Elementary Statistics Course

Alfredo Sánchez Alberca (asalber@ceu.es)

Feb 2016

Department of Applied Math and Statistics
CEU San Pablo

License terms

Contents

Random variables

The process of drawing a sample randomly is a random experiment and any variable measured in the sample is a *random variable* cause the values taken by the variable in the individuals of the sample is a matter of chance.

Definition (Random variable)

A *random variable* X is a function that maps every element of the sample space of a random experiment to a real number.

$$X : \Omega \to \mathbb{R}$$

The set of values that the variable can assume is called range and is represented R.

In essence, a random variable is a variable whose values come from a random experiment, and every value have a probability of occurrence.

Example. The variable X that measures the outcome of rolling a dice is a random variable and its range is

$$\text{Range}(X) = \{1, 2, 3, 4, 5, 6\}$$

Types of random variables

There are two types of random variables:

Discrete They take isolated values, and their range is numerable.
Example. Number of children of a family, number of smoked cigarettes, number of subjects passed, etc.

Continuous They can take any value in an real interval, and their range is non-numerable.
Example. Weight, height, age, cholesterol level, etc.

The way of modelling every type of variable is different. In this chapter we are going to study how to model discrete variables.

Probability distribution of a discrete random variable

As values of a discrete random variable are linked to the elementary events of a random experiment, every value have a probability.

Definition (Probability function)

The *probability function* of a discrete random variable X is the function $f(x)$ that maps every value x_i of the variable to its probability

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

We can also accumulate probabilities the same way that we accumulated sample frequencies.

Definition (Distributin function)

The *distribution function* of a discrete random variable X is the function $F(x)$ that maps every value x_i of the variable to the probability of having a value less than or equal to x_i

$$F(x_i) = P(X \leq x_i) = f(x_1) + \cdots + f(x_i).$$

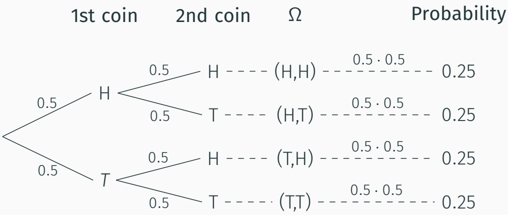
The range of a discrete random variable and its probability function is known as **Probability Distribution** of the variable, and usually it is presented in a table

X	x_1	x_2	$\cdots$	x_n	$\sum$
$f(x)$	$f(x_1)$	$f(x_2)$	$\cdots$	$f(x_n)$	1
$F(x)$	$F(x_1)$	$F(x_2)$	$\cdots$	$F(x_n) = 1$	

The same way that the sample frequency table shows the distribution of values of a variable in the sample, the probability distribution of a discrete random variable shows the distribution of values in the whole population.

Probability distribution of a discrete random variable
Example of tossing two coins

Let X be the discrete random variable that measures the number of heads after tossing two coins. The probability tree of the random experiment is

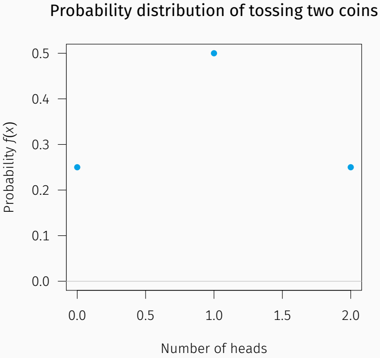


According to this, the probability distribution of X is

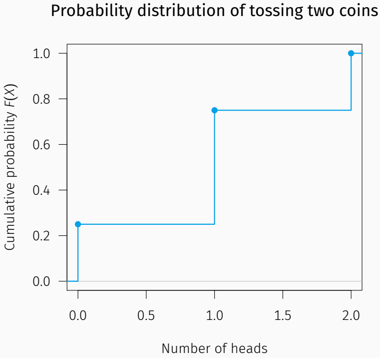
X	0	1	2
$f(x)$	0.25	0.5	0.25
$F(x)$	0.25	0.75	1

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 0.25 & \text{si } 0 \leq x < 1 \\ 0.75 & \text{si } 1 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

Probability function



Distribution function



Population statistics

The same way we use sample statistics to describe the sample frequency distribution of a variable, we use population statistics to describe the probability distribution of a random variable in the whole population.

