A practical overview on probability distributions

Andrea Viti[1], Alberto Terzi[2], Luca Bertolaccini[2]

[1]Thoracic Surgery Unit, S. Croce Carle Hospital, Cuneo, Italy; [2]Thoracic Surgery Unit, Sacro Cuore Research Hospital, Negrar Verona, Italy
Correspondence to: Andrea Viti, MD, PhD. Thoracic Surgery Unit, S. Croce e Carle Hospital, Via Michele Coppino 26, 12100 Cuneo, Italy.
Email: vitimassa@hotmail.it.

Abstract: Aim of this paper is a general definition of probability, of its main mathematical features and the features it presents under particular circumstances. The behavior of probability is linked to the features of the phenomenon we would predict. This link can be defined probability distribution. Given the characteristics of phenomena (that we can also define variables), there are defined probability distribution. For categorical (or discrete) variables, the probability can be described by a binomial or Poisson distribution in the majority of cases. For continuous variables, the probability can be described by the most important distribution in statistics, the normal distribution. Distributions of probability are briefly described together with some examples for their possible application.

Keywords: Probability distributions; discrete variables; continuous variables

Submitted Nov 09, 2014. Accepted for publication Dec 17, 2014.
doi: 10.3978/j.issn.2072-1439.2015.01.37
View this article at: http://dx.doi.org/10.3978/j.issn.2072-1439.2015.01.37

A short definition of probability

We can define the probability of a given event by evaluating, in previous observations, the incidence of the same event under circumstances that are as similar as possible to the circumstances we are observing [this is the frequentistic definition of probability, and is based on the relative frequency of an observed event, observed in previous circumstances (1)]. In other words, probability describes the possibility of an event to occur given a series of circumstances (or under a series of pre-event factors). It is a form of inference, a way to predict what may happen, based on what happened before under the same (never exactly the same) circumstances. Probability can vary from 0 (our expected event was never observed, and should never happen) to 1 (or 100%, the event is almost sure). It is described by the following formula: if X = probability of a given x event (Eq. [1]):

$$\sum P(X=x)=1 \quad [1]$$

This is one of the three axioms of probability, as described by Kolmogorov (2):

(I) If under some circumstances, a given number of events (E) could verify (E_1, E_2, E_3, …, E_n), the probability (P) of any E is always more than zero;

(II) The sum of the probabilities of $E = P(E_1) + (E_2) + \cdots + P(E_n)$ is 100%;

(III) If E_1 and E_3 are two possible events, the probability that one or the other could happen P (E_1 or E_3) is equal to the sum of the probability of E_1 and the probability of E_3 (Eq. [2]):

$$P(E_1 \text{ or } E_2)=P(E_1)+P(E_3) \quad [2]$$

Probability could be described by a formula, a graph, in which each event is linked to its probability. This kind of description of probability is called probability distribution.

Binomial distribution

A classic example of probability distribution is the binomial distribution. It is the representation of the probability when only two events may happen, that are mutually exclusive. The typical example is when you toss a coin. You can only have two results. In this case, the probability is 50% for both events. However, binomial distribution may describe also two events that are mutually exclusive but are not equally possible (for instance that a newborn baby will be left-handed or right-handed). The probability that x individuals present a given characteristic, p, that is mutually exclusive of another one, called q, depends on the possible number

of combinations of x individuals within the population, called C. If my population is composed of five5 individuals, that can be p or q, I have ten possible combinations of, for instance, three individuals with p is (Eq. [3]):

pppqq, *pqqpp*, *ppqpq*, *ppqqp*, *pqpqpq*, *qpppq*, *qpqpp*, *qppqp*, *qqppp*

$$P(p,p,p,q,q) = pppqq = p^3q^2 \quad [3]$$

Then p^3q^2 will be multiplied for the number of combinations (ten times).

If, in experimental population, I had a big number of individuals (n), the number of combinations of x individuals within the population will be (Eq. [4]):

$$nC_x = \frac{n!}{x!(n-x)!} \quad [4]$$

Therefore, the probability that a group of x individuals within the population of n individuals presents the characteristic p, that excludes q, will be described by the following formula (Eq. [5]):

$$f(x) = nC_x p^x q^{n-x} \quad [5]$$

that describes the binomial distribution. It follows the Kolmogorow's rules (Eq. [6]):

$$f(x) > 1$$
$$\sum f(x) = 1 \quad [6]$$

In a given population, 30% of the people are left-handed. If we select ten individuals from this population, what is the probability that four out of ten individuals are left handed?

