is the **parameter** which identifies the distribution.

Examples are:

- toss a coin with success=head and failure=tail;
- roll a die, success=odd number, failure=even number;
- examine a component produced by an assembly line with success=acceptable, failure=defective;
- transmit a binary digits by a communication channel: success=digit received correctly, failure= digit received uncorrectly.

Bernoulli distribution II

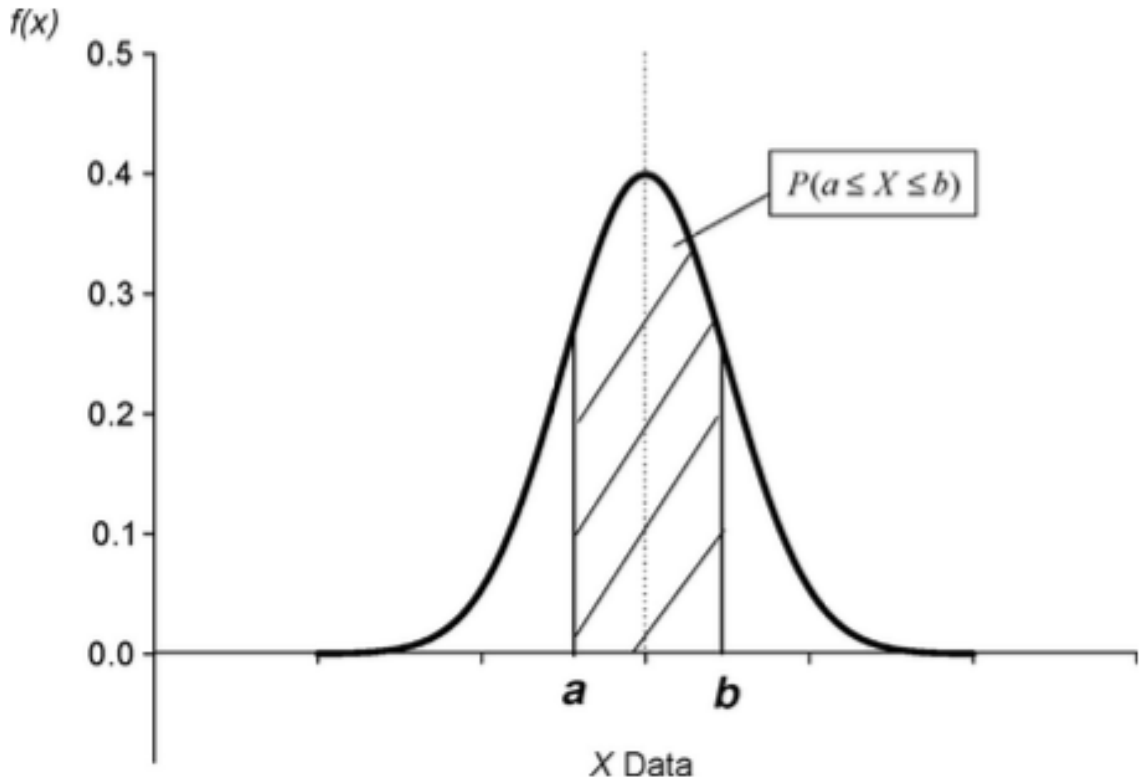
Probability mass function of a Bernoulli distribution with $p = 0.25$.



Continuous random variables

A **continuous** random variable is a random variable which assumes uncountably many values.

Probabilities are assigned using the **probability density** function which is a continuous non-negative function. Specifically for a random variable X with density f, the probability that X assumes values in a continuous set $[a, b]$ is given by the area underlying f on $[a, b]$.



Normal distribution I

The Normal distribution plays an important role in probability and statistics. Many phenomena in nature have approximately Normal distributions, some example are

- blood pressure
- measurement error
- IQ scores.

