ECS315 2018/1 Part III.3 Dr.Prapun

9 Expectation and Variance

Two numbers are often used to summarize a probability distribution for a random variable X. The mean is a measure of the center or middle of the probability distribution, and the variance is a measure of the dispersion, or variability in the distribution. These two measures do not uniquely identify a probability distribution. That is, two different distributions can have the same mean and variance. Still, these measures are simple, useful summaries of the probability distribution of X.

9.1 Expectation of Discrete Random Variable

The most important characteristic of a random variable is its expectation. Synonyms for expectation are expected value, mean, and first moment.

The definition of expectation is motivated by the conventional idea of numerical average. Recall that the numerical average of n numbers, say $a_1, a_2, \ldots, a_n$ is

$$\frac{1}{n}\sum_{k=1}^{n} a_k.$$

We use the average to summarize or characterize the entire collection of numbers $a_1, \ldots, a_n$ with a single value.

Example 9.1. Consider 10 numbers: 5, 2, 3, 2, 5, -2, 3, 2, 5, 2.

The average is

$$\frac{5+2+3+2+5+(-2)+3+2+5+2}{10} = \frac{27}{10} = 2.7.$$

We can rewrite the above calculation as

$$-2 \times \frac{1}{10} + 2 \times \frac{4}{10} + 3 \times \frac{2}{10} + 5 \times \frac{3}{10}$$

Definition 9.2. Suppose X is a discrete random variable, we define the **expectation** (or ***mean*** or ***expected value***) of X by

$$\mathbb{E}X = \sum_x x \times P[X = x] = \sum_x x \times p_X(x). \tag{15}$$

In other words, The expected value of a discrete random variable is a weighted mean of the values the random variable can take on where the weights come from the pmf of the random variable.

- Some references use m_X or μ_X to represent $\mathbb{E}X$.
- For conciseness, we simply write x under the summation symbol in (15); this means that the sum runs over all x values in the support of X. (Of course, for x outside of the support, $p_X(x)$ is 0 anyway.)

9.3. Analogy: In mechanics, think of point masses on a line with a mass of $p_X(x)$ kg. at a distance x meters from the origin.

In this model, $\mathbb{E}X$ is the center of mass (the balance point).

This is why $p_X(x)$ is called probability mass function.

Example 9.4. When $X \sim \text{Bernoulli}(p)$ with $p \in (0,1)$,

Note that, since X takes only the values 0 and 1, its expected value p is "never seen".

9.5. Interpretation: The expected value is in general not a typical value that the random variable can take on. It is often helpful to interpret the expected value of a random variable as the ***long-run average value*** of the variable over many independent repetitions of an experiment

Example 9.6. $p_X(x) = \begin{cases} 1/4, & x = 0 \\ 3/4, & x = 2 \\ 0, & \text{otherwise} \end{cases}$

Example 9.7. For $X \sim \mathcal{P}(\alpha)$,

$$\begin{aligned} \mathbb{E}X &= \sum_{i=0}^{\infty} i e^{-\alpha} \frac{(\alpha)^i}{i!} = \sum_{i=1}^{\infty} e^{-\alpha} \frac{(\alpha)^i}{i!} i + 0 = e^{-\alpha}(\alpha) \sum_{i=1}^{\infty} \frac{(\alpha)^{i-1}}{(i-1)!} \\ &= e^{-\alpha} \alpha \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} = e^{-\alpha} \alpha e^{\alpha} = \alpha. \end{aligned}$$

Example 9.8. For $X \sim \mathcal{B}(n,p)$,

$$\begin{aligned} \mathbb{E}X &= \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} i \frac{n!}{i!\,(n-i)!} p^i (1-p)^{n-i} \\ &= n \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!\,(n-i)!} p^i (1-p)^{n-i} = n \sum_{i=1}^{n} \binom{n-1}{i-1} p^i (1-p)^{n-i} \end{aligned}$$

Let $k = i - 1$. Then,

$$\mathbb{E}X = n \sum_{k=0}^{n-1} \binom{n-1}{k} p^{k+1} (1-p)^{n-(k+1)} = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

We now have the expression in the form that we can apply the binomial theorem which finally gives

$$\mathbb{E}X = np(p + (1-p))^{n-1} = np.$$

We shall revisit this example again using another approach in Example 11.41.

