

# CHAPTER 3: DISTRIBUTIONS OF RANDOM VARIABLES (PART 1)

Day 6 topics:

- Section 3.1: Random variables
- Section 3.2: Binomial distribution

EDA  
 ↓  
 Center, shape,  
 and spread  
 of a distribution

## 3.1. Random variables.

**Definition 3.1.** A **random variable (r.v.)** assigns numerical values to the outcome of a random phenomenon.

Notation:

A random variable is usually denoted with a capital letter such as  $X$ ,  $Y$ , or  $Z$ .

**Coin toss example:**  $\begin{cases} X=1 & \text{if toss heads} \\ X=-1 & \text{if toss tails} \end{cases}$

### Example 3.2. Data points

Suppose you have a dataset of size 3 ( $n = 3$ ) with  $x_1 = 5$ ,  $x_2 = 3$ , and  $x_3 = 6$ .

- Each data point is the outcome of a random phenomenon
- Each data point is a numerical value
- The data points are examples of values of a sequence of random variables  $X_1$ ,  $X_2$ , and  $X_3$
- For datasets, we almost always assume the data points came from random variables that are independent and have the same distribution.
- To calculate the likelihood of data, we need to know the distribution of the random variable that models the data.
- First, let's remind ourselves how to calculate them mean and variance of a dataset:

average

- What is the mean of the data points?

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3}{3} = \frac{5 + 3 + 6}{3} = \frac{14}{3} = 4.67$$

- What are the variance and standard deviation of the data points?

$$\begin{aligned} \text{Variance } s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(5 - \frac{14}{3})^2 + (3 - \frac{14}{3})^2 + (6 - \frac{14}{3})^2}{3-1} \\ &= 2.33 \text{ units}^2 \\ \text{SD } s &= \sqrt{s^2} = \sqrt{2.33} = 1.53 \text{ units} \end{aligned}$$

### Example 3.3. Rolling a die

Suppose you roll a fair die. Let the random variable (r.v.)  $X$  be the outcome of the roll, i.e. the value of the face showing on the die.

(1) What is the probability distribution of the r.v.  $X$ ?

$x$	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$P(X=x) = \frac{1}{6} \quad \text{for } x=1, 2, \dots, 6$$

expected value

(2) What is the expected outcome of the r.v.  $X$ ?

Weighted average =  $\frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$  ↗ Not a possible outcome!

$$= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)$$

(3) Now suppose the 6-sided die is not fair. How would we calculate the expected outcome?

expected value

$$= 1(0.10) + 2(0.20) + 3(0.05) + 4(0.05) + 5(0.25) + 6(0.35)$$

$x$	$P(X=x)$
1	0.10
2	0.20
3	0.05
4	0.05
5	0.25
6	0.35

40% ← 60%

$$= 4.2$$

(From Textbook § 2.1.5)

**Definition 3.4.** A **probability distribution** consists of all disjoint **outcomes** and their associated **probabilities**.

### Rules for a probability distribution

A probability distribution is a list of all possible outcomes and their associated probabilities that satisfies three rules:

- (1) The outcomes listed must be **disjoint**.
- (2) Each probability must be **between 0 and 1**.
- (3) The probabilities must **total to 1**.

Probability distributions are usually either **discrete** or **continuous**, depending on whether the random variable is discrete or continuous.

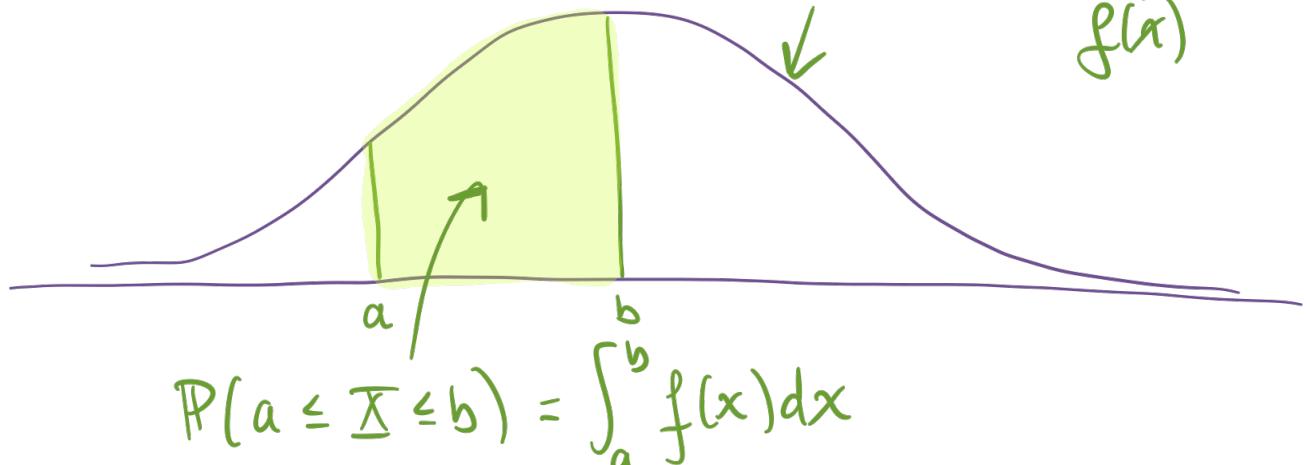
(Back to Textbook § 3.1.1)

