Module 3: Distributions of random variables

Lecture 6: Expected Value and Variance

Derivative of OpenIntro Slides developed by Mine Çetinkaya-Rundel.

Overview

Table 1: Supplementary Reading

Topics	**Relevant Ch**
Expectation	OIS: 3.4.1, 3.4.3
Variance	OIS: 3.4.2, 3.4.4

RbE = R by Example

OIS = OpenIntro Statistics

Expected Values

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the *expected value* (mean), and it is a weighted average of the possible outcomes.
- The term expected value (EV) is somewhat misleading.
- The expected value of X is not necessarily
  - a value that you would *expect* X to take on
  - the most probable value
- If fact, the EV may be an unlikely value or even an impossible value! (e.g. the expected number of children born per woman in Canada is 1.58 children/woman).

Expected Values

Recall the following example from last lecture:

Example 1

Consider selecting a random student at UBCO among the 9000 registered students. Let X = the number of courses for which that student is registered. Suppose the pmf is given by:

Table 2: *pmf*

	x_1	x_2					x_7
x	1	2	3	4	5	6	7
$P(X = x)$	0.01	0.03	0.13	0.25	0.39	0.17	0.02
	p_1	p_2					p_7

Motivating Example

If we understand $p(x)$ to be the *proportion* of the finite population with each X value then we could write:

x	1	2	3	4	5	6	7
# students register to x courses	0.01∗ 9000	0.03∗ 9000	0.13∗ 9000	0.25∗ 9000	0.39∗ 9000	0.17∗ 9000	0.02∗ 9000
	90	270	1170	2250	3510	1530	180

To find the average number of course per student, we calculate the total number of courses and divide by the total number of students:

$$\frac{1 * 90 + 2 * 270 + 3 * 1170 + 4 * 2250 + 5 * 3510 + 6 * 1530 + 7 * 180}{9000}$$

$$= 4.57$$

Motivating Example

Notice how

$$\frac{1 * 90 + 2 * 270 + 3 * 1170 + 4 * 2250 + 5 * 3510 + 6 * 1530 + 7 * 180}{9000}$$

$$= 1\left(\frac{90}{9000}\right) + 2\left(\frac{270}{9000}\right) + 3\left(\frac{1170}{9000}\right) + 4\left(\frac{2250}{9000}\right)$$

$$+ 5\left(\frac{3510}{9000}\right) + 6\left(\frac{1530}{9000}\right) + 7\left(\frac{180}{9000}\right)$$

$$= 1 \cdot p(1) + 2 \cdot p(2) + 3 \cdot p(3) + 4 \cdot p(4) + 5 \cdot p(5) + 6 \cdot p(6) + 7 \cdot p(7)$$

$$= 1 \cdot 0.01 + 2 \cdot 0.03 + 3 \cdot 0.13 + 4 \cdot 0.25 + 5 \cdot 0.39 + 6 \cdot 0.17 + 7 \cdot 0.02$$

$$= 4.57$$

Motivating Example

- As it turns out, the population size is irrelevant in this question
- Alternatively, imagine that your population is made up of 1's, 2's, 3's 4's, 5's, 6's and 7's instead of students and $p(x)$ gives the proportion of x's in the population.
- Then, the average number of courses per student would be the average value of X's in the population having pmf given by Table 2.

Expected Value (EV)

Definition 1 (expected value)

The *expected value* *of a discrete random variable X is given by*

E[x]

$$\mathbb{E}[X] = \mu_X = \sum_x x \cdot P(X = x).$$

E(x)

over the support of X

The average or mean value of X or *expected value of X* can be denote by $\mathbb{E}[X]$ (or $E(X)$ or $E[X]$ or μ_X or simply μ if it is obvious we are talking about RV X). More generally...

Proposition 1

Note that, if $g(X)$ is some function of X, then

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

Some interpretations of $E(X)$

EV can be interpreted as a long run average, a weighted average of all possible values, or as a centre of mass for the distribution.