Example 9.9. Calculation of Expected Profit

Game #1: Flip a fair coin.

- H: You get \$200
- T: You lose \$100

Game #2: Flip a biased coin with $P(\{\text{H}\}) = 10^{-6}$.

- H: You get \$2,000,000
- T: You lose \$0

Game #3: Pay \$50 to play the game.
Flip a biased coin with $P(\{\text{H}\}) = 10^{-6}$.

- H: You get \$2,000,000
- T: You lose \$0

Example 9.10. ***Pascal's wager***: Suppose you concede that you don't know whether or not God exists and therefore assign a 50 percent chance to either proposition. How should you weigh these odds when deciding whether to lead a pious life? If you act piously and God exists, Pascal argued, your gain–eternal happiness–is infinite. If, on the other hand, God does not exist, your loss, or negative return, is small–the sacrifices of piety. To weigh these possible gains and losses, Pascal proposed, you multiply the probability of each possible outcome by its payoff and add them all up, forming a kind of average or expected payoff. In other words, the mathematical expectation of your return on piety is one-half infinity (your gain if God exists) minus one-half a small number (your loss if he does not exist). Pascal knew enough about infinity to know that the answer to this calculation is infinite, and thus the expected return on piety is infinitely positive. Every reasonable person, Pascal concluded, should therefore follow the laws of God. [14, p 76]

- Pascals wager is often considered the founding of the mathematical discipline of game theory, the quantitative study of optimal decision strategies in games.

Example 9.11. A sweepstakes sent through the mail offered a grand prize of $5 million. All you had to do to win was mail in your entry. There was no limit on how many times you could enter, but each entry had to be mailed in separately. The sponsors were apparently expecting about 200 million entries, because the fine print said that the chances of winning were 1 in 200 million. Does it pay to enter this kind of "free sweepstakes offer"?

Multiplying the probability of winning times the payoff, we find that each entry was worth 1/40 of $1, or $0.025 far less than the cost of mailing it in. In fact, the big winner in this contest was the post office, which, if the projections were correct, made nearly $80 million in postage revenue on all the submissions. [14, p 77]

9.12. Technical issue: Definition (15) is only meaningful if the sum is well defined.

The sum of infinitely many nonnegative terms is always well-defined, with $+\infty$ as a possible value for the sum.

- ***Infinite Expectation***: Consider a random variable X whose pmf is defined by

$$p_X(x) = \begin{cases} \frac{1}{cx^2}, & x = 1, 2, 3, \ldots \\ 0, & \text{otherwise} \end{cases}$$

  Then, $c = \sum_{n=1}^{\infty} \frac{1}{n^2}$ which is a finite positive number $(\pi^2/6)$. However,

$$\mathbb{E}X = \sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k \frac{1}{c} \frac{1}{k^2} = \frac{1}{c} \sum_{k=1}^{\infty} \frac{1}{k} = +\infty.$$

Some care is necessary when computing expectations of signed random variables that take infinitely many values.

- The sum over countably infinite many terms is not always well defined when both positive and negative terms are involved.
- For example, the infinite series $1-1+1-1+\ldots$ has the sum 0 when you sum the terms according to $(1-1)+(1-1)+\cdots$, whereas you get the sum 1 when you sum the terms according to $1+(-1+1)+(-1+1)+(-1+1)+\cdots$.

- Such abnormalities cannot happen when all terms in the infinite summation are nonnegative.