The population statistics definition is analogous to the to the sample statistics definition, but using probabilities instead of relative frequencies.

The most important are [1]:

- Mean:

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

- Variance:

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2$$

- Standard deviation:

$$\sigma = +\sqrt{\sigma^2}$$

[1] To distinguish population statistics from sample statistics we use Greek letters

In the random experiment of tossing two coins the probability distribution is

X	0	1	2
$f(x)$	0.25	0.5	0.25
$F(x)$	0.25	0.75	1

The main population statistics are

$$\mu = \sum_{i=1}^{n} x_i f(x_i) = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1 \text{ heads},$$

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2 = (0^0 \cdot 0.25 + 1^2 \cdot 0.5 + 2^2 \cdot 0.25) - 1^2 = 0.5 \text{ heads}^2,$$

$$\sigma = +\sqrt{0.5} = 0.71 \text{ heads}.$$

According to the type of experiment where the random variable is measured, there are different probability distributions models. The most common are

- Discrete uniform
- Binomial
- Poisson

Discrete uniform probability distribution model $U(a, b)$

When all the values of a random variable X have equal probability, the probability distribution of X is uniform.

Definition (Discrete uniform distribution $U(a, b)$)

A discrete random variable X follows a *discrete uniform distribution model* with parameters a and b, noted $X \sim U(a, b)$, if its range is $Ran(X) = \{a, a+1, \ldots, b\}$ and its probability function is

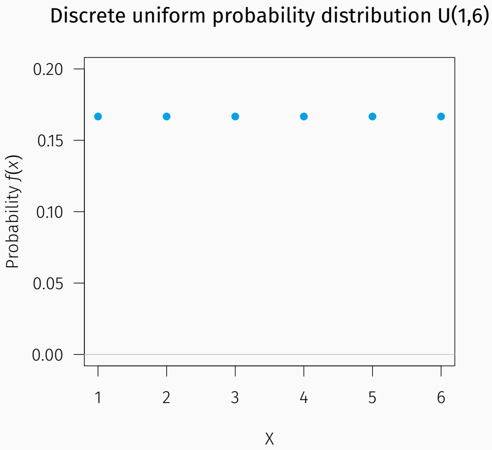
$$f(x) = \frac{1}{b-a+1}.$$

Observe that a and b are the minimum and the maximum of the range respectively.

The mean and the variance are

$$\mu = \sum_{i=0}^{b-a} \frac{a+i}{b-a+1} = \frac{a+b}{2} \qquad \sigma^2 = \sum_{i=0}^{b-a} \frac{(a+i-\mu)^2}{b-a+1} = \frac{(b-a+1)^2-1}{12}$$

The variable that measures the outcome of rolling a dice follows a discrete uniform distribution model $U(1, 6)$.



Usually the binomial distribution correspond to a variable measured in a random experiment with the following features:

- The experiment consist in a sequence of n repetitions of the same trial.
- Each trial is repeated in identical conditions and produces two possible outcomes known as *Success* or *Failure*.
- The trials are independent among them.
- The probability of Success is the same in all the trials and is $P(\text{Success}) = p$.

Under these conditions, the discrete random variable X that measures the number of successes in the n trials follows a *binomial distribution model* with parameters n and p.