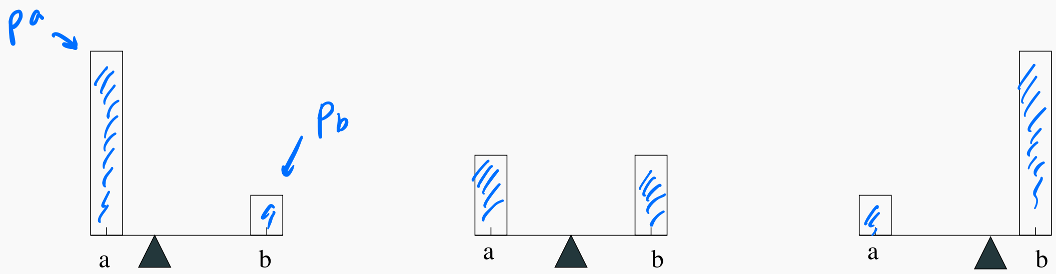
Long-run Average: $E(X)$ can be interpreted as the long-run average value of the random variable X over many independent repetitions of an example.

Weighted average: $E(X)$ can be viewed as a weighted average of the potential values where the “weights” are probabilities.

Centre of Mass: It can be visualized as a center of mass of the corresponding probability histogram.

Center of Mass interpretation

The picture below shows that pmf of X with various possible values of p_a and p_b. Consider a rod with weights placed at a and b, where the mass of the weights are p_a and p_b respectively. Then the rod has centre of mass or balance point equal to $E(X)$, i.e., the rod will balance at $E(X)$, shown below as a triangle.



Expected value of a discrete random variable

In a game of cards you win \$1 if you draw a heart, (not ace) \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning. Standard deck

Always unless specified otherwise.

X = winnings \$

pmf

x	1	5	10	0
$P(X=x)$	$\frac{12}{52}$	$\frac{4}{52}$	$\frac{1}{52}$	$\frac{35}{52}$
	draw a heart (non ace)	draw an ace	draw K♠	anything else

$$E[X] = \sum_{x} x \cdot P(X=x) = 1\left(\frac{12}{52}\right) + 5\left(\frac{4}{52}\right) + 10\left(\frac{1}{52}\right) + 0\left(\frac{35}{52}\right)$$

