“A basic explanation of the Expectation of a Random Variable”

Alan Berger October 6, 2020

Available at https://github.com/AlanBerger/Statistical_Inference_GitHub_files

Introduction

I’m going to give an explanation of what the **expectation** of a random variable is (without getting into any meaure theory). A heuristic definition of a **random variable** is a function X that assigns a number to each of the members of some “sample space” (for example the height of every person, or the temperature at every location in some geographic region or in some object, or the measured amount of antibodies raised in response to a vaccination for each subject in a clinical trial. The level of this discussion corresponds to the textbook “Statistical inference for data science”, Brian Caffo, Leanpub, last updated on 2016-05-23, https://leanpub.com/LittleInferenceBook used in the Statistical Inference Course in the Johns Hopkins University Data Science Specialization on Coursera. For this discussion, the random variable will be assumed to be either **discrete** or **continuous**.

Discrete Random Variables

A discrete random variable X takes on either a finite number of values (for example the outcome of tossing a coin is either 1 for heads or 0 for tails; and the outcome of rolling a single standard die is one of the values {1,2,3,4,5,6}); or X can have a countable number of possible values (for example the number of possible counts of some event governed by a random variable X having a Poisson distribution is the set of non-negative integers $\{0, 1, 2, \ldots\}$). The **probability mass function** (**pmf**) for a given discrete random variable X evaluated at a given value x (denoted by $p(x)$) is the probability that X has the value x, or phrased another way, $p(x)$ is the probability of the outcome or measurement value x. Hence $p(x)$ is only non-zero at a finite or countable number of values of x (commonly integers), and one can think of $p(x)$ as the fraction of time that the outcome x occurs.

Then the **expected value** of the discrete random variable X, denoted by $E(X)$, is the sum of $x * p(x)$ over the finite or countable values x that are possible values for X (or equivalently, over x such that $p(x) > 0$).

$$E(x) = \sum_{x \ni p(x)>0} x\, p(x)$$

So $E(X)$ can be thought of as the average value for X; it is the average of the possible outcome values x weighted by the fraction of the time (i.e., the probability) that each x occurs. For example if X was the amount one won or lost depending on the outcome of some “contest”, the average amount one would win (or lose) for each run of the “contest” would be $E(X)$. In practical terms, if one “played this game” N times and N was relatively large, one would expect, on average, to win (or lose if E was negative) $N * E(X)$.

Continuous Random Variables

A continuous random variable, for the purposes of the discussion here, is a random variable X that can take on any real value, or any value in one or several subintervals such as a finite interval $[a, b]$ with $a < b$, or a half interval $[a, \infty)$ or (a, ∞) or $(-\infty, a]$ or $(-\infty, a)$ with a a real number. For the commonly encountered continuous random variables, they can take on either all real values (for example the normal (Gaussian) distribution and the t-distribution), or values in a finite interval $[a, b]$ (the uniform distribution) or values in the half interval $[a, \infty)$ or (a, ∞) (for example the lognormal, exponential and Weibull distributions).

Thie rest of this section on continuous random variables is taken from (with some modifications) a pinned post of mine “The d p r and q functions for important statistical distributions – what they do and how to use them Part I" in the Week 2 Discussion Forum for the Statistical Inference Course in the Johns Hopkins Data Science Specialization on Coursera.

A probability density function (**pd**f) for a given continuous random variable X is a corresponding function $f(x)$ defined on the real line (for all real values x) which satisfies these three conditions:

1. For each number x, $f(x) \geq 0$. Note $f(x)$ may be 0 for none or some or many values of x.
2. The area under the entire graph of the probability density function $f(x)$ is 1 (meaning the area that is at and above the x-axis and at and below the graph of the function).
3. The probability that values of the random variable lie between any two given values of x, call them x_1 and x_2 with $x_1 \leq x_2$, is the area under the graph of $f(x)$ between x_1 and x_2 (mathematically, the integral of $f(x)$ from x_1 to x_2; $\int_{x_1}^{x_2} f(x)dx$). For our purposes the integral here is the Riemann integral of calculus and so f is assumed to be "reasonably well behaved" (which it is for the commonly used statistical distributions). Note this implies that the probability of any single value, and indeed, the probability of any set S consisting of a finite or countable set of values is 0, since the integral of f over an interval of length 0 is 0.

Item 3 is analogous to calculating the weight of a section of a cylinder of material where the cylinder has a cross section area of 1 (in whatever units of length are being used) and one is given the density of the material of the cylinder as a function of length along the cylinder. For the continuous random variables encountered in the Statistical Inference course (and commonly used in general), their probability density function (pdf) is either a continuous function (actually a "smooth" function − having continuous derivatives) on the real line ($-\infty < x < \infty$) (for example normal and t distribution functions); or is a continuous (and smooth) function defined on a finite interval $a \leq x \leq b$ with $a < b$ (for example uniform distribution functions), or on an infinite half interval $x_o \leq x < \infty$ or $x_o < x < \infty$ with x_o a given constant and with the pdf being 0 for all x outside the specified interval (for example the lognormal and exponential and Weibull distributions).

Sometimes a pdf is called a **probability distribution function** but note that term is also used for the cumulative distribution function given by $\text{cdf}(v) = \int_{-\infty}^{v} f(x)\, dx$.

One can think of a random variable determining (inducing) the corresponding pdf or pmf which determines the probability that the random variable has values in a given subset of the real line. Statistics tends to deal with the pdf or pmf more than the (often hidden) random variable function.

The expectation of a Continuous Random Variable

The formal definition of the expectation $E(X)$ for a continuous random variable X having pdf $f(x)$ is the integral over the real line (or equivalently, over an interval I when it is the case that $f(x) = 0$ for x not in I) of $x * f(x)$. This is

$$E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx \quad \text{or when applicable} \quad \int_I x\, f(x)\, dx \qquad \text{eqn 1}$$

There are random variables (i.e., pdf's) for which the expected value is in fact not defined (not finite), for example the Cauchy distribution, but that is not the case for the commonly used distributions.

The definition of $E(X)$ in equation 1 above can be viewed as a natural extension of the definition of $E(X)$ for discrete random variables above, in the following sense. Divide the real line up into subintervals $[x_i - h, x_i + h)$ each of length $2h$ which do not overlap and whose union is all real values. Given a continuous random variable X, "convert" it to a corresponding discrete random variable Y defined on the same sample space, such that whenever X has a value in $[x_i - h, x_i + h)$, the value of Y is equal x_i. So in effect, we are taking the pdf of X and in each subinterval $[x_i - h, x_i + h)$, we are "concentrating" the probability p_i that X is in that subinterval, $\int_{x_i - h}^{x_i + h} f(x)\, dx$, to the probability $p(x_i) = p_i$ that Y takes on the value x_i at the center of this interval. The "average value" for Y is, as described above, $E(Y) = \sum_i x_i \, p(x_i)$ with the probability mass function p for Y as just described. For continuous random variables X with "nice" pdf's, the value of $E(Y)$ will converge to $E(X)$ as h gets close to 0.

One can find the next two results in many statistics textbooks.

If X and Y are random variables (defined on the same sample space), and $g(x, y)$ is a "reasonably nice" function defined for pairs of real numbers (x, y), then g(X,Y) is also a random variable. The expectation of $aX + bY + c$ where a, b and c are constants, i.e., real numbers, is

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

For **independent random variables** X and Y it is the case that the expectation of the random variable X^*Y is the product of their expected values: $E(X)\,E(Y)$.

License and legal notice