CSE 312

Foundations of Computing II

Lecture 10: Variance and Independence of RVs

Anonymous questions: www.slido.com/[illegible]

2671111

Agenda

- Recap + LOTUS
- Variance
- Properties of Variance
- Independent Random Variables
- Properties of Independent Random Variables

Review Expected Value of a Random Variable

Definition. Given a discrete RV $X: \Omega \to \mathbb{R}$, the **expectation** or **expected value** or **mean** of X is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot P(\omega)$$

or equivalently

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x \cdot P(X = x) = \sum_{x \in \Omega_X} x \cdot p_X(x)$$

Intuition: "Weighted average" of the possible outcomes (weighted by probability)

Linearity of Expectation – Even stronger

Theorem. For any random variables $X_1, \dots, X_n$, and real numbers $a_1, \dots, a_n, b \in \mathbb{R}$,

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n + b] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n] + b.$$

Very important: In general, we do <u>not</u> have $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Linearity is special!

In general $\mathbb{E}[g(X)] \neq g\big(\mathbb{E}(X)\big)$

E.g., $X = \begin{cases} +1 \text{ with prob } 1/2 \\ -1 \text{ with prob } 1/2 \end{cases}$

Then: $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$

How DO we compute $\mathbb{E}[g(X)]$?

Expected Value of $g(X)$

g $Y = g(X)$ $\frac{X \to x}{Y \to g(x)}$

Definition. Given a discrete RV $X: \Omega \to \mathbb{R}$, the **expectation** or **expected value** or **mean** of $g(X)$ is

$$\mathbb{E}[g(X)] = \sum_{\omega \in \Omega} g\big(X(\omega)\big) \cdot P(\omega)$$

or equivalently

$$\mathbb{E}[\overset{Y}{g(X)}] = \sum_{x \in X(\Omega)} g(x) \cdot P(X = x) = \sum_{x \in \Omega_X} g(x) \cdot p_X(x)$$

Also known as **LOTUS:** "Law of the unconscious statistician

(nothing special going on in the discrete case)

$Y = 4X^2 - 1$

$X=0$ $X=1$ $X=3$

$Y=-1$ $Y=4\cdot1^2-1=3$ $Y=4\cdot3^2-1=35$

$g(x)=4x^2-1$

Example: Returning Homeworks

$$\mathbb{E}[X] = \sum_{x\in\Omega_X} x \cdot P(X = x)$$

$$\mathbb{E}[g(X)] = \sum_{x\in\Omega_X} g(x) \cdot P(X = x)$$

- Class with 3 students, randomly hand back homeworks. All permutations equally likely.
- Let X be the number of students who get their own HW

$\Pr(\omega)$	ω	$X(\omega)$
1/6	1, 2, 3	3
1/6	1, 3, 2	1
1/6	2, 1, 3	1
1/6	2, 3, 1	0
1/6	3, 1, 2	0
1/6	3, 2, 1	1

$$\mathbb{E}[X] = 3 \cdot P(X = 3) + 1 \cdot P(X = 1) + 0 \cdot P(X = 0)$$

$$P_X(x) = \begin{cases} \frac{1}{3} & x=0 \\ \frac{1}{2} & x=1 \\ \frac{1}{6} & x=3 \\ 0 & \text{otherwise} \end{cases}$$

$E(Y) = g(3)P(X=3) + g(1)P(X=1) + g(0)P(X=0)$

$= 35\cdot\frac{1}{6} + 3\cdot\frac{1}{2} + (-1)\frac{1}{3}$

$=$

$g(X)$

Agenda

- LOTUS
- Variance
- Properties of Variance
- Independent Random Variables
- Properties of Independent Random Variables

Which game would you rather play?

Game 1: *In every round, you win \$2 with probability 1/3, lose \$1 with probability 2/3.*

W_1 = payoff in a round of Game 1

$P(W_1 = 2) = \frac{1}{3}$, $P(W_1 = -1) = \frac{2}{3}$

$E(W_1) = 2 \cdot \frac{1}{3} + (-1) \cdot \frac{2}{3} = 0$

Which game would you rather play?

Game 1: *In every round, you win \$2 with probability 1/3, lose \$1 with probability 2/3.*

W_1 = payoff in a round of Game 1

$P(W_1 = 2) = \frac{1}{3}$, $P(W_1 = -1) = \frac{2}{3}$

$\mathbb{E}[W_1] = 0$

Game 2: *In every round, you win \$10 with probability 1/3, lose \$5 with probability 2/3.*