Binomial distribution model $B(n, p)$

Definition (Binomial distribution $(B(n, p))$)

A discrete random variable X follows a *binomial distribution model* with parameters n and p, noted $X \sim B(n, p)$, if its range is $Ran(X) = \{0, 1, \ldots, n\}$ and its probability function is

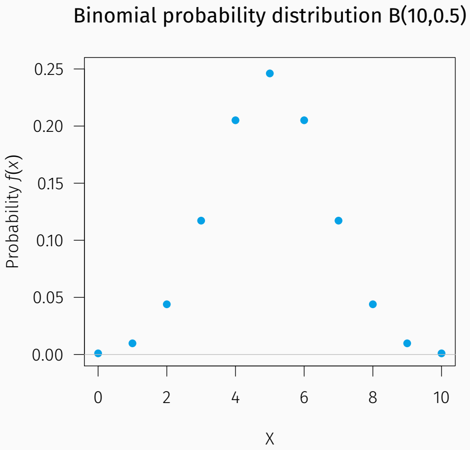
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

Observe that n is known as the number of repetitions of a trial and p is known as the probability of Success in every repetition.

The mean and the variance are

$$\mu = n \cdot p \qquad \sigma^2 = n \cdot p \cdot (1-p).$$

The variable that measures the number of heads after tossing 10 coins follows a binomial distribution model $B(10, 0.5)$.



If $X \sim B(10,\ 0.5)$ is the random variable that measures the number of heads after tossing 10 coins, then

- The probability of getting 4 heads is

$$f(4) = \binom{10}{4} 0.5^4 (1 - 0.5)^{10-4} = \frac{10!}{4!6!} 0.5^4 0.5^6 = 210 \cdot 0.5^{10} = 0.2051.$$

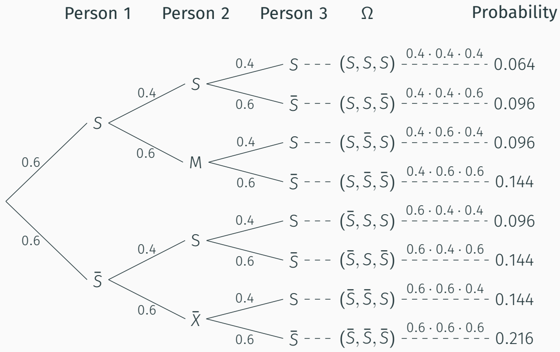
- The probability of getting 2 or less heads is

$$\begin{aligned} F(2) &= f(0) + f(1) + f(2) = \\ &= \binom{10}{0} 0.5^0 (1 - 0.5)^{10-0} + \binom{10}{1} 0.5^1 (1 - 0.5)^{10-1} + \binom{10}{2} 0.5^2 (1 - 0.5)^{10} \\ &= 0.0547. \end{aligned}$$

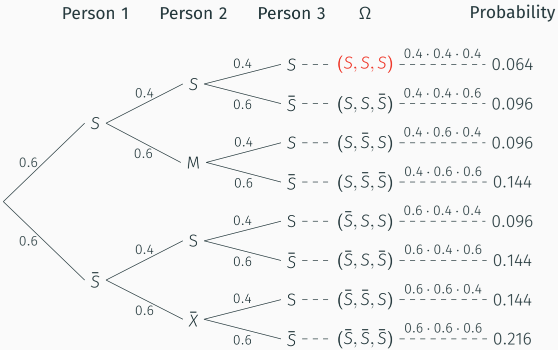
- And the expected number of heads is

$$\mu = 10 \cdot 0.5 = 5 \text{ heads}.$$

In a population there are a 40% of smokers. The variable X that measures the number of smokers in a random sample with replacement of 3 persons follows a binomial distribution model $X \sim B(3,\ 0.4)$.



In a population there are a 40% of smokers. The variable X that measures the number of smokers in a random sample with replacement of 3 persons follows a binomial distribution model $X \sim B(3,\ 0.4)$.