It is the convention in probability theory that $\mathbb{E}X$ should be evaluated as

$$\mathbb{E}X = \sum_{x \geq 0} x p_X(x) - \sum_{x<0} (-x) p_X(x),$$

- If at least one of these sums is finite, then it is clear what value should be assigned as $\mathbb{E}X$.
- If both sums are $+\infty$, then no value is assigned to $\mathbb{E}X$, and we say that $\mathbb{E}X$ is **undefined**.

Example 9.13. Undefined Expectation: Let

$$p_X(x) = \begin{cases} \frac{1}{2cx^2}, & x = \pm 1, \pm 2, \pm 3, \ldots \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$\mathbb{E}X = \sum_{k=1}^{\infty} k p_X(k) - \sum_{k=-\infty}^{-1} (-k) p_X(k).$$

The first sum gives

$$\sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k \frac{1}{2ck^2} = \frac{1}{2c} \sum_{k=1}^{\infty} \frac{1}{k} = \frac{\infty}{2c}.$$

The second sum gives

$$\sum_{k=-\infty}^{-1} (-k) p_X(k) = \sum_{k=1}^{\infty} k p_X(-k) = \sum_{k=1}^{\infty} k \frac{1}{2ck^2} = \frac{1}{2c} \sum_{k=1}^{\infty} \frac{1}{k} = \frac{\infty}{2c}.$$

Because both sums are infinite, we conclude that $\mathbb{E}X$ is undefined.

9.14. More rigorously, to define $\mathbb{E}X$, we let $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$. Then observe that $X = X^+ - X^-$ and that both X^+ and X^- are nonnegative r.v.'s. We say that a random variable X ***admits an expectation*** if $\mathbb{E}X^+$ and $\mathbb{E}X^-$ are not both equal to $+\infty$. In which case, $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$.

9.2 Function of a Discrete Random Variable

Given a random variable X, we will often have occasion to define a new random variable by $Y \equiv g(X)$, where $g(x)$ is a real-valued function of the real-valued variable x. More precisely, recall that a random variable X is actually a function taking points of the sample space, $\omega \in \Omega$, into real numbers $X(\omega)$. Hence, we have the following definition:

Definition 9.15. The notation $Y = g(X)$ is actually shorthand for $Y(\omega) := g(X(\omega))$.

- The random variable $Y = g(X)$ is sometimes called **derived** random variable.

Example 9.16. Let

$$p_X(x) = \begin{cases} \frac{1}{c}x^2, & x = \pm 1, \pm 2 \\ 0, & \text{otherwise} \end{cases}$$

and

$$Y = X^4.$$

Find $p_Y(y)$ and then calculate $\mathbb{E}Y$.

9.17. For discrete random variable X, the pmf of a derived random variable $Y = g(X)$ is given by

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

Note that the sum is over all x in the support of X which satisfy $g(x) = y$.

Example 9.18. A "binary" random variable X takes only two values a and b with

$$P[X = b] = 1 - P[X = a] = p.$$

X can be expressed as $X = (b - a)I + a$, where I is a Bernoulli random variable with parameter p.

9.3 Expectation of a Function of a Discrete Random Variable

Recall that for discrete random variable X, the pmf of a derived random variable $Y = g(X)$ is given by

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

If we want to compute $\mathbb{E}Y$, it might seem that we first have to find the pmf of Y. Typically, this requires a detailed analysis of g which can be complicated, and it is avoided by the following result.

9.19. Suppose X is a discrete random variable.

$$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x).$$

This is referred to as the **law/rule of the lazy/unconcious statistician** (LOTUS) [22, Thm 3.6 p 48],[9, p. 149],[8, p. 50] because it is so much easier to use the above formula than to first find the pmf of Y. It is also called **substitution rule** [21, p 271].

Example 9.20. Back to Example 9.16. Recall that

$$p_X(x) = \begin{cases} \frac{1}{c}x^2, & x = \pm 1, \pm 2 \\ 0, & \text{otherwise} \end{cases}$$

(a) When $Y = X^4$, $\mathbb{E}Y =$

(b) $\mathbb{E}[2X-1]$

9.21. Caution: A frequently made *mistake* of beginning students is to set $\mathbb{E}[g(X)]$ equal to $g(\mathbb{E}X)$. In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}X)$.