W_2 = payoff in a round of Game 2

$P(W_2 = 10) = \frac{1}{3}$, $P(W_2 = -5) = \frac{2}{3}$

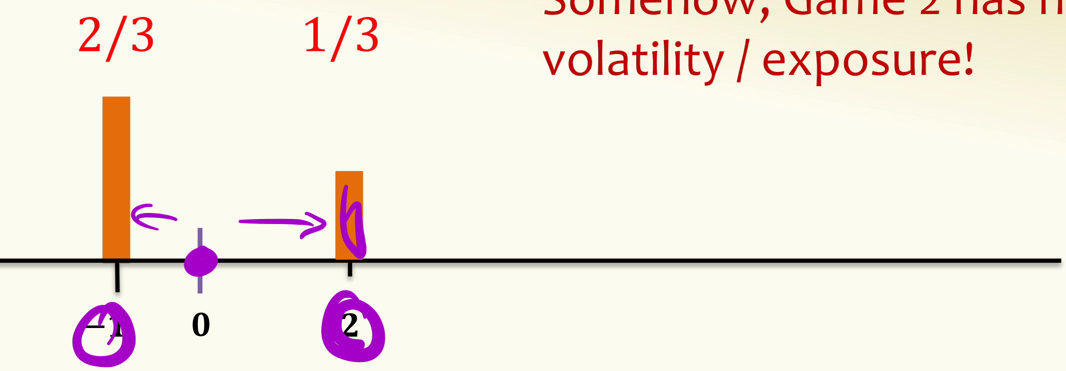
$\mathbb{E}[W_2] = 0$

$E(W_2) = 10 \cdot \frac{1}{3} - 5 \cdot \frac{2}{3} = 0$

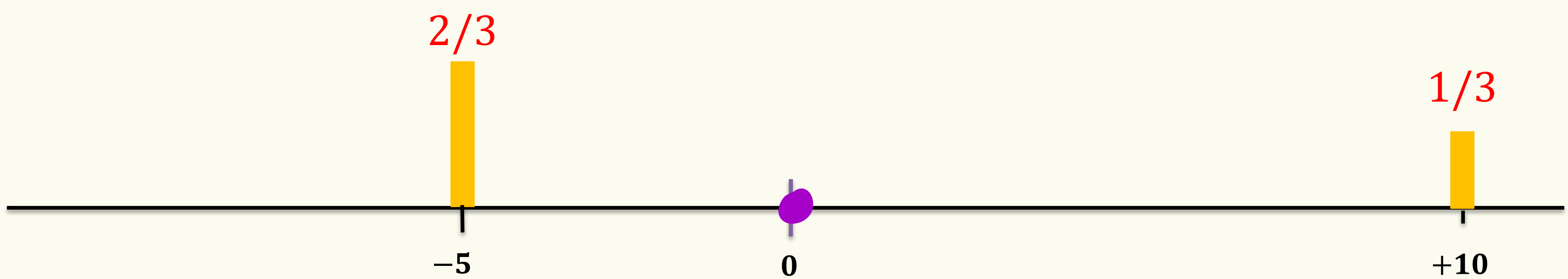
Two Games

Somehow, Game 2 has higher volatility / exposure!

$P(W_1 = 2) = \frac{1}{3}$, $P(W_1 = -1) = \frac{2}{3}$



$P(W_2 = 10) = \frac{1}{3}$, $P(W_2 = -5) = \frac{2}{3}$



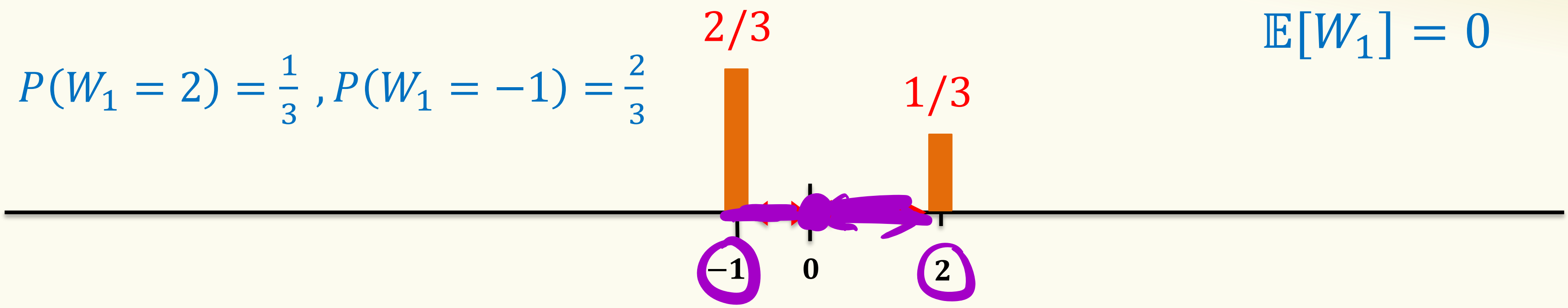
Same expectation, but clearly a very different distribution.

We want to capture the difference – New concept: **Variance**

Variance (Intuition, First Try)

$P(W_1 = 2) = \frac{1}{3}$, $P(W_1 = -1) = \frac{2}{3}$

$\mathbb{E}[W_1] = 0$



New quantity (random variable): How far from the expectation?