$$\approx 0.81 = \frac{42}{52}$$

$x \in \{0, 1, 5, 10\}$

support of x

Expected value of a discrete random variable

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning. $E[X]$

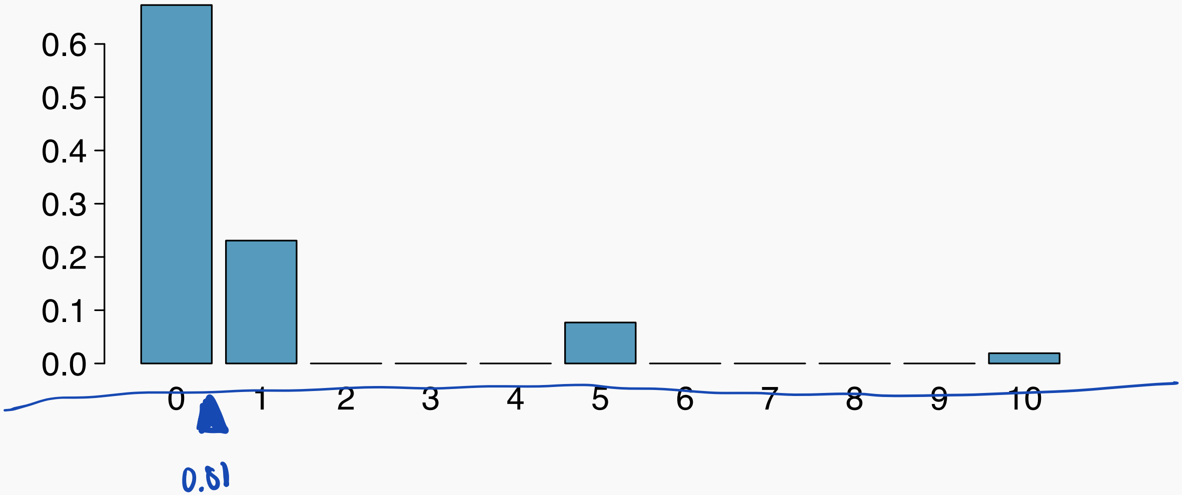
X = # of dollars we win

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

$$E[X] = \sum_{x \in \{0,1,5,10\}} x\,P(X=x) = 1\left(\frac{12}{52}\right) + 5\left(\frac{4}{52}\right) + 10\left(\frac{1}{52}\right) + 0\left(\frac{35}{52}\right) = \$0.81$$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:



On a related note, a *fair* game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

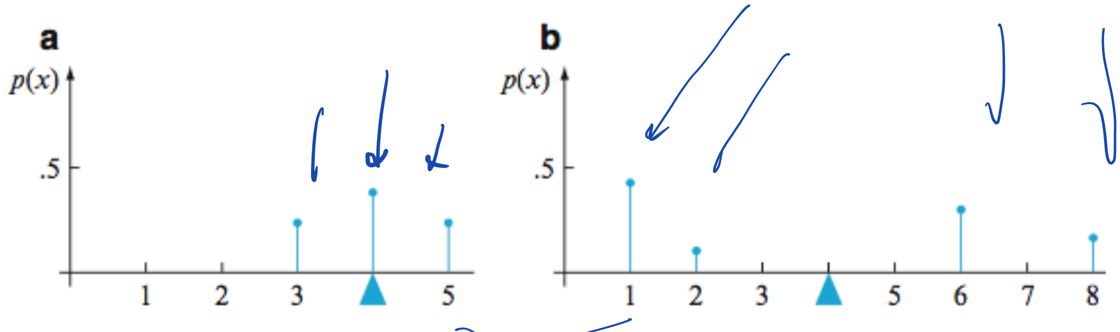


Figure 3.7 Two different probability distributions with $\mu = 4$

- The EV, i.e. μ, is a key characteristic that describes the distribution's center or "location"
- While these distributions have the same mean, they one on the right is much more spread out than the on the left.
- To get a better understanding of the shape of a distribution, we therefore often talk its *variance* as a measure of dispersion.

Variance and Standard Deviation

The variance is defined in terms of two expectations. In words, it's the expected value of the squared deviation of X from its mean.

Definition 2 (variance)

The variance of a random variable X is given

$$\sigma^2_{\cancel{X}} = Var(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \text{ or}$$

$$= \sum_{i=1}^{k} (x_i - \mathbb{E}(X))^2 P(X = x_i) \text{ (for discrete RV)}$$

Interpretation: If most of the possible values of the RV are close to (resp. far from) μ, the variance will be small (resp. large).

$E[X] = \mu_X$

Variance and Standard Deviation

SD is a related measure of the spread or scale.

Definition 3 (standard deviation)

The standard deviation (SD) of a random variable X can be written as follows:

$$SD(X) = \sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var[X]}.$$

You will see how to calculate these by hand on the next slide, however, in practice, we will be using a software like R to do these cumbersome calculations for us.

