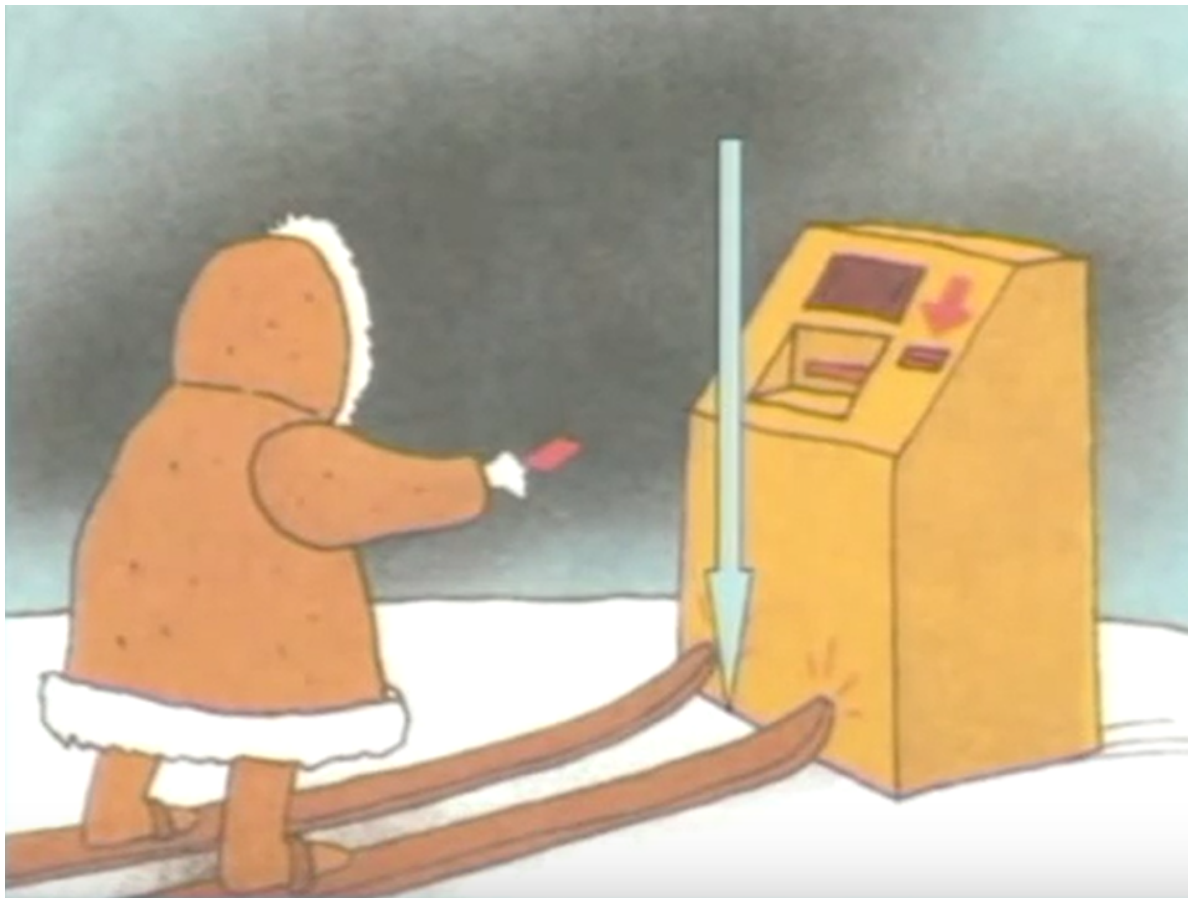
Information Theory
Degree in Data Science and Engineering
Lesson 1: Discrete random variables

Jordi Quer, Josep Vidal

Mathematics Department, Signal Theory and Communications Department
{jordi.quer, josep.vidal}@upc.edu

2019/20 - Q1

Why probability?



Why probability?

Letters in an English text appear with different *frequencies*:

- e is the most frequent: 12.702%
- t is the second most frequent: 9.056%
- z is the less frequent: 0.074%

The same happens with words: the 10 most frequent words in English are, in the given order:

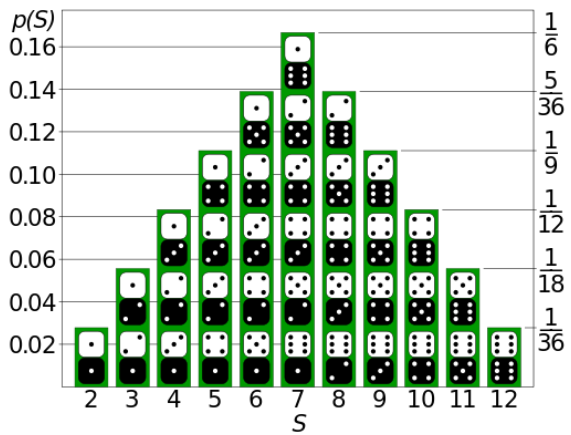
the, be, to, of, and, a, in, that, have, I.