(a) In particular, $\mathbb{E}\left[\frac{1}{X}\right]$ is not the same as $\frac{1}{\mathbb{E}X}$.

(b) An exception is the case of an affine function $g(x) = ax + b$. See also (9.27).

Example 9.22. Continue from Example 9.16 and Example 9.20.

Example 9.23. Continue from Example 9.4. For $X \sim \text{Bernoulli}(p)$,

(a) $\mathbb{E}X = p$

(b) $\mathbb{E}\left[X^2\right] = 0^2 \times (1-p) + 1^2 \times p = p \neq (\mathbb{E}X)^2$.

Example 9.24. Continue from Example 9.7. Suppose $X \sim \mathcal{P}(\alpha)$.

$$\mathbb{E}\left[X^2\right] = \sum_{i=0}^{\infty} i^2 e^{-\alpha} \frac{\alpha^i}{i!} = e^{-\alpha}\alpha \sum_{i=1}^{\infty} i \frac{\alpha^{i-1}}{(i-1)!} \tag{16}$$

We can evaluate the infinite sum in (16) by rewriting i as $i-1+1$:

$$\begin{aligned}\sum_{i=1}^{\infty} i\frac{\alpha^{i-1}}{(i-1)!} &= \sum_{i=1}^{\infty} (i-1+1)\frac{\alpha^{i-1}}{(i-1)!} = \sum_{i=1}^{\infty} (i-1)\frac{\alpha^{i-1}}{(i-1)!} + \sum_{i=1}^{\infty} \frac{\alpha^{i-1}}{(i-1)!} \\ &= \alpha \sum_{i=2}^{\infty} \frac{\alpha^{i-2}}{(i-2)!} + \sum_{i=1}^{\infty} \frac{\alpha^{i-1}}{(i-1)!} = \alpha e^{\alpha} + e^{\alpha} = e^{\alpha}(\alpha+1).\end{aligned}$$

Plugging this back into (16), we get

$$\mathbb{E}\left[X^2\right] = \alpha(\alpha+1) = \alpha^2 + \alpha.$$

9.25. Continue from Example 9.8. For $X \sim \mathcal{B}(n,p)$, one can find $\mathbb{E}\left[X^2\right] = np(1-p) + (np)^2$.

Example 9.26. Let $p_X(x) = \begin{cases} 1/3, & x \in \{-1, 1\}, \\ 1/6, & x \in \{-2, 2\}, \\ 0, & \text{otherwise.} \end{cases}$ Find $\mathbb{E}X$ and $\mathbb{E}\left[X^2\right]$.

9.27. Some Basic Properties of Expectations

(a) For $c \in \mathbb{R}$, $\mathbb{E}[c] = c$

(b) For $c \in \mathbb{R}$, $\mathbb{E}[X + c] = \mathbb{E}X + c$ and $\mathbb{E}[cX] = c\mathbb{E}X$

(c) For constants a, b, we have

$$\mathbb{E}[aX + b] = a\mathbb{E}X + b.$$

(d) For constants c_1 and c_2,

$$\mathbb{E}[c_1 g_1(X) + c_2 g_2(X)] = c_1 \mathbb{E}[g_1(X)] + c_2 \mathbb{E}[g_2(X)].$$

(e) For constants $c_1, c_2, \ldots, c_n$,

$$\mathbb{E}\left[\sum_{k=1}^{n} c_k g_k(X)\right] = \sum_{k=1}^{n} c_k \mathbb{E}[g_k(X)].$$

Definition 9.28. Some definitions involving expectation of a function of a random variable:

(a) ***Absolute moment***: $\mathbb{E}\left[|X|^k\right]$, where we define $\mathbb{E}\left[|X|^0\right] = 1$

(b) ***Moment***: $m_k = \mathbb{E}\left[X^k\right]$ = the k^{th} moment of X, $\quad k \in \mathbb{N}$.