We can apply the binomial distribution, since we suppose that a person may be either left-handed or right-handed.

Se we can use our formula (Eq. [7]):

$$f(4) = 10C_4 (0.3)^4 (0.7)^6 = 0.2001 \quad [7]$$

Poisson distribution

Another important distribution of probability is the Poisson distribution. It is useful to describe the probability that a given event can happen within a given period (for instance, how many thoracic traumas could need the involvement of the thoracic surgeon in a day, or a week, etc.). The events that may be described by this distribution have the following characteristics:

(I) The events are independent from one another;

(II) Within a given interval the event may present from 0 to infinite times;

(III) The probability of an event to happen increases when the period of observation is longer.

To predict the probability, I must know how the events behave (this data comes from previous, or historical, observations of the same event before the time I am trying to perform my analysis). This parameter, that is a mean of the events in a given interval, as derived from previous observations, is called λ.

The Poisson distribution follows the following formula (Eq. [8]):

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad [8]$$

where the number e is an important mathematical constant that is the base of the natural logarithm. It is approximately equal to 2.71828.

For example, the distribution of major thoracic traumas needing intensive care unit (ICU) recovery during a month in the last three years in a Third Level Trauma Center follows a Poisson distribution, were λ=2.75. In a future period of one month, what is the probability to have three patients with major thoracic trauma in ICU? (Eq. [9]):

$$P = (X = 3) = \frac{2.75^3 e^{-2.75}}{3!} = 0.221 \quad [9]$$

Therefore, the probability is 22.1%.

The binomial distribution refers only to discrete variables (that present a limited number of values within a given interval). However, in nature, many variables may present an infinite distribution of values, within a given interval. These are called continuous variables (3).

Distributions of continuous variables

An example of continuous variable is the systolic blood pressure. Within a given cohort of systolic blood pressure can be presented as in *Figure 1*. Each single histogram length represents an interval of the measure of interest between two intervals on the x-axis, while the histogram height represents the number of measured values within the interval. When the number of observation becomes very large (tends to infinite) and the length of the histogram becomes narrower (tends to 0), the above representation becomes more similar to a curved line (*Figure 2*). This curve describes the distribution of probability, f (density of probability) for any given value of x, the continuous variable. The area under the curve is equal to 1 (100% of probability). We can now assume that the value of our continuous variable X depends on a very large number of other factors (in many cases beyond our possibility of

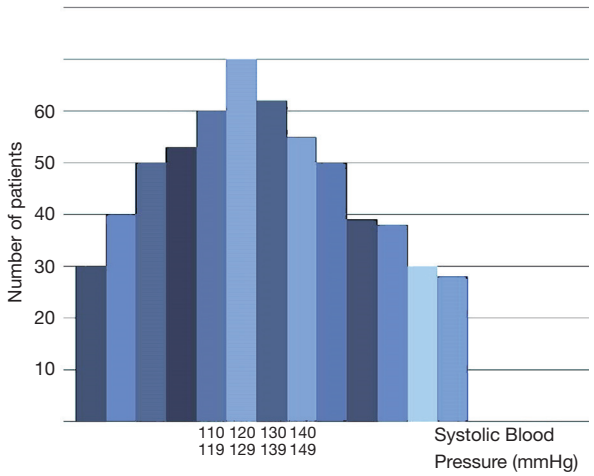

Figure 1 Graphical description of the distribution of systolic blood pressure in a given population.

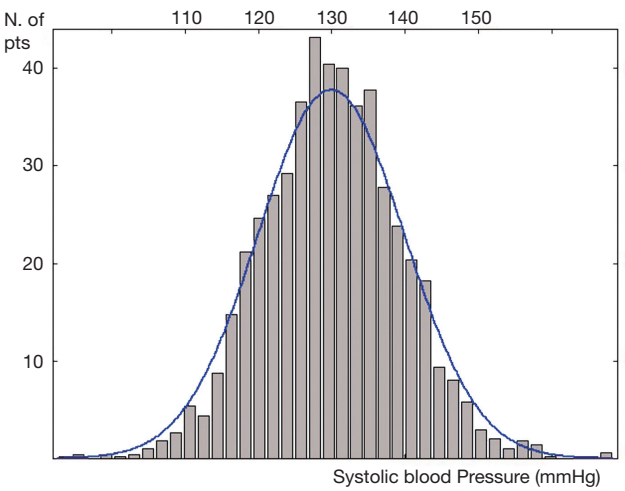

Figure 2 Graphical description of the normal distribution.

direct analysis), the probability distribution of X becomes similar to a particular form of distribution, called normal distribution or Gauss distribution. The aforementioned concept is the famous Central Limit Theorem. The normal distribution represents a very important distribution of probability because *f*, that is the distribution of probability of our variables, can be represented by only two parameters:

- μ = mean;
- σ = standard deviation.