$W_1 - \mathbb{E}[W_1]$

$$E(W_1 - E(W_1))$$

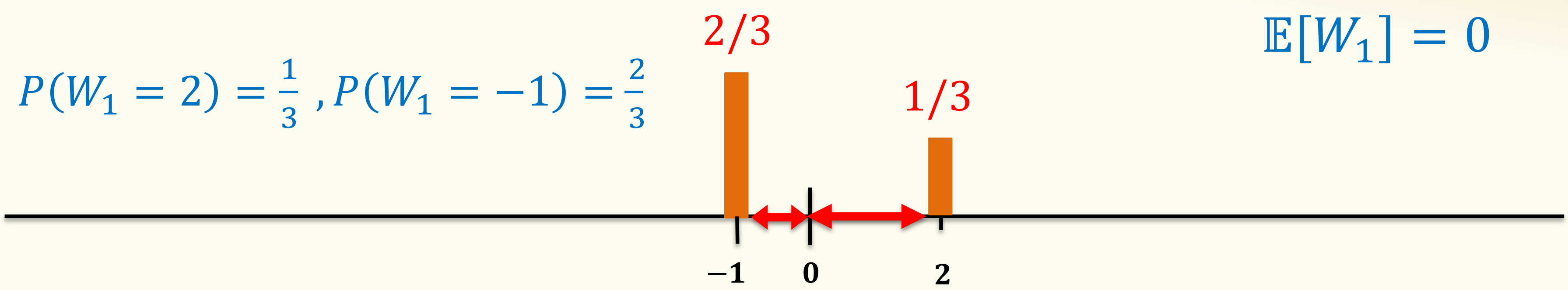
$$\underset{\text{LOE}}{=} E(W_1) - E(E(W_1))$$

$$= E(W_1) - E(W_1) = 0$$

Variance (Intuition, First Try)

$P(W_1 = 2) = \frac{1}{3}, P(W_1 = -1) = \frac{2}{3}$

$\mathbb{E}[W_1] = 0$



New quantity (random variable): How far from the expectation?

$W_1 - \mathbb{E}[W_1]$

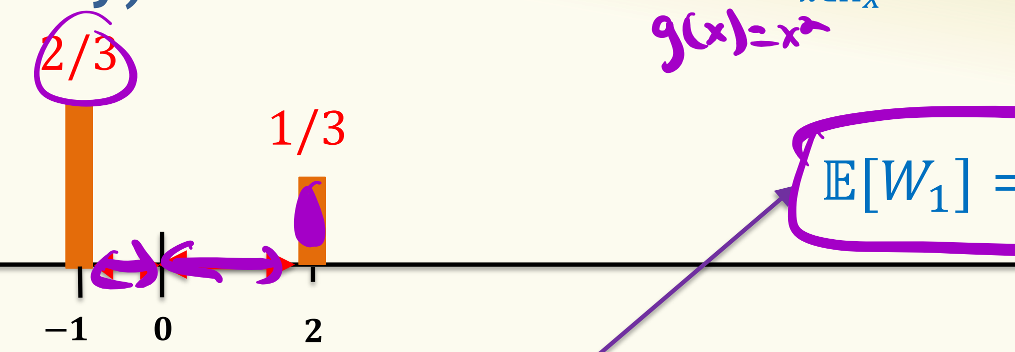
$$\begin{aligned}
&\mathbb{E}[W_1 - \mathbb{E}[W_1]] \\
&\quad = \mathbb{E}[W_1] - \mathbb{E}\big[\mathbb{E}[W_1]\big] \\
&\quad = \mathbb{E}[W_1] - \mathbb{E}[W_1] \\
&\quad = 0
\end{aligned}$$

Variance (Intuition, Better Try)

$$\mathbb{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot P(X = x)$$

$g(x) = x^2$

$P(W_1 = 2) = \frac{1}{3}, P(W_1 = -1) = \frac{2}{3}$



$\mathbb{E}[W_1] = 0$

A better quantity (random variable): How far from the expectation?

$$\mathbb{E}[(W_1 - \mathbb{E}[W_1])^2]$$

$= \mathbb{E}(W_1^2)$

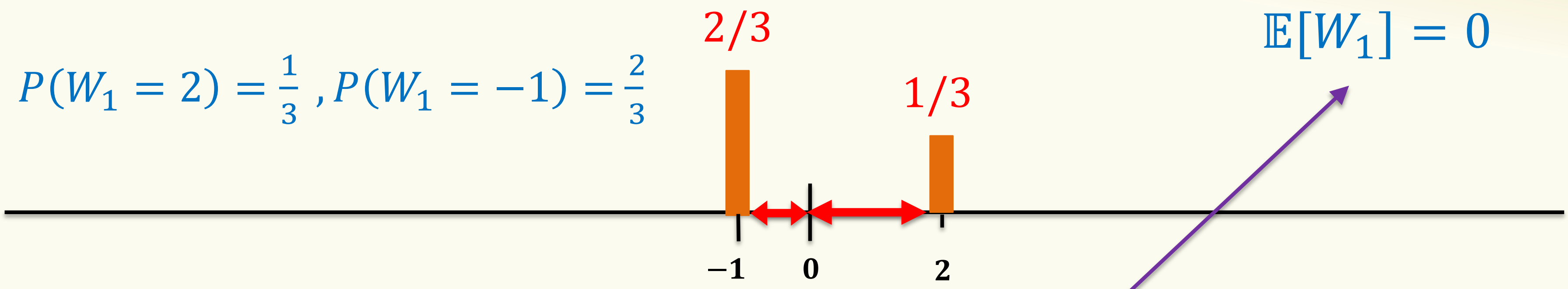
$= 2^2 \cdot \frac{1}{3} + (-1)^2 \cdot \frac{2}{3}$