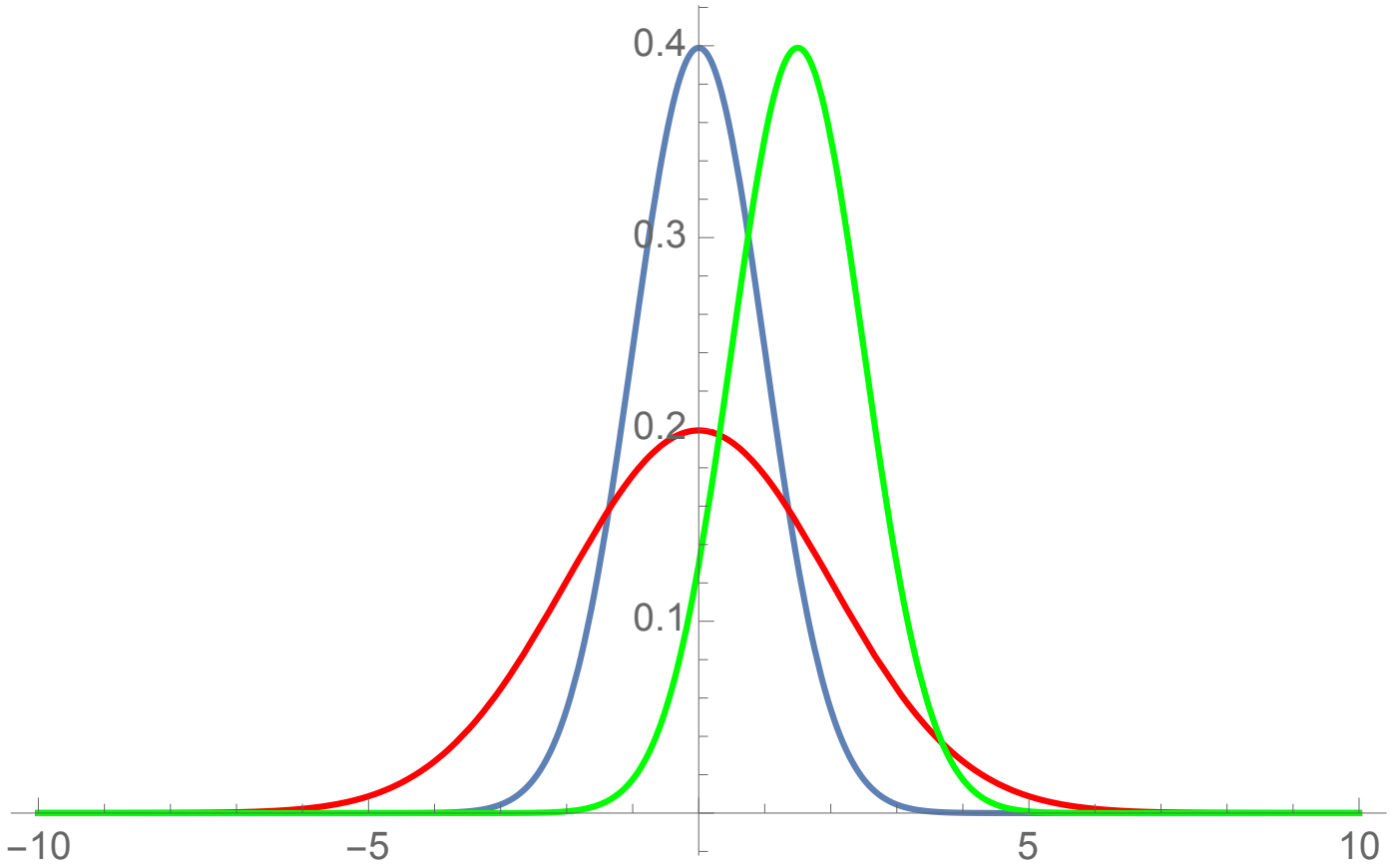
Further, many distributions *converges* to the Normal one.

It is a symmetric distribution with more likely values around the center, and extreme values in both tails which are similarly unlikely.

The parameters characterizing the distribution are μ which is the "center" (mean), and σ^2 which is the "spread" (variance). If X has normal distribution with mean μ and variance σ^2, we write $X \sim N(\mu, \sigma^2)$.

Normal distribution II

Density functions of $N(0,1)$ (blue), $N(1.5,1)$ (green), $N(0,4)$ (red).



Expectation and variance

The expectation (or expected value) of a random variable X is the *average* value of X. It is a one-number summary of the distribution and it is denoted as $E[X]$.

For example, if X is a discrete random variable taking values $x_1, \ldots, x_k$, then

$$E[X] = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_k P(X = x_k).$$

The variance measures the "spread" of the distribution, and it is defined as

$$V[X] = E[(X - E[X])^2].$$

For example, if X is a discrete random variable taking values $x_1, \ldots, x_k$

$$V[X] = (x_1 - E[X])^2 P(X = x_1) + \ldots + (x_k - E[X])^2 P(X = x_k).$$

If $X \sim N(\mu, \sigma^2)$, $E[X] = \mu$ and $V[X] = \sigma^2$.

Bivariate distribution I

Given a pair of discrete random variables X and Y, the joint mass function is defined by $P(X = x \text{ and } Y = y)$ and denoted as $P(X = x, Y = y)$.

Example. The joint distribution of two discrete random variables X and Y each taking values 0 or 1 is, for example

	Y=0	Y=1	
X=0	1/9	3/9	4/9
X=1	1/9	4/9	5/9
	2/9	7/9	1

From the joint distribution, one can obtain the **marginal** ones.

Example. From the above example, one can obtain the marginal distributions

X	$P(X = x)$
0	4/9
1	5/9
	1

Y	$P(Y = y)$
0	2/9
1	7/9
	1

Bivariate distribution II

In the continuous case, we have a joint density function to assign probabilities such has $P(X \in (a, b), Y \in (c, d))$.
Also in this case one can define marginal densities for X and Y.

Two random variables X and Y are **independent** if, for every A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any set A, B.

If X and Y are independent the joint mass (or density) function is equal to the product of the marginal ones.

Conditional distribution

If X and Y are discrete, we can compute the conditional distribution of Y given that we have observed $X = x$. Specifically, the conditional probability mass function of $Y \mid X = x$ is defined as

$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}.$$

Example. In the previous example, the distribution of Y given $X = 0$ is

Y	$P(Y = y \mid X = 0)$
0	1/4
1	3/4

Similarly, for the continuous case, one can define the conditional density function of Y given $X = x$.

The conditional expectation and the conditional variance of Y given $X = x$ can be obtained as before but by substituting the conditional mass function (or conditional density function) in the definition of expectation and variance, respectively.

Multivariate distributions and iid variables

Let $X_1, \ldots, X_n$ be random variables. Given a joint distribution for $X_1, \ldots, X_n$, it is possible to define their marginals, conditionals etc. in the same way as in the bivariate case. We say that $X_1, \ldots, X_n$ are independent if, for every $A_1, A_2, \ldots, A_n$,

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \ldots P(X_n \in A_n).$$

If $X_1, \ldots, X_n$ are independent and each has the same marginal distribution, then $X_1, \ldots, X_n$ are said to be iid (independent and identically distributed).