Expected Value & Standard Deviation of Random Variables

Stacey Hancock

1 Definition of Random Variable

When we assign a number to the outcome of a random process, we call this a **random variable**. For example, flipping a coin once is a random process, and we could define the random variable X to take on the value 0 if the coin lands on tails and 1 if the coin lands on heads. Examples of other random variables include:

1. The amount of claims (in hundreds of dollars) an insurance company has to pay out to its clients in a randomly selected month.
2. The time (in hours) a randomly chosen MSU student spends studying for final exams.
3. The number of girls in the next three births at the Bozeman hospital.
4. The number of heads in nine tosses of a coin.

Not only are questions about the distribution of a random variable, or its probabilities, of interest, but we may want to determine the "average" or **expected value** of a random variable as well as how far it tends to vary from its expected value, or its **standard deviation**. We will only study expected value and standard deviation for **discrete** random variables which are random variables whose set of possible values form a countable list of distinct values. For example, the number of girls in the next three births at the Bozeman hospital is a discrete random variable since it can only take on the values 0, 1, 2, or 3. Discrete random variables can take on an infinite number of possible values, as long as we can list them in an ordered list. For example, the number of tosses of a coin until the first head appears is a discrete random variable with possible values 1, 2, 3, 4, ... Random variables that can take on any value in an interval (e.g., time, length, interest rates, height) are called **continuous** random variables.

We will use the following notation to specify a probability for a possible outcome of a discrete random variable:

- X = the random variable (e.g., number of girls in the next three births)
- k = a possible value of the random variable (e.g., 2)
- $P(X = k)$ = the probability that the random variable X equals the value k

2 Expected Value: What is the average value over many observations of the random variable?

If we know the probabilities of each of the possible values of a discrete random variable, then we can calculate the long-run average of the variable, or its mean value over an infinite number of observations of the random variable. We call this its **expected value**. Suppose the random variable X can take on possible values k_1, k_2, k_3,... Then we define the expected value of X as

Mathematical definition:

$$E(X) = k_1 \times P(X = k_1) + k_2 \times P(X = k_2) + k_3 \times P(X = k_3) + ...$$

$$= \text{sum of "value}\times\text{probability" summed over all possible values}$$

Interpretation: The expected value of X, $E(X)$, is the mean value that would be obtained from an infinite number of observations of the random variable, or its long-run average.

Example: Suppose that in a gambling game, it costs \$1 to play, and you win \$2 with probability 0.37 or \$10 with probability 0.02 (otherwise you lose your dollar). Let X be the net amount won. Then the probability distribution of X is

X = net amount won	\$9	\$1	$-\$1$
Probability	0.02	0.37	$1 - 0.37 - 0.02 = 0.61$

If you were to play this game a large number of times, how much money would you win per game on average?

The expected net amount won is

$$E(X) = 9(0.02) + 1(0.37) - 1(0.61) = -0.06.$$

That is, in the long-run, you can expect to *lose* about \$0.06 per game, on average.

3 Standard Deviation: About how far away from the expected value does the random variable lie in the long run?

Mathematical definition:

$$SD(X) = \sqrt{(k_1 - E(X))^2 \times P(X = k_1) + (k_2 - E(X))^2 \times P(X = k_2) + ...}$$

= square root of the sum of "(value − expected value) squared ×probability"

summed over all possible values

Interpretation: The standard deviation of X, $SD(X)$, is the approximate average distance we would expect an observation of the random variable to be away from its mean in an infinite number of observations of the random variable.

Example (continued): Consider the gambling game from the previous example. On average, how far away from $-\$0.06$ are your net winnings per game in the long run?

The standard deviation of the net amount won per game is

$$SD(X) = \sqrt{(9 - (-0.06))^2 \times 0.02 + (1 - (-0.06))^2 \times 0.37 + (-1 - (-0.06))^2 \times 0.61} = 1.6113.$$

Thus, in the long run, the net winnings per game are about \$1.61 away from $-\$0.06$.