$= 2$

Variance (Intuition, Better Try)

$P(W_1 = 2) = \frac{1}{3}, P(W_1 = -1) = \frac{2}{3}$

$\mathbb{E}[W_1] = 0$



A better quantity (random variable): How far from the expectation?

$$\mathbb{E}[(W_1 - \mathbb{E}[W_1])^2]$$

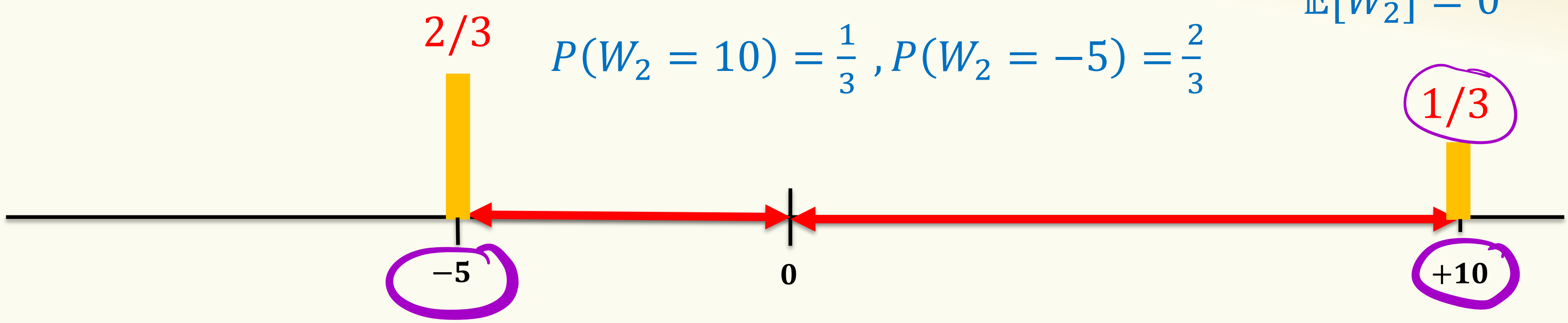
$$= \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 4$$

$$= 2$$

Variance (Intuition, Better Try)

$\mathbb{E}[W_2] = 0$

$P(W_2 = 10) = \frac{1}{3}$, $P(W_2 = -5) = \frac{2}{3}$



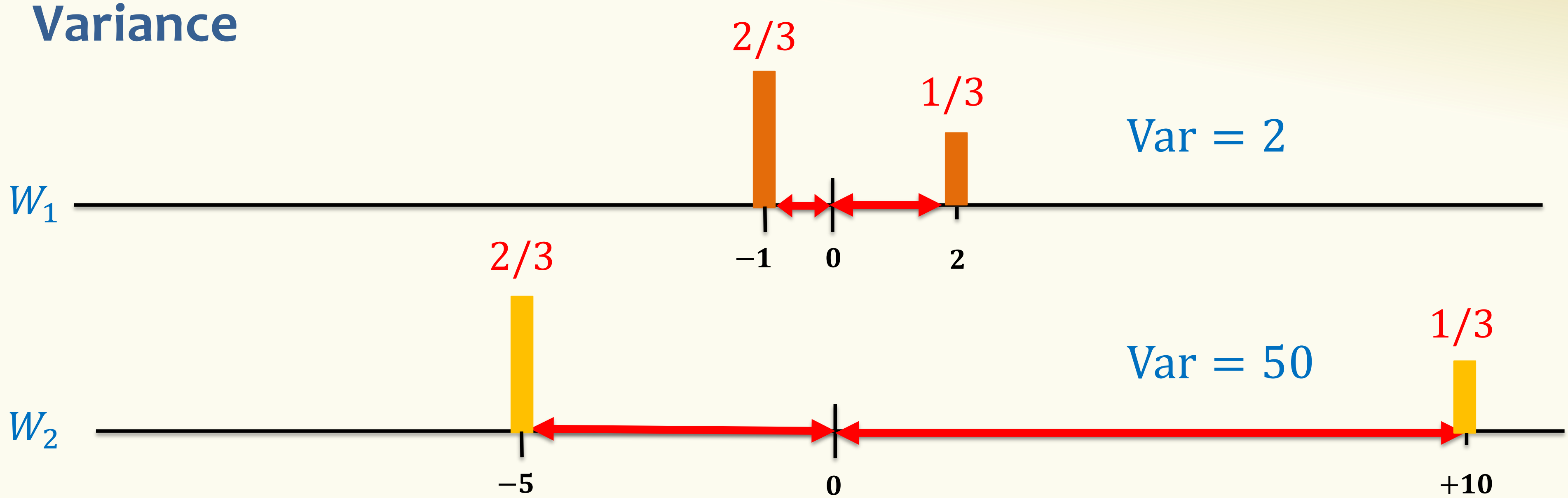
A better quantity (random variable): How far from the expectation?