Also for: instructions in a computer program, pixels in an image, samples in a digitized wave sound, nucleotides in a DNA sequence, data content in a database, weather forecast, etc.

Probability

Probability provides a way to model random phenomena: random experiments are associated to a value within a set of possible outcomes, each occurring with some probability. For example:

- *Flip a coin*. Two possible outcomes: heads and tails, each with the same probability $1/2$.
- *Throw a dice*. Six possible outcomes: 1 to 6 dots, each with the same probability $1/6$.
- Throw two dice and take the sum. Eleven possible outcomes: $S = \{2, \ldots, 12\}$ with probabilities shown in the table at the right:



Probabilities are numbers in $[0, 1]$ and their sum must be $= 1$.

Random variables

Mathematical formalization in terms of *random variables*: for us, a discrete random variable X consists of

- a discrete set $\mathcal{X} = \{x_1, x_2, \ldots, x_q\}$ of possible values x_i,
- each occurring with a given probability

$$p_i = p(x_i) = \Pr(X = x_i) \in [0, 1],$$

- and $\sum_{i=1}^{q} p(x_i) = 1$.

If x_i are numbers, the *expected value* (or expectation) of X is:

$$\mu_X = \mathbb{E}[X] = \sum_{i=1}^{q} x_i \, p(x_i)$$

and the variance of X is

$$Var(X) = \sigma_X^2 = \mathbb{E}\left[|X - \mathbb{E}[X]|^2\right] = \sum_{i=1}^{q} |x_i - \mathbb{E}[X]|^2 \, p(x_i)$$

Random variables: examples

- *Bernoulli distribution*. X has two outcomes: $\mathcal{X} = \{1, 0\}$ with probabilities $p(1) = p \in [0, 1]$ and $p(0) = 1 - p$.

$$\Pr(X = x) = p(x) = p^x (1 - p)^{(1-x)}$$

  When flipping a coin $\mathcal{X} = \{\text{head}, \text{tail}\}$. For a fair coin $p = 1/2$.

- *Uniform distribution*. Y has n outcomes: $\mathcal{Y} = \{y_1, \ldots, y_q\}$, each with the same probability $p(y_i) = 1/q$. Throwing a dice corresponds to $q = 6$.

- *Binomial distribution*. Z has $n + 1$ possible outcomes: $\mathcal{Z} = \{0, 1, \ldots, n\}$, with probabilities:

$$\Pr(Z = z) = p(z) = \binom{n}{z} p^k (1 - p)^{n-z},$$

  that counts the number of 1's in n independent samples of a random variable X with Bernoulli distribution.
  If p is the probability of heads in a coin flipping, the average number of heads when flipping n coins is $E[z] = pn$.

A mathematical model for Information Theory

In *information theory*, a data source is a device that produces letters or symbols that can be considered as observations of a random variable X taking values in a finite alphabet $\mathcal{X}, (|\mathcal{X}| = q)$ with certain probabilities.

Concrete values x_i are not important. The relevant information about the variable X is given by the *probability distribution* given by q numbers

$$p_1, p_2, \ldots, p_q \qquad \text{with} \qquad p_i \in [0,1] \qquad \text{and} \qquad \sum_{i=1}^{q} p_i = 1.$$

When X is sampled several consecutive times, it produces a text: a string of letters of the alphabet.

We shall see more sophisticated models that consider strings produced a sequence $X_1 X_2 X_3 \ldots$ of non-independent random variables.

Pairs of random variables

One may consider two or more random variables X and Y associated to the same experiment. For example:

The experiment consists of rolling two different dice, and

- X and Y are the outcomes of each dice;
- X is the outcome of the first dice and Y is the sum;
- X is the sum and Y says whether the two outcomes are equal or different;
- X is the sum and Y is the parity of the second dice; etc.

In each situation there are several relevant probability distributions.
Let $\mathcal{X} = \{x_1, \ldots, x_q\}$ and $\mathcal{Y} = \{y_1, \ldots, y_r\}$ be their values.