4 Expected value and standard deviation of a random variable versus sample mean and standard deviation

A **population** is the entire collection of all individuals about which information is desired. Often it is infeasible to measure every individual in the population, so we take a **sample**, or subset, of individuals from the population, and use measurements on the sample to infer information about the larger population. The expected value (mean) and standard deviation defined above are for a

population. We could also calculate the mean and standard deviation of a sample, which are defined using the values of the sample, not true probabilities. Let's explore this idea through an example.

A new pet store is interested in the probability distribution of the number of dogs in each household, X, in a city of 5,000 households. Unknown to the pet store, 2400 of these households (48%) do not own a dog, 1750 own one dog (35%), 650 own two dogs (13%), and 200 own three dogs (4%). The mean number of dogs in the population of households is

$$\begin{aligned} E(X) &= 0 \times (2400/5000) + 1 \times (1750/5000) + 2 \times (650/5000) + 3 \times (200/5000) \\ &= 0(.48) + 1(.35) + 2(.13) + 3(.04) = 0.73. \end{aligned}$$

The average number of dogs per household in this city is 0.73 dogs. (Do not round this to 1! The expected value is a long-run average, or the average in a large population, thus often it will not be a possible value of X. One household cannot have 0.73 dogs, but the average over a large number of households can be 0.73 dogs.) We could have also calculated the population mean number of dogs by adding up all the values in the population and dividing by the population size:

$$\begin{aligned} E(X) &= \frac{(0+0+\cdots+0)+(1+1+\cdots+1)+(2+2+\cdots+2)+(3+3+\cdots+3)}{5000} \\ &= \frac{0(2400)+1(1750)+2(650)+3(200)}{5000} = 0.73 \end{aligned}$$

since multiplying each possible value by its probability and adding them up is equivalent to multiplying each possible value by the number of observations in the population that resulted in that value then dividing the sum of those values by the population size. The standard deviation of the number of dogs in the population is

$$\begin{aligned} SD(X) &= \sqrt{(0-0.73)^2(.48) + (1-0.73)^2(.35) + (2-0.73)^2(.13) + (3-0.73)^2(.04)} \\ &= \sqrt{0.6971} = 0.8349. \end{aligned}$$

On average, the number of dogs in a typical household in this city is about 0.8349 away from the average number of dogs of 0.73. (Again, a single household cannot have $0.73 + 0.8349 = 1.5649$ dogs. Instead, the value 0.8349 represents an approximate long-run average distance away from the mean over many households.)

The pet store does not have enough money to survey all of the 5,000 households in the population, so they survey a random sample of 60 households in the city. Suppose that in their sample of 60 households, 29 do not own a dog, 21 own one dog, 9 own two dogs, and 1 owns three dogs. The *sample* mean and standard deviation are

$$\bar{x} = \frac{(0+\cdots+0)+(1+\cdots+1)+(2+\cdots+2)+3}{60} = \frac{0(29)+1(21)+2(9)+3}{60} = 0.70$$

and

$$s = \sqrt{\frac{(0-0.70)^2(29) + (1-0.70)^2(21) + (2-0.70)^2(9) + (3-0.70)^2}{60-1}} = 0.7876.$$

(See Appendix A in Tintle, et. al. for formulas and calculation details for a sample mean and sample standard deviation.) These **sample statistics** are not equal to the **population parameters**, but they are close. In fact, if we were to take many many random samples of size 60 from the population of 5,000 households, the center of the distribution of sample means would be around 0.73, and the center of the distribution of the sample standard deviations would be around 0.8349. Note that each time you take a different random sample of 60 households, you will get different values for the sample mean and sample standard deviation, but the population mean and population standard deviation are fixed values and do not change.

5 References

- Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2014). *Introductory Statistics with Randomization and Simulation*, Appendix A. Creative Commons license, openintro.org.
- Tintle, N. L., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T. M., & VanderStoep, J. L. (2016). *Introduction to Statistical Investigations*. Hoboken, NJ: Wiley.
- Utts, J. M., & Heckard, R. F. (2015). *Mind on Statistics*, 5th ed., Chapters 7 and 8. Stamford, CT: Cengage Learning.