$$\mathbb{E}[(W_2 - \cancelto{0}{\mathbb{E}[W_2]})^2]$$

$$E(w_2^2) \quad = \frac{2}{3} \cdot 25 + \frac{1}{3} \cdot 100$$

$$= 50$$

Variance



We say that W_2 has **"higher variance"** than W_1.

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

Variance

$Y = g(X)$ $g(x) = (x - E(X))^2$

Definition. The **variance** of a (discrete) RV X is

$$\mathrm{Var}(X) = \mathbb{E}[\underbrace{(X - \mathbb{E}[X])^2}_{Y}] = \sum_{x} p_X(x) \cdot (x - \mathbb{E}[X])^2$$

$x \in \Omega_X$

Standard deviation: $\sigma(X) = \sqrt{\mathrm{Var}(X)}$

Recall $\mathbb{E}[X]$ is a **constant**, not a random variable itself.

Intuition: Variance (or standard deviation) is a quantity that measures, in expectation, how "far" the random variable is from its expectation.

Variance – Example 1

$$\mathbb{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot P(X = x)$$

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

X fair die

- $P(X = 1) = \cdots = P(X = 6) = 1/6$
- $\mathbb{E}[X] = 3.5$

$$\mathrm{Var}(X) = \sum_{x=1}^{6} P(X = x) \cdot (x - \mathbb{E}[X])^2$$

$$\frac{1}{6}(1-3.5)^2 + \frac{1}{6}(2-3.5)^2 + \cdots + \frac{1}{6}(6-3.5)^2$$

$$=$$

Variance – Example 1

X fair die

- $P(X = 1) = \cdots = P(X = 6) = 1/6$
- $\mathbb{E}[X] = 3.5$

$$\mathrm{Var(X)} = \sum_x P(X = x) \cdot (x - \mathbb{E}[X])^2$$

$$= \frac{1}{6}[(1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2]$$

$$= \frac{2}{6}[2.5^2 + 1.5^2 + 0.5^2] = \frac{2}{6}\left[\frac{25}{4} + \frac{9}{4} + \frac{1}{4}\right] = \frac{35}{12} \approx 2.91677 \ldots$$

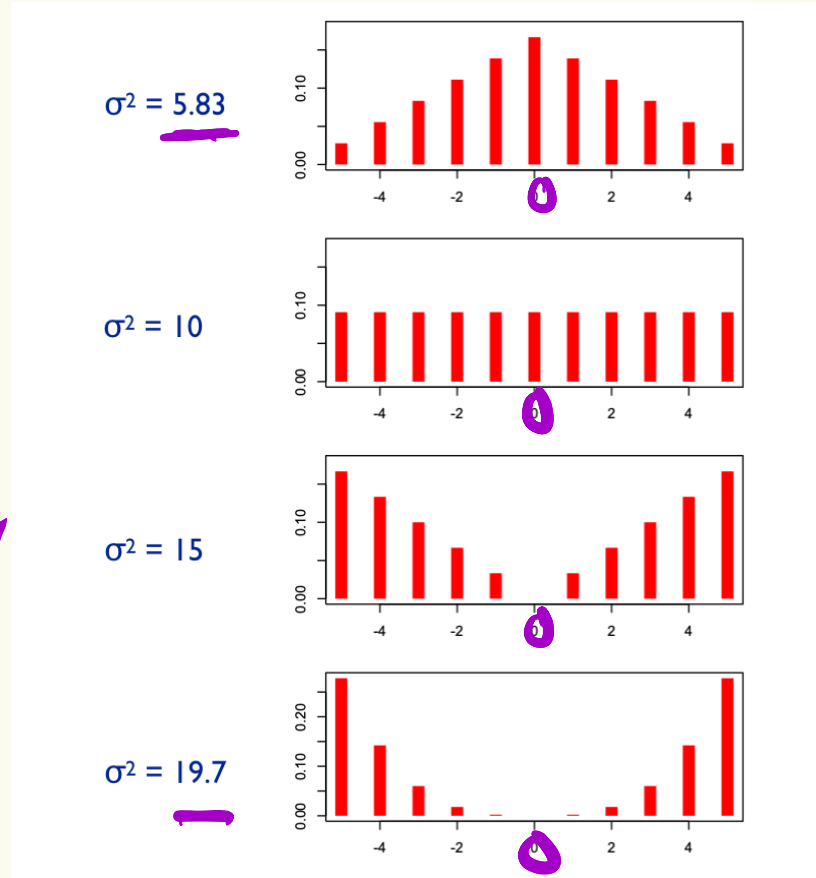
$Var(X) = \sigma^2(X)$

Std dev σ

Variance in Pictures

Captures how much "spread' there is in a pmf

All pmfs have same expectation



Agenda

- LOTUS
- Variance
- Properties of Variance
- Independent Random Variables
- Properties of Independent Random Variables