Joint and marginal probabilities

Everything is governed by the *joint probability distribution*:

$$p(x_i, y_j) = \Pr(X = x_i, Y = y_j),$$

which satisfies:

$$\sum_{i=1}^{q} \sum_{j=1}^{r} p(x_i, y_j) = 1.$$

The probabilities of each separate variables are called *marginal probabilities*. They are computed from the joint distribution as:

$$p(x_i) = \Pr(X = x_i) = \sum_{j=1}^{r} p(x_i, y_j),$$

$$p(y_j) = \Pr(Y = y_j) = \sum_{i=1}^{q} p(x_i, y_j).$$

Conditional probabilities

Conditional probabilities can be defined as

$$p(x_i|y_j) = \Pr(X = x_i|Y = y_j) := \frac{p(x_i, y_j)}{p(y_j)}, \quad \text{if} \quad p(y_j) \neq 0,$$

to be read as the probability of x_i given y_j or under the condition y_j.

Must be interpreted as the probability that, in a sample of the pair of variables, X takes the value x_i under the condition that Y takes the value y_j.
Analogously,

$$p(y_j|x_i) = \Pr(Y = y_j|X = x_i) := \frac{p(x_i, y_j)}{p(x_i)}, \quad \text{if} \quad p(x_i) \neq 0,$$

Marginal and conditional probability distributions

Theorem

The marginal probabilities defined from a joint probability distribution satisfy:

$$\sum_{i=1}^{q} p(x_i) = \sum_{i=1}^{q} \sum_{j=1}^{q} p(x_i, y_j) = 1, \qquad \sum_{j=1}^{q} p(y_j) = 1$$

Theorem

The conditional probabilities satisfy:

$$\forall j \quad \sum_{i=1}^{q} p(x_i|y_j) = 1, \qquad \forall i \quad \sum_{j=1}^{r} p(y_j|x_i) = 1$$

Independent variables

Two variables X and Y are *independent* if the joint probability is always the product of the marginal probabilities:

$$p(x_i, y_j) = p(x_i)p(y_j) \quad \forall i, j.$$

This property is derived from the fact that the conditional probabilities do not really depend on the condition:

$$p(x_i|y_j) = p(x_i) \quad \text{and} \quad p(y_j|x_i) = p(y_j) \quad \forall i, j.$$

The dependency measures the degree of correlation between the two variables.

Pairs of random variables: examples

Experiment: Let X and Y be the outcomes of rolling two different dices, $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$.

$$p(x_i, y_j) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p(x_i)p(y_j) \quad \forall\, i, j.$$

They are independent, and their conditional distributions are:

$p(y_j \mid x_i)$	1	2	3	4	5	6
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
3	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
4	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
5	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
6	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$p(x_j \mid y_i)$	1	2	3	4	5	6
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
3	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
4	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
5	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
6	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Pairs of random variables: examples

Let S be the sum of the two dices, $\mathcal{S} = \{2, 3, \ldots, 12\}$.

The joint distribution of X and S is:

$p(x_i, s_j)$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{6}$
2	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	0	$\frac{1}{6}$
3	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	0	$\frac{1}{6}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$p(s_j)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$	1

The sum of all table entries $= 1$ (joint probability distribution).
The marginal probabilities are the sums of rows and columns.

Pairs of random variables: examples

Conditional distribution of S with respect to X: knowing the first dice, probability of value of the sum.

$p(s_j \mid x_i)$	2	3	4	5	6	7	8	9	10	11	12
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0	0
2	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0
3	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0
4	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0
5	0	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0
6	0	0	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The sum of each row must be $= 1$ (conditional probability distributions).

Pairs of random variables: examples

Conditional distribution of X with respect to S: knowing the sum, probability of value of first dice.

$p(x_j \mid s_i)$	1	2	3	4	5	6
2	1	0	0	0	0	0
3	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0
4	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0
5	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0
6	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0
7	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
8	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
9	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
10	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
11	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
12	0	0	0	0	0	1

Bayes' Theorem

What is the relation between $p(x_i|y_j)$ and $p(y_j|x_i)$? It is given by the *Bayes' theorem*:

$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{p(y_j)}$$

where

$$p(y_j) = \sum_{i=1}^{q} p(x_i, y_j) = \sum_{i=1}^{q} p(y_j|x_i)p(x_i)$$

Example: When getting the result of a medical test on AIDS, what can be said about our condition? Take as random variables:

Patient condition: $X \in \{healthy, sick\}$
Result of the test: $Y \in \{-, +\}$

Bayes' Theorem

At the pharmaceutical lab...

X: healthy

TEST

Y: − − − − − + − − − −

$p(-\,|\,healthy)$
Specificity

X: sick

TEST

Y: + − + + + + + + + −

$p(+\,|\,sick)$
Sensitivity

Me at the doctor's...

TEST

Y: +

$p(x\,|\,+)$?

Use Bayes' theorem and check the dependency with the prevalence $p(sick)$.
Get a result using *AIDS prevalence data in Spain*.