Random variables

Josemari Sarasola

Statistics for Business

Gizapedia

Random variables and probability distributions

A *random variable* (rv) (eusk. *zorizko aldagai*; gazt., *variable aleatoria*) is a variable that takes its values randomly; e.g., the results of throwing a die: 1,2,3,4,5,6.

When we list all the values of a rv along with their probabilities, we say we have set the probability distribution of that rv (eusk., *probabilitate-banakuntza*, gazt., *distribución de probabilidad*).

Random variables and probability distributions

Adibidea: points after throwing a dice,

random variable

x	$p(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$
	1

probability distribution

Discrete rv.s

A discrete rv takes isolated or separate values (the number of children of a couple, e.g.). We can define them in two ways: by means of their probability mass function and by means of their distribution function.

The *probability mass function* gives directly the probability of a particular value. We can define it as a function or as a table. E.g., $P[X = x] = (x + 1)/10; x = 0, 1, 2, 3.$, is a probability mass function that can be expressed also as a table.

For a given probability mass function it holds that the sum of probabilities is 1.

Discrete random variables

Distribution functions, denoted by $F(x)$, gives the probability of being x or less:

$$F(X) = P[X \leq x]$$

We can express them as a function or as a table:

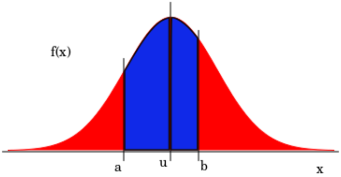
$$F(x) = 1 - (1/2)^x ; x = 1, 2, 3, \dots$$

They hold these properties:

- it's always an increasing function;
- it starts at 0, and ends at 1.

Continuous random variables

They can take any value in an interval (e.g., tenperature). We can define them by means of a density function or a distribution function.



The image above is a *density function* and we express it as $f(x)$. For $f(x)$ functions we calculaste probabilities in this way kalkulatzen da probabilitatea:

$$P[a < X < b] = \int_a^b f(x)dx$$

It's quite intuitive: the bigger is the function, the bigger is the probability. But be careful: probabilities are areas in that interval,

Continuous random variables

Probability for a point is always 0, as points don't have any area, and the infinite points in the interval must share a probability of 1:

$$P[X = x] = 0$$

Density functions must hold these conditions:

(1) area under universe (possible values) is 1:

$$\int_{\Omega} f(x)dx = 1$$

(2) it's always positive (as probabilities must be positive):

$$f(x) \geq 0, \forall x \in \Omega$$

Continuous random variables: distribution function

It gives the probabiity of being less than x (Do we include x? It's trivial, as the probability of a point is 0):

$$F(x) = P[X < x] = \int_{inf}^{x} f(x)dx$$

So, the distribution function gives the cumulated probabilities. It's like the density function but the integral is already made.

*inf: the smallest possible value, sup: the bigger possible value

So, distribution functions give *directly* probability in an interval, in this way:

- $P[X{<}a]{=}F(x{=}a)$
- $P[X{>}b]{=}1{-}P[X{<}b]{=}1{-}F(x{=}b)$
- $P[a{<}X{<}b]{=}P[X{<}b]{-}P[X{<}a]{=}F(x{=}b){-}F(x{=}a)$

Continuous random variables: distribution function

Distribution functions must hold these conditions:

(1) $F(x = inf) = 0$: the probability under inf is 0.

(2) $F(x = sup) = 1$: the probability under sup is 1.

(3) Always increasing, as probability cumulates.

Continuous random variables

Briefly, density functions (pdfs) and distribution functions (cdfs) relate each other in this way:

- to get cdf from pdf, we integrate:

$$F(x) = \int_{inf}^{x} f(x)dx$$

- so, to get pdf from cdf we derivate:

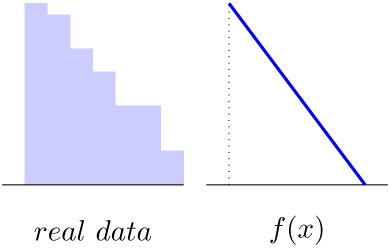
$$f(x) = \frac{dF(x)}{dx}$$

Parameters

- Parameters are constants that take different values in a probability distribution.
- If we don't concrete them, they definite a distribution in a general manner.
- It's usual in statistics to take parameters as unknown, and to estimate them from data.

Distributions as models

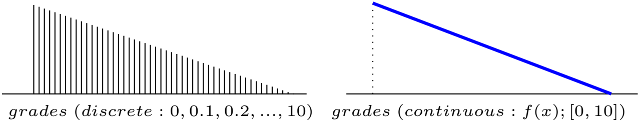
Take this data about a variable and a density function about the same variable.



Density function is "too perfect" to be true, specially we compare it to data.
So, are probability distributions realistic?
That's not the point if they realistic or not. We always look to good approximations to data, that is to say, we take probability distributions (pdfs and cdfs) as models or simplified representations of reality.

Continuous approximations

When a discrete rv takes many values, it's better to represent it as a continuous rv:



How do we calculate $P[grade = 5]$?

- on the discrete domain, directly (looking at the height of the bar):

$$P[X = 5] = 0.04$$

- on the continuous domain, as 5 value is just one point:

$$P[X = 5] = 0$$

- So, there is a lag between discrete values and the continuous interval. How to fix it?

Continuous approximations

In fact, there are students with 5 grade. So, the right one is that from the discrete domain.

How to get the discrete probability from fron the continuous interval?

$$P[X = 5] = P[4.95 < X < 5.05] = \int_{4.95}^{5.05} f(x)dx$$

Actually, grades in the continuous domain round to the nearest discrete value:

- $4.9765 \to 5$
- $5.0398 \to 5$
- baina, $4.9327 \to 4.9$
- baina, $5.0649 \to 5.1$

Jarraituzko hurbilketak

Some other examples

- students that signed up in a course:

$$P[X = 1200] = P[1199.5 < X < 1200.5]$$

- weight of an apple in a balance with a 10 gr precision:

$$P[X = 100] = P[95 < X < 105]$$