Variance – Properties

Definition. The **variance** of a (discrete) RV X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x p_X(x) \cdot (x - \mathbb{E}[X])^2$$

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

$(a-b)^2 = a^2 - 2ab + b^2$

$$\mathrm{Var}(X) = E\Big[\big(\overset{a}{X} - \overset{b}{E(X)}\big)^2\Big]$$

$$= E\big[X^2 - 2XE(X) + (E(X))^2\big]$$

LOE $\quad = E(X^2) + E(-2E(X)X) + E\big(E(X)^2\big)$

$E(aX) = aE(X)$

$$= E(X^2) - 2\underbrace{E(X)\,E(X)}_{E(X)^2} + E(X)^2$$

$= E(X^2) - 2E(X)^2 + E(X)^2$

Variance

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

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

Recall $\mathbb{E}[X]$ is a **constant**

$$= \mathbb{E}[X^2 - 2\mathbb{E}[X] \cdot X + \mathbb{E}[X]^2]$$

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

(linearity of expectation!)

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

$\mathbb{E}[X^2]$ and $\mathbb{E}[X]^2$ are different !

Variance – Example 1

X fair die

- $\mathbb{P}(X=1)=\cdots=\mathbb{P}(X=6)=1/6$
- $\mathbb{E}[X]=\frac{21}{6}$
- $\boxed{\mathbb{E}[X^2]}=\frac{91}{6}$ $= 1^2 \cdot \frac{1}{6} + 2^2 \frac{1}{6} + \cdots + 6^2 \frac{1}{6}$

$$\mathrm{Var}(X)=\mathbb{E}[X^2]-\mathbb{E}[X]^2=\frac{91}{6}-\left(\frac{21}{6}\right)^2=\frac{105}{36}\approx 2.91677$$

Variance – Properties

$Z = aX$

$E(Z) = aE(X)$

$Y = X + b.$

$E(Y) = E(X) + b$

Definition. The **variance** of a (discrete) RV X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x p_X(x) \cdot (x - \mathbb{E}[X])^2$$

$ax \quad aE(X)$

$[a(X - E(X))]^2$

$\mathrm{Var}(X+b) = \mathrm{Var}(X)$

const.

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

Theorem. For any $a, b \in \mathbb{R}$, $\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X)$

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

$X_A \nearrow 1 \quad X_A^2 = 1$

$\searrow 0 \quad 0$

Variance of Indicator Random Variables

Suppose that X_A is an indicator RV for event A with $P(A) = p$ so

$$\boxed{\mathbb{E}[X_A] = P(A) = p}$$

$$\mathbb{E}(X_A^2) = p$$

$$\mathrm{Var}(X_A) = \underbrace{\mathbb{E}[X_A^2]}_{p} - \underbrace{\mathbb{E}[X_A]^2}_{p^2} = p - p^2 = p(1-p)$$

Variance of Indicator Random Variables

Suppose that X_A is an indicator RV for event A with $P(A) = p$ so

$$\mathbb{E}[X_A] = P(A) = p$$

Since X_A only takes on values 0 and 1, we always have $X_A^2 = X_A$ so

$$\mathrm{Var}(X_A) = \mathbb{E}[X_A^2] - \mathbb{E}[X_A]^2 = \mathbb{E}[X_A] - \mathbb{E}[X_A]^2 = p - p^2 = p(1-p)$$

$$\mathbb{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot P(X = x)$$

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \textstyle\sum_x p_X(x) \cdot (x - \mathbb{E}[X])^2$$

$$\mathrm{Var}(X) = E(X^2) - E(X)^2$$

In General, $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proof by counter-example:

- Let X be a r.v. with pmf $P(X = 1) = P(X = -1) = 1/2$
  - What is $\mathbb{E}[X]$ and $\mathrm{Var}(X)$?

$E(X) = 1 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} = 0$

$\mathrm{Var}(X) = E(X^2) = 1^2 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2} = 1$

- Let $\boxed{Y = -X}$
  - What is $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$?

$E(Y) = 0$ $\quad$ $\mathrm{Var}(Y) = 1$

$\mathrm{Var}(X) + \mathrm{Var}(Y) = 2$

What is $\mathrm{Var}(X + Y)$?

$\mathrm{Var}(X + (-X)) = \mathrm{Var}(0) = 0$


Brain Break

Agenda

- LOTUS
- Variance
- Properties of Variance
- Independent Random Variables
- Properties of Independent Random Variables

$\Omega_X = \{-1, 1\}$ $\Omega_Y = \{1, 3, 5\}$

Random Variables and Independence

Comma is shorthand for AND

Definition. Two random variables X, Y are **(mutually) independent** if for all x, y,

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

Intuition: Knowing X doesn't help you guess Y and vice versa

$P(A \cap B) = P(A)P(B)$

Definition. The random variables $X_1, \ldots, X_n$ are **(mutually) independent** if for all $x_1, \ldots, x_n$,

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

Note: No need to check for all subsets, but need to check <u>for all outcomes</u>!