**Definition 3.5.** A **discrete** r.v.  $X$  takes on a finite number of values or **countably infinite** number of possible values. **hospital stays**

**Definition 3.6.** A **continuous** r.v.  $X$  can take on any real value in an interval of values or unions of intervals. **height**

Normal distribution

distribution function  
 $f(x)$



## § 3.1.2 Expectation For discrete random variables (r.v.'s)

- We call the mean of a random variable its **expected value**
- The expected value is calculated as a **weighted average**

**Definition 3.7. Expected value** of a discrete random variable

If  $X$  takes on outcomes  $x_1, \dots, x_k$  with probabilities  $P(X = x_1), \dots, P(X = x_k)$ , the expected value of  $X$  is the sum of each outcome multiplied by its corresponding probability:

$$\mu = E[X] = x_1 P(X=x_1) + x_2 P(X=x_2) + \dots + x_k P(X=x_k)$$

↑ "mu"

$$= \sum_{i=1}^k x_i P(X=x_i)$$

## § 3.1.3 Variability of random variables

- Just like with data, the variability of a r.v. is described with its variance or standard deviation.
- Squared deviations from the mean are weighted by their respective probabilities

**Definition 3.8. Variance** of a discrete random variable

If  $X$  takes on outcomes  $x_1, \dots, x_k$  with probabilities  $P(X = x_1), \dots, P(X = x_k)$  and expected value  $\mu = E(X)$ , then the variance of  $X$ , denoted by  $\text{Var}(X)$  or  $\sigma^2$ ,

is

$$\sigma^2 = \text{Var}(X) = \sum_{i=1}^k (x_i - \mu)^2 P(X=x_i)$$

↑ "sigma"

$$= (x_1 - \mu)^2 P(X=x_1) + (x_2 - \mu)^2 P(X=x_2) + \dots + (x_k - \mu)^2 P(X=x_k)$$

**Definition 3.9. Standard deviation** of a discrete random variable

The standard deviation of  $X$ , labeled  $SD(X)$  or  $\sigma$ , is

$$\sigma = SD(X) = \sqrt{\text{Var}(X)}$$

$s = \text{sample SD}$   
 $\sigma = \text{population SD}$

**Example 3.10. Rolling a fair die: variance**

Suppose you roll a **fair 6-sided die**. Let the random variable (r.v.)  $X$  be the outcome of the roll, i.e. the value of the face showing on the die.

Find the **variance and standard deviation** of  $X$ .

$$\sigma^2 = \sum_{i=1}^k (x_i - \mu)^2 \cdot P(X=x_i)$$

$x$	$P(X=x)$	$(x_i - \mu)^2$	$P(X=x_i)$
1	1/6	$(1-3.5)^2$	$\cdot 1/6$
2	1/6	$+ (2-3.5)^2$	$\cdot 1/6$
3	1/6	$+ (3-3.5)^2$	$\cdot 1/6$
4	1/6	$+ (4-3.5)^2$	$\cdot 1/6$
5	1/6	$+ (5-3.5)^2$	$\cdot 1/6$
6	1/6	$+ (6-3.5)^2$	$\cdot 1/6$

$$\text{Total } \sigma^2 = 2.92 = \sum_{i=1}^k (x_i - 3.5)^2 \cdot 1/6$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

## Class discussion

### Example 3.11. Vaccinated people testing positive for Covid-19

About 25% of people that test positive for Covid-19 are vaccinated for Covid-19.

Define the r.v.  $X$  to be 1 if someone that tests positive is vaccinated and 0 if they are not vaccinated.

(1) Make a table for the probability distribution for the r.v.  $X$

(2) What is the expected value of  $X$ ?

(3) What is the variance of  $X$ ?

### § 3.1.4 Linear combinations of random variables

Line  $y = mx + b$  constants

**Definition 3.12.** **Linear combinations** of random variables.

If  $X$  and  $Y$  are random variables and  $a$  and  $b$  are constants, then

$$aX + bY + cZ + d$$

Example:  $\bar{X}$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$= \left(\frac{1}{n}\right)X_1 + \left(\frac{1}{n}\right)X_2 + \dots + \left(\frac{1}{n}\right)X_n$$

is a linear combination of the random variables.

**Theorem 3.13.** **Expected value of a linear combination** of random variables.

If  $X$  and  $Y$  are random variables and  $a$  and  $b$  are constants, then

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y] = a\mu_X + b\mu_Y$$

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

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

**Example 3.14.** **Expected money for rolling 3 dice**

Let the random variables  $X_1, X_2, X_3$  be the values shown on 3 fair 6-sided dice rolls.

Suppose you are given in dollars the amount of the first roll, plus twice the value of the second roll, plus 4 times the value of the third roll.