- The first moment of X is its expectation $\mathbb{E}X$.
- The second moment of X is $\mathbb{E}\left[X^2\right]$.

9.4 Variance and Standard Deviation

An average (expectation) can be regarded as one number that summarizes an entire probability model. After finding an average, someone who wants to look further into the probability model might ask, "How typical is the average?" or, "What are the chances of observing an event far from the average?" A measure of **dispersion/deviation/spread** is an answer to these questions wrapped up in a single number. (The opposite of this measure is the **peakedness**.) If this measure is small, observations are likely to be near the average. A high measure of dispersion suggests that it is not unusual to observe events that are far from the average.

Example 9.29. Consider your score on the midterm exam. After you find out your score is 7 points above average, you are likely to ask, "How good is that? Is it near the top of the class or somewhere near the middle?".

Example 9.30. In the case that the random variable X is the random payoff in a game that can be repeated many times under identical conditions, the expected value of X is an informative measure on the grounds of the law of large numbers. However, the information provided by $\mathbb{E}X$ is usually not sufficient when X is the random payoff in a nonrepeatable game.

Suppose your investment has yielded a profit of \$3,000 and you must choose between the following two options:

- the first option is to take the sure profit of \$3,000 and
- the second option is to reinvest the profit of \$3,000 under the scenario that this profit increases to \$4,000 with probability 0.8 and is lost with probability 0.2.

The expected profit of the second option is

$$0.8 \times \$4,000 + 0.2 \times \$0 = \$3,200$$

and is larger than the \$3,000 from the first option. Nevertheless, most people would prefer the first option. The downside ***risk*** is too big for them. A measure that takes into account the aspect of risk is the *variance* of a random variable. [21, p 35]

9.31. The most important ***measures of dispersion*** are the standard deviation and its close relative, the variance.

Definition 9.32. ***Variance***:

$$\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]. \tag{17}$$

- Read "the variance of X"
- *Notation*: D_X, or $\sigma^2(X)$, or σ_X^2, or $\mathbb{V}X$ [22, p. 51]
- In some references, to avoid confusion from the two expectation symbols, they first define $m = \mathbb{E}X$ and then define the variance of X by

$$\operatorname{Var} X = \mathbb{E}\left[(X - m)^2\right].$$

- We can also calculate the variance via another identity:

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2$$

- The units of the variance are squares of the units of the random variable.

9.33. Basic properties of variance:

- $\operatorname{Var} X \geq 0$.
- $\operatorname{Var} X \leq \mathbb{E}\left[X^2\right]$.
- $\operatorname{Var}[cX] = c^2 \operatorname{Var} X$.
- $\operatorname{Var}[X + c] = \operatorname{Var} X$.
- $\operatorname{Var}[aX + b] = a^2 \operatorname{Var} X$.

Example 9.34. In Example 9.26, we found that $\mathbb{E}X = 0$ and $\mathbb{E}\left[X^2\right] = 2$. Therefore,

Definition 9.35. ***Standard Deviation***:

$$\sigma_X = \sqrt{\text{Var}[X]}.$$

- It is useful to work with the standard deviation since it has the same units as $\mathbb{E}X$.
- Informally we think of outcomes within $\pm\sigma_X$ of $\mathbb{E}X$ as being in the center of the distribution. Some references would informally interpret sample values within $\pm\sigma_X$ of the expected value, $x \in [\mathbb{E}X - \sigma_X, \mathbb{E}X + \sigma_X]$, as "typical" values of X and other values as "unusual".
- $\sigma_{aX+b} = |a|\,\sigma_X$.

9.36. σ_X and $\sqrt{\text{Var}\,X}$: Note that the $\sqrt{\cdot}$ function is a strictly increasing function. Because $\sigma_X = \sqrt{\text{Var}\,X}$, if one of them is large, another one is also large. Therefore, both values quantify the amount of spread/dispersion in RV X (which can be observed from the spread or dispersion of the pmf or the histogram or the relative frequency graph). However, $\text{Var}\,X$ does not have the same unit as the RV X.