Example

$n = 5$

$\Omega_X = \{0, 1, 2, 3, 4, 5\}$

$\Omega_Y = \{0, 1\}$

Let X be the number of heads in n *independent* coin flips of the same coin. Let $Y = X \bmod 2$ be the parity (even/odd) of X.

Are X and Y independent?

$x \in \Omega_X, \ y \in \Omega_Y$

$$P(X = x, Y = y) \overset{?}{\neq} P(X = x)P(Y = y)$$

$$P(X = 2, Y = 1) = 0$$

$$P(X = 2) \neq 0$$

$$P(Y = 1) \neq 0$$

Example

p. of H.

Make $2n$ *independent* coin flips of the same coin.

Let X be the number of heads in the first n flips and Y be the number of heads in the last n flips.

Are X and Y independent?

$$P(X=i, Y=j) = P(X=i)\,P(Y=j)$$

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}$$

$$p^i (1-p)^{n-i}$$

$\underbrace{HH\ldots H}_{i}\ \underbrace{T \ldots T}_{n-i}$

Agenda

- LOTUS
- Variance
- Properties of Variance
- Independent Random Variables
- Properties of Independent Random Variables

Important Facts about Independent Random Variables

Theorem. If X, Y independent, $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Theorem. If X, Y independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Corollary. If $X_1, X_2, \ldots, X_n$ mutually independent,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i}^{n} \mathrm{Var}(X_i)$$

(Not Covered) Proof of $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Theorem. If X, Y independent, $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Proof

Let $x_i, y_i, i = 1, 2, \ldots$ be the possible values of X, Y.

$$\begin{aligned}
\mathbb{E}[X \cdot Y] &= \sum_i \sum_j x_i \cdot y_j \cdot P(X = x_i \wedge Y = y_j) \\
&= \sum_i \sum_j x_i \cdot y_i \cdot P(X = x_i) \cdot P(Y = y_j) \\
&= \sum_i x_i \cdot P(X = x_i) \cdot \left(\sum_j y_j \cdot P(Y = y_j) \right) \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y]
\end{aligned}$$

independence

Note: *NOT* true in general; see earlier example $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$

(Not Covered) Proof of $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Theorem. If X, Y independent, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Proof

$Var(X+Y)$

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

$= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2$

$= \mathbb{E}[X^2] + 2\,\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2\,\mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[Y]^2)$

$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 + 2\,\mathbb{E}[XY] - 2\,\mathbb{E}[X]\,\mathbb{E}[Y]$

$= Var(X) + Var(Y) + 2\,\mathbb{E}[XY] - 2\,\mathbb{E}[X]\,\mathbb{E}[Y]$

$= Var(X) + Var(Y)$

linearity

equal by independence

Example – Coin Tosses

We flip n *independent* coins, each one heads with probability p

- $X_i = \begin{cases} 1, & i^{\text{th}} \text{ outcome is heads} \\ 0, & i^{\text{th}} \text{ outcome is tails.} \end{cases}$
- $Z =$ number of heads

Fact. $Z = \sum_{i=1}^{n} X_i$

$P(X_i = 1) = p$
$P(X_i = 0) = 1 - p$

$P(Z = k) =$

What is $\mathbb{E}[Z]$? What is $\mathrm{Var}(Z)$?

Example – Coin Tosses

We flip n *independent* coins, each one heads with probability p

- $X_i = \begin{cases} 1, & i^{\text{th}} \text{ outcome is heads} \\ 0, & i^{\text{th}} \text{ outcome is tails.} \end{cases}$
- $Z =$ number of heads

Fact. $Z = \sum_{i=1}^{n} X_i$

$P(X_i = 1) = p$
$P(X_i = 0) = 1 - p$

$P(Z = k) = \binom{n}{k} p^k (1-p)^{n-k}$

What is $\mathbb{E}[Z]$? What is $\mathrm{Var}(Z)$?

Note: $X_1, \dots, X_n$ are <u>mutually</u> independent! [Verify it formally!]

$$\mathrm{Var}(Z) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n \cdot p(1-p)$$

Note $\mathrm{Var}(X_i) = p(1-p)$

Questions

The **variance** of a (discrete) RV X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x p_X(x) \cdot (x - \mathbb{E}[X])^2$$

- Can the variance of a random variable be negative?
- Is $\mathrm{Var}(X + 5) = \mathrm{Var}(X) + 5$?
- Is it true that if $\mathrm{Var}(X) = 0$, then X is a constant?
- What is the relationship between $E(X^2)$ and $[E(X)]^2$?

Independence of random variables

- Suppose X and Y are independent indicator random variables taking the value 1 with probability ½, and let $Z = XY$
  - Are X and Z independent?
  - Are Y and Z independent?

