Esssentials of Applied Data Analysis IPSA-USP Summer School 2018

Expected Value

Leonardo Sangali Barone
leonardo.barone@usp.br

jan/18

Expectation and mean value of a discrete random variable

Expectation and a simple game

Imagine a game where you toss a coin and get $1 if the result is head and 0 if the result is tail. How much should you expect to earn?

Remember:

x_i	$P(X = x_i)$
0	0.50
1	0.50

The expected value of a discrete random variable can be easily obtained by summing each result multiplied by probability of that result ocurring.

$$E[X] = x_1 * P(X = x_1) + x_2 * P(X = x_2) = 0 * 0.5 + 1 * 0.5 = 0.5$$

Now, imagine a game where you roll a dice and you can get $1 times the number you get on the dice. How much should you expect to earn?

$$E[X] = x_1 * P(X = x_1) + x_2 * P(X = x_2) + x_3 * P(X = x_3)$$
$$+x_4 * P(X = x_4) + x_5 * P(X = x_5) + x_6 * P(X = x_6) =$$
$$= \frac{1}{6} * 1 + \frac{1}{6} * 2 + \frac{1}{6} * 3 + \frac{1}{6} * 4 + \frac{1}{6} * 5 + \frac{1}{6} * 6 + \frac{1}{6} * 1 = 3.666$$

Expectation and mean value of a discrete random variable

In more general term, the expectation or mean of a discrete random variable is:

$$E[X] = \sum_{i=1}^{n} x_i * P(X = x_i) = \sum_{i=1}^{n} x_i * f(x_i)$$

where x_i is an occurence of the variable X and $f(x_i)$ is the probability mass function (the probability that x_i will occur).

Note that, since the set of all x_i is the set of all possible values for X, then

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

When all $P(X = x_i)$ are the same for every x_i, we can simplify the expression of $E[X]$ to:

$$E[X] = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n}$$

which is what we normally do to calculate avareges in daily life.

Expectation and mean value examples

Example 1: educational level in Fakeland

i	x_i	$P(X = x_i)$	f_i
1	"No High School Degree"	0.10	0.10
2	"High School Degree"	0.40	0.47
3	"College Incomplete"	0.20	0.2
4	"College Degree or more"	0.30	0.23

Can we compute $E[X]$? Not really. Expected value just makes sense for numerical variables (either Discrete-Integer or Continous. For the rest we can work only with proportions of each category.

What if we had "years of education" instead of educational level?

Example 2: educational level in Fakeland

$i = x_i$	$P(X = x_i)$
0	0.30
1	0.10
2	0.10
3	0.20
4	0.30

Can we compute $E[X]$? Yes:

$$E[X] = x_0 * P(X = x_0) + x_1 * P(X = x_1) + x_2 * P(X = x_2) + x_3 * P(X = x_3) + x_4 * P(X = x_4) =$$

$$= 0 * 0.30 + 1 * 0.10 + 2 * 0.10 + 3 * 0.2 + 4 * 0.30 = 0 + 0.1 + 0.2 + 0.6 + 1.2 = 2.1$$

which means that we expect that a Fakeland citizen voted in 2.1 of the last 4 presidential elections. Another way to read this is: on average, fakelandians voted in 2.1 of the last 4 presidential elections.

Properties of the mean

Some properties of the mean:

$$E[a * X] = a * E[X]$$

$$E[X + b] = E[X] + b$$

So what? Well, if you multiply a variable by a number (a) to generate a new variable, the mean of the new variable is the mean of the old variable times a.

Also, if you sum a quantity b to a random variable, the mean of the result variable will be the mean of the original variable plus b.

Let's try it later in Stata!

Expectation and variance of a discrete random variable

Another important quantity of a random variable is the variance. The name is self-explanatory: the variance measures how spread-out a variable is.

The variance of a random variable is also an expectation:

$$Var[X] = \sum_{i=1}^{n} [x_i - E[X]]^2 * P(X = x_i) = \sum_{i=1}^{n} [x_i - E[X]]^2 * f(x_i)$$

where x_i is an occurence of the variable X, E[X] is the expected value of X and $f(x_i)$ is the probability mass function (the probability that x_i will occur).

Look at the formula again and pay attention to the squared term, $[x_i - E[X]]$. Reading it aloud we would say "the value of variable X for observation i minus the mean of variable X". Or, more simply, "how much observation i *deviate* from the mean in X".

Why squared? Well, it is a mathematical trick. Deviations from the mean can be positive or negative. When squared, they all became positive (because the square of a number is always positive).

We are left with the last term, which is the probability of $X = x_i$, which should be familiar to us by now.

So the variance can be understood as the sum of squared deviations from the mean weigthed by their probability of ocurrence. Much easier to remember, right?

Variance of a discrete random variable

Going back to the coin game (where heads pays off $1 and tails pays off $0), the variance of X is:

$$Var[X] = \sum_{i=1}^{n} [x_i - E[X]]^2 * f(x_i) = [0.5 - 0.5]^2 * 0 + 0.5 * 1 = 0.5$$

Properties of the variance

Some properties of the variance:

$$Var[a * X] = a^2 * E[X]$$

$$Var[X + b] = Var[X]$$

So what? Well, if you multiply a variable by a number (a) to generate a new variable, the variance of the new variable is the variance of the old variable times a^2.

Also, if you sum a quantity b to a random variable, the variance of the result variable will equal to the variance of the original variable.

Let's try it later in Stata!

Expectation, mean, variance and standard deviation

Notation:

$$E[X] = \mu[X] = \mu$$

$$Var[X] = \sigma^2[X] = \sigma^2$$

The standard deviation (σ) of a variable is

$$\sigma = \sqrt{Var[X]} = \sqrt{\sigma^2}$$

Another way to calculate the variance is simply doing:

$$Var[X] = E[X^2] - (E[X])^2$$