9.37. In finance, standard deviation is a key concept and is used to measure the ***volatility*** (risk) of investment returns and stock returns.

It is common wisdom in finance that diversification of a portfolio of stocks generally reduces the total risk exposure of the investment. We shall return to this point in Example 11.60.

Example 9.38. Continue from Example 9.29. If the standard deviation of exam scores is 12 points, the student with a score of +7 with respect to the mean can think of herself in the middle of the class. If the standard deviation is 3 points, she is likely to be near the top.

Example 9.39. Suppose $X \sim \text{Bernoulli}(p)$.

(a) $\mathbb{E}\left[X^2\right] = 0^2 \times (1-p) + 1^2 \times p = p.$

(b) $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$.
Alternatively, if we directly use (17), we have

$$\begin{aligned}\operatorname{Var} X &= \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = (0-p)^2 \times (1-p) + (1-p)^2 \times p \\ &= p(1-p)(p + (1-p)) = p(1-p).\end{aligned}$$

Example 9.40. Continue from Example 9.7 and Example 9.24. When $X \sim \mathcal{P}(\alpha)$, we have

$$\operatorname{Var} X = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2 = \alpha^2 + \alpha - \alpha^2 = \alpha.$$

Therefore, for Poisson random variable, the expected value is the same as the variance.

9.41. Continue from Example 9.8 and Example 9.25.
When $X \sim \mathcal{B}(n, p)$, we have $\operatorname{Var} X = np(1-p)$.

Example 9.42. Consider the two pmfs shown in Figure 20. The random variable X with pmf at the left has a smaller variance than the random variable Y with pmf at the right because more probability mass is concentrated near zero (their mean) in the graph at the left than in the graph at the right. [9, p. 85]

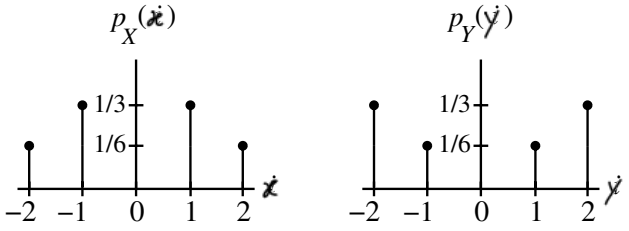

Figure 20: Example 9.42 shows that a random variable whose probability mass is concentrated near the mean has smaller variance. [9, Fig. 2.9]

9.43. We have already talked about variance and standard deviation as a number that indicates spread/dispersion of the pmf. More specifically, let's imagine a pmf that shapes like a bell curve. As the value of σ_X gets smaller, the spread of the pmf will be smaller and hence the pmf would "look sharper". Therefore, the probability that the random variable X would take a value that is far from the mean would be smaller.

The next property involves the use of σ_X to bound "the tail probability" of a random variable.

9.44. *Chebyshev's Inequality*:

$$P\left[|X - \mathbb{E}X| \geq \alpha\right] \leq \frac{\sigma_X^2}{\alpha^2}$$

or equivalently

$$P\left[|X - \mathbb{E}X| \geq n\sigma_X\right] \leq \frac{1}{n^2}$$

- Useful only when $\alpha > \sigma_X$

Example 9.45. If X has mean m and variance σ^2, it is sometimes convenient to introduce the normalized random variable

$$Y = \frac{X - m}{\sigma}.$$

Definition 9.46. *Central Moments*: A generalization of the variance is the nth central moment which is defined to be

$$\mu_n = \mathbb{E}\left[(X - \mathbb{E}X)^n\right].$$

(a) $\mu_1 = \mathbb{E}\left[X - \mathbb{E}X\right] = 0.$

(b) $\mu_2 = \sigma_X^2 = \operatorname{Var} X$: the second central moment is the variance.