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$\mu = E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246} = 1.85$

Sometimes using the following shortcut will make this variance calculation easier.

x

Proposition 2 (Variance shortcut)

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

Proof: $Var(x) = E[(x - E[x])^2]$ by def

$E[x] = \text{constant} = \mu$

constant

$E[b] = b$

$= E[x^2 - 2xE[x] + (E[x])^2]$

$= E[x^2] - E[2xE[x]] + E[(E[x])^2]$ μ

$= E[x^2] - E[2\mu x] + \mu^2$

$= E[x^2] - 2\mu E[x] + \mu^2$

$\mu = E[x]$

$= E[x^2] - 2(E[x])^2 + (E[x])^2$

$$\boxed{= E[x^2] - (E[x])^2}$$

$\blacksquare$

	4	6	8
$P(X=x)$	0.5	0.3	0.2

$$E[g(x)] = \sum g(x) \cdot P(X=x)$$

$$E[X] = 4(0.5) + 6(0.3) + 8(0.2)$$
$$= 5.4$$

$$E[X^2] = \sum x^2 P(X=x)$$
$$= 4^2(0.5) + 6^2(0.3) + 8^2(0.2)$$
$$= 31.6$$

$$\Rightarrow \operatorname{var}(x) = E[X^2] - (E[X])^2 = 31.6 - (5.4)^2 = 2.44$$

$$\operatorname{var}(x) = (4 - 5.4)^2(0.5) + (6 - 5.4)^2(0.3) + (8 - 5.4)^2(0.2)$$
$$\Rightarrow \quad = \ldots$$
$$= 2.44$$

Linear combinations

- A *linear combination* of random variables X and Y is given by

RV

constant

$$aX + bY$$

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

where a and b are some fixed numbers.

- The average value of a linear combination of random variables is given by

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

$$y = aX + b$$

$$E[y] = E[aX+b]$$
$$= E[aX] + E[b]$$
$$= aE[X] + b$$

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and chemistry homework for the week?

X = total # of time spent on stat/chem homework

$X = 5S + 4C$

S = time it takes to complete a stat q

C = ... chem q

$$E[X] = E[5S + 4C]$$
$$= 5E[S] + 4E[C]$$
$$= 5(10) + 4(15)$$
$$= 50 + 60$$
$$= 110 \text{ min}$$

Calculating the expectation of a function of X

A computer store has purchased three computers at \$500 apiece. It will sell them for \$1000 apiece. The manufacturer has agreed to repurchase any computers still unsold after a specified period at \$200 apiece. Let X denote the number of computers sold, and suppose that $p(0) = 0.1$, $p(1) = 0.2$, $p(2) = 0.3$, and $p(3) = 0.4$. Find the expected profit, where profit is given by

X = # comp sold

$n = 3$

x	$P(X=x)$
0	0.1
1	0.2
2	0.3
3	0.4

$$\begin{aligned} h(X) &= \text{revenue} - \text{cost} \\ &= 1000X + 200(3 - X) - 1500 \\ &= 800X - 900. \end{aligned}$$

# of computers sold

# of computers unsold

$E[h(x)] = E[800X] - E[900]$

$= 800\,E[x] - 900$

$= 800(2) - 900 = \$700$

$E[x] = 0(0.1) + 1(0.2) + 2(0.3) + 3(0.4) = 2$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X) + \cancel{\mathrm{Var}(b)}^{\,0}$$

$$= a^2\,\mathrm{Var}(X)$$

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

$sd(aX + b)$

Note: If the random variables are not independent, the variance calculation gets a little more complicated.

Calculating the variance of a linear combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and chem ~~physics~~ homework for the week if you have 5 statistics and 4 chemistry homework problems assigned? Suppose that the time it takes to complete each problem is independent of another.

(a) 15.5

(b) 27.25

(c) 69.5

(d) 120.25

(e) None of the above

Calculating the variance of a linear combination

SD(S)

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and physics homework for the week if you have 5 statistics and 4 chemistry homework problems assigned? Suppose that the time it takes to complete each problem is independent of another.