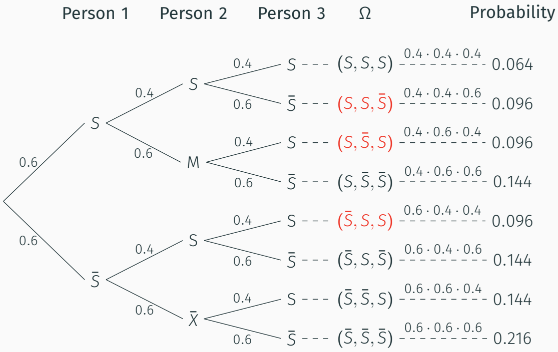


$f(0) = \binom{3}{0} 0.4^0 (1 - 0.4)^{3-0} = 0.6^3,$

Binomial distribution function $B(n, p)$

Example of random sampling with replacement

In a population there are a 40% of smokers. The variable X that measures the number of smokers in a random sample with replacement of 3 persons follows a binomial distribution model $X \sim B(3,\ 0.4)$.

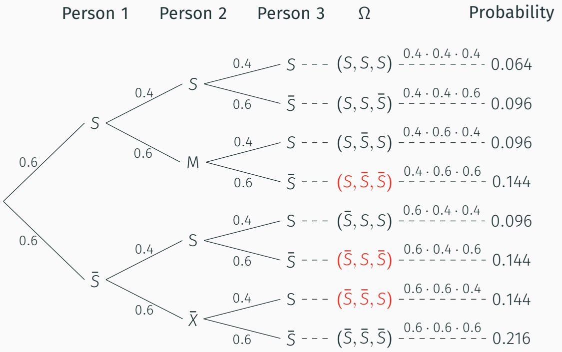


$f(0) = \binom{3}{0} 0.4^0 (1-0.4)^{3-0} = 0.6^3,$ $\quad$ $f(1) = \binom{3}{1} 0.4^1 (1-0.4)^{3-1} = 3 \cdot 0.4 \cdot 0.6^2,$

Binomial distribution function $B(n, p)$

Example of random sampling with replacement

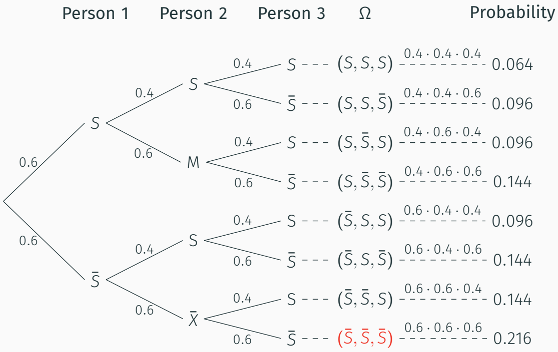
In a population there are a 40% of smokers. The variable X that measures the number of smokers in a random sample with replacement of 3 persons follows a binomial distribution model $X \sim B(3,\ 0.4)$.



$f(0) = \binom{3}{0} 0.4^0 (1-0.4)^{3-0} = 0.6^3,$ $\qquad f(1) = \binom{3}{1} 0.4^1 (1-0.4)^{3-1} = 3 \cdot 0.4 \cdot 0.6^2,$

$f(2) = \binom{3}{2} 0.4^2 (1-0.4)^{3-2} = 3 \cdot 0.4^2 \cdot 0.6,$

In a population there are a 40% of smokers. The variable X that measures the number of smokers in a random sample with replacement of 3 persons follows a binomial distribution model $X \sim B(3,\ 0.4)$.



$f(0) = \binom{3}{0}0.4^0(1-0.4)^{3-0} = 0.6^3,$ $\quad$ $f(1) = \binom{3}{1}0.4^1(1-0.4)^{3-1} = 3 \cdot 0.4 \cdot 0.6^2,$

$f(2) = \binom{3}{2}0.4^2(1-0.4)^{3-2} = 3 \cdot 0.4^2 \cdot 0.6,$ $\quad$ $f(3) = \binom{3}{3}0.4^3(1-0.4)^{3-3} = 0.4^3.$