The mean is a so-called measure of central tendency (it represents the more central value of our curve), while the standard deviation represents how dispersed are the values of probability around the central value (is a measure of dispersion).

(I) The main characteristics of this distribution are:

(II) It is symmetric around the μ;

(III) The area under the curve is 1;

If we consider the area under the curve between $\mu \pm \sigma$, this area will cover 68% of all the possible values of X, while the area between $\mu \pm 2\sigma$, it will cover 95% of all the values.

The two parameters of the distribution are linked in the formula (Eq. [10]):

$$f(x) = \frac{1}{\sqrt{2\sigma\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \quad [10]$$

For $\mu = 0$, and $\sigma = 1$, the curve is called standardized normal distribution. All the possible normal distributions of x may be "normalized" by defining a derived variable called z. (Eq. [11]):

$$Z = \frac{x-\mu}{\sigma}$$
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, -\infty < z < \infty \quad [11]$$

To calculate the probability that our variable falls within a given interval, for instance z_0 and z_1, we should calculate the following definite integral calculus (Eq. [12]):

$$\int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e \frac{-z^2}{2} dz \quad [12]$$

Fortunately, for the standard normalized distribution of z every possible interval has been tabulated.

In a given population of adult men, the mean weight is 70 kg, with a standard deviation of 3 kg. What is the probability that a randomly selected individual from this population would have a weight of 65 kg or less?

To "normalize" our distribution, we should calculate the value of z (Eq. [13]):

$$z = \frac{65-70}{3} = -1.67 \quad [13]$$

Then, we should calculate the area under the curve (Eq. [14]):

$$\int_{-\infty}^{-1.67} \frac{1}{\sqrt{2\pi}} e^{\frac{1.67^2}{2}} dz \quad [14]$$

The value of our interval has been already calculated and tabulated [the tables can be easily found in any text of statistics or in the web (4)]. Our probability is 0.0475 (4.75%). We may also calculate the probability to find, within the same population, someone whose weight is between 65 and 74 kg. This probability can be seen as the difference of distribution between those whose weight is 74 kg

or less and those whose weight is 65 kg or less: (Eq. [15]):

$$P(65 \le x \le 74) = P\left(\frac{65-70}{3} \le z \le \frac{74-70}{3}\right) = P(-1.67 \le z \le 1.33)$$

$$= P(-\infty \le z \le 1.33) - P(-\infty \le z \le 1.67) \quad [15]$$

We already know that (Eq. [16]):

$$P(-\infty \le z \le -1.67) = 0.0475 \quad [16]$$

In the table we can find also the value for (Eq. [17]):

$$P(-\infty \le z \le 1.33) = 0.9082 \quad [17]$$

Our probability is (Eq. [18]):

$$P = 0.9082 - 0.0475 = 0.8607(86.07\%) \quad [18]$$

Conclusions

The probability distributions are a common way to describe, and possibly predict, the probability of an event. The main point is to define the character of the variables whose behaviour we are trying to describe, trough probability (discrete or continuous). The identification of the right category will allow a proper application of a model (for instance, the standardized normal distribution) that would easily predict the probability of a given event.

Cite this article as: Viti A, Terzi A, Bertolaccini L. A practical overview on probability distributions. J Thorac Dis 2015;7(3):E7-E10. doi: 10.3978/j.issn.2072-1439.2015.01.37.

Acknowledgements

Disclosure: The authors declare no conflict of interest.

References

1. Daniel WW. eds. Biostatistics: a foundation for analysis in the health sciences. New York: John Wiley & Sons, 1995.
2. Kolmogorov AN. eds. Foundations of Theory of Probability. Oxford: Chelsea Publishing, 1950.
3. Lim E. Basic statistics (the fundamental concepts). J Thorac Dis 2014;6:1875-8.
4. Standard Normal Distribution Table. Available online: http://www.mathsisfun.com/data/standard-normal-distribution-table.html