$Var(S) = sd(S)^2$

(a) 15.5

(b) 27.25

(c) 69.5

(d) ~~120.25~~

(e) None of the above

$Y =$

$$V(5S + 4C)$$
$$= 5^2 \times V(S) + 4^2 \times V(C)$$
$$= 5^2 \times 1.5^2 + 4^2 \times 2^2 =$$ [redacted]

$= 120.25$

$sd = \sqrt{120.25} = 10.966$

Practice

A casino game costs \$5 to play. If the first card you draw is red, then you get to draw a second card (without replacement). If the second card is the ace of clubs, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(a) A profit of 5¢

(b) A loss of 10¢

(c) A loss of 25¢

(d) A loss of 30¢

$$P(\text{Win}) = P(R_1 \cap A♣) = P(R_1) \cdot P(A♣ \mid R_1) = \left(\frac{26}{52}\right) \cdot \left(\frac{1}{51}\right)$$

$$= 0.0098$$

X = Profit

$$P(\text{Loose}) = 1 - 0.0098 = 0.9902$$

X = profit

	x	$P(X=x)$
Win	$500 - 5 = 495$	0.0098
Loose	$0 - 5 = -5$	0.9902

$$E[\text{Profit}] = E[X] = 495(0.0098) + (-5)(0.9902) = -0.1$$

Practice

A casino game costs \$5 to play. If the first card you draw is red, then you get to draw a second card (without replacement). If the second card is the ace of clubs, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(a) A profit of 5¢

$\boxed{\text{(b) A loss of 10¢}}$

(c) A loss of 25¢

(d) A loss of 30¢

Event	Win	Profit: X	$P(X)$	$X \times P(X)$
Red, A♣	500	500 - 5 = 495	$\frac{26}{52} \times \frac{1}{51} = 0.0098$	$495 \times 0.0098 = 4.851$
Other	0	0 - 5 = -5	$1 - 0.0098 = 0.9902$	$-5 \times 0.9902 = -4.951$
				$E(X) = -0.1$

Clicker question

Is the game from the previous slide a "fair" game?

" fair "

$E[X] = 0$

a) Yes this game is fair

b) No this game is not fair

c) There is not enough information to answer

Clicker question

Is the game from the previous slide a “fair” game?

a) Yes this game is fair
b) No this game is not fair
c) There is not enough information to answer

Mean and Variance for Special Distributions

- The mean and variance are often quoted statistics when summarizing a distribution.

Parameters	$0 \le p \le 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
pmf	$\begin{cases} q = 1-p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1-p & \text{if } 0 \le k < 1 \\ 1 & \text{if } k \ge 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0,1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0, 1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Variance	$p(1-p) = pq$

Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$

Notation	$\mathrm{Pois}(\lambda)$
Parameters	$\lambda \in \mathbb{R}^+$ (rate)
Support	$k \in \mathbb{N}_0$
pmf	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \ge 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ

Figure 1: A selection of the Wikipedia summary page for the Bernoulli (left), Binomial (middle) and Poisson (right) distribution

Bernoulli Distribution Example

Example 2 (Mean/Variance of a Bernoulli)

Find the expectation and variance of X where $X \sim$ Bernoulli with success probability p.

$X \sim \text{Bernoulli}(p)$

	0	1
$P(X=x)$	$(1-p)$	p

$$E[X] = \sum_{x \in (0,1)} x P(X=x)$$

$$= 0\,P(X=0) + 1 \cdot P(X=1)$$

$$= 0(1-p) + 1p = p$$

$$\text{Var}(X) = \sum_{0,1} (x - E[X])^2 P(X=x)$$

$$= (0-p)^2(1-p) + (1-p)^2 p$$

$$= p^2(1-p) + (1-p)^2 p$$

$$= p(1-p)\left[\underbrace{\cancel{p} + (1-\cancel{p})}_{1}\right] \rightarrow p(1-p)$$

Exercises

Exercise 1 Let $X \sim \text{Poisson}(\lambda)$. Find the expected value and variance of X.

Exercise 2 Let $X \sim \text{Binomial(n,p)}$. Find the expected value and variance of X.

Exercise 3 The distribution of a random variable Y is:

y	50	100	200
$P(Y = y)$	0.2	0.5	0.3

Calculate the variance of Y.