- Is it true that if X and Y are independent, and Y and Z are independent, then X and Z are independent?

Agenda

- LOTUS
- Variance
- Properties of Variance
- Independent Random Variables
- Properties of Independent Random Variables
- An Application: Bloom Filters!

Basic Problem

Problem: Store a subset S of a large set U.

Example. $U =$ set of 128 bit strings $\quad |U| \approx 2^{128}$

$S =$ subset of strings of interest $\quad |S| \approx 1000$

Two goals:

1. **Very fast** (ideally constant time) answers to queries "Is $x \in S$?" for any $x \in U$.
2. **Minimal storage** requirements.

Naïve Solution I – Constant Time

Idea: Represent S as an array A with 2^{128} entries.

$$\mathrm{A}[x] = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}$$

$S = \{0, 2, \dots, K\}$

0	1	2	...	K	...		
1	0	1	0	1	...	0	0

Membership test: To check. $x \in S$ just check whether $\mathrm{A}[x] = 1$.

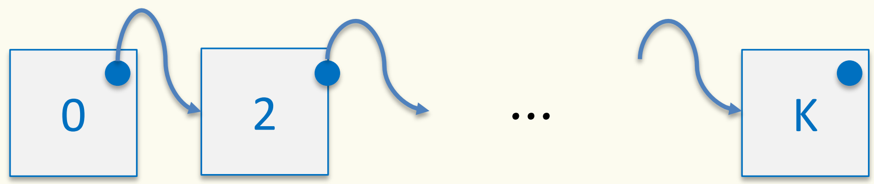
$\longrightarrow$ **constant time!** 👍 😀

Storage: Require storing 2^{128} bits, even for small S. 👎 😢

Naïve Solution II – Small Storage

Idea: Represent S as a list with $|S|$ entries.

$S = \{0,2, \ldots, K\}$



Storage: Grows with $|S|$ only 👍 😀

Membership test: Check $x \in S$ requires time linear in $|S|$

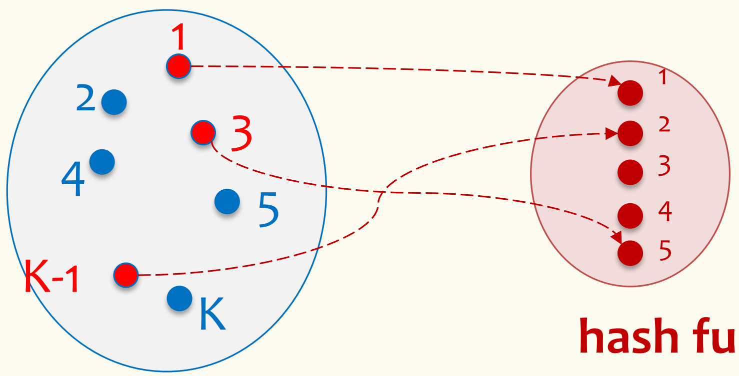
(Can be made logarithmic by using a tree) 👎 😢

Hash Table

Idea: Map elements in S into an array A of size m using a hash function $\mathbf{h}$

Membership test: To check $x \in S$ just check whether $A[\mathbf{h}(x)] = x$

Storage: m elements (size of array)



hash function h: $U \to [m]$

Hash Table

Idea: Map elements in S into an array A of size m using a hash function $\mathbf{h}$

Membership test: To check $x \in S$ just check whether $A[\mathbf{h}(x)] = x$

Storage: m elements (size of array)

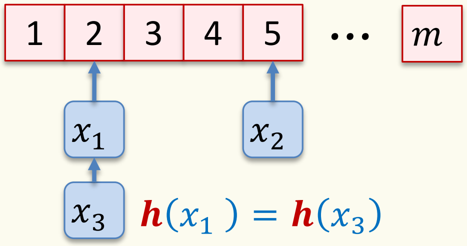
Challenge 1: Ensure $\boldsymbol{h}(x) \neq \boldsymbol{h}(y)$ *for most* $x, y \in S$

Challenge 2: Ensure $m = O(|S|)$

Hashing: collisions

Collisions occur when $\boldsymbol{h}(x) = \boldsymbol{h}(y)$ for some distinct $x, y \in S$, i.e., two elements of set map to the same location

- Common solution: <u>chaining</u> – at each location (bucket) in the table, keep linked list of all elements that hash there.



Good hash functions to keep collisions low

- The hash function $\boldsymbol{h}$ is good iff it
  - distributes elements uniformly across the m array locations so that
  - pairs of elements are mapped independently

"Universal Hash Functions" – see CSE 332

Hashing: summary

Hash Tables

- They store the data itself
- With a good hash function, the data is well distributed in the table and lookup times are small.
- However, they need at least as much space as all the data being stored, i.e., $m = \Omega(|S|)$

In some cases, $|S|$ is huge, or not known a-priori …

Can we do better!?

Next time: Bloom Filters

- Probabilistic data structure.
- Close cousins of hash tables.
  - But: Ridiculously space efficient
- Occasional errors, specifically false positives.