How much money do you expect to get?

M = money we get from 3 rolls

$$M = 1 \cdot X_1 + 2X_2 + 4X_3$$

$$\mathbb{E}[M] = \mathbb{E}[X_1 + 2X_2 + 4X_3]$$

↓ linearity property

$$= \mathbb{E}[X_1] + 2\mathbb{E}[X_2] + 4\mathbb{E}[X_3]$$

$$= 3.5 + 2(3.5) + 4(3.5)$$

$$= 7(3.5)$$

$$= \$24.50$$

From before:

$$\mathbb{E}[X_i] = 3.5$$

Make sure to correct typo in textbook!!!

### 3.1. RANDOM VARIABLES

145

#### 3.1.4 Linear combinations of random variables

Sums of random variables arise naturally in many problems. In the health insurance example, the amount spent by the employee during her next five years of employment can be represented as  $X_1 + X_2 + X_3 + X_4 + X_5$ , where  $X_1$  is the cost of the first year,  $X_2$  the second year, etc. If the employee's domestic partner has health insurance with another employer, the total annual cost to the couple would be the sum of the costs for the employee ( $X$ ) and for her partner ( $Y$ ), or  $X + Y$ . In each of these examples, it is intuitively clear that the average cost would be the sum of the average of each term.

Sums of random variables represent a special case of linear combinations of variables.

#### LINEAR COMBINATIONS OF RANDOM VARIABLES AND THEIR EXPECTED VALUES

If  $X$  and  $Y$  are random variables, then a linear combination of the random variables is given by

$$aX + bY,$$

where  $a$  and  $b$  are constants. The mean of a linear combination of random variables is

$$E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y.$$

The formula easily generalizes to a sum of any number of random variables. For example, the average health care cost for 5 years, given that the cost for services remains the same, is

$$= E[X_1] + E[X_2] + \dots + E[X_5] \quad \text{all have same expected value}$$

$E(X_1 + X_2 + X_3 + X_4 + X_5) = E(5X_1) = 5E(X_1) = (5)(1010) = \$5,050.$

The formula implies that for a random variable  $Z$ ,  $E(a + Z) = a + E(Z)$ . This could have been used when calculating the average health costs for the employee by defining  $a$  as the fixed cost of the premium ( $a = \$948$ ) and  $Z$  as the cost of the physician visits. Thus, the total annual cost for a year could be calculated as:  $E(a + Z) = a + E(Z) = \$948 + E(Z) = \$948 + .30(1 \times \$20) + .40(3 \times \$20) + .20(4 \times \$20) + 0.10(8 \times \$20) = \$1,010.00$ .

**Theorem 3.15. Variance of a linear combination of random variables.**

If  $X$  and  $Y$  are INDEPENDENT random variables and  $a$  and  $b$  are constants, then

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

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

**Example 3.16. Variance of money for rolling 3 dice**

Let the random variables  $X_1, X_2, X_3$  be the values shown on 3 fair 6-sided dice rolls. Suppose you are given in dollars the amount of the first roll, plus twice the value of the second roll, plus 4 times the value of the third roll.

What are the variance and standard deviation of the amount you get from the 3 rolls?

$M = \text{total amount from 3 rolls}$

$$M = X_1 + 2X_2 + 4X_3$$

$$\begin{aligned}\sigma^2 &= \text{Var}(M) = \text{Var}(X_1 + 2X_2 + 4X_3) \\ &= \text{Var}(X_1) + 2^2 \text{Var}(X_2) + 4^2 \text{Var}(X_3)\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} = \sqrt{61.32} \\ &= \$7.83\end{aligned}$$

## Class discussion

### Example 3.17. Vaccinated people testing positive for Covid-19 (revisited)

About 25% of people that test positive for Covid-19 are vaccinated for Covid-19.

Define the r.v.  $X_i$  to be 1 if someone that tests positive is vaccinated and 0 if they are not vaccinated.

Suppose 3 people have tested positive for Covid-19 (independently of each other).

Let  $T$  denote the number of people that are vaccinated amongst the 3 that tested positive.

(1) Using the r.v.'s  $X_i$ , write a mathematical equation for calculating  $T$ .

(2) What is the expected value of  $T$ ?

(3) What is the variance of  $T$ ?

(4) What is the probability distribution of  $T$ ?

### 3.2. Binomial distribution.

- Many situations involve modeling independent random events that have 2 possible outcomes (binary), such as
  - Repeatedly flipping a coin
  - Whether a person that tested positive with Covid-19 is vaccinated or not
- Repeated events are referred to as trials  $X=1$   $X=0$
- The 2 possible outcomes are referred to as successes and failures.
- We denote the probability of a success as  $p$ .
- We denote the probability of a failure as  $q = 1 - p$ .

#### 3.2.1. Bernoulli distribution.

**Definition 3.18.** Bernoulli random variable.

If  $X$  is a random variable that takes value 1 with probability of success  $p$  and 0 with probability  $1 - p$ , then  $X$  is a Bernoulli random variable.

$$\begin{array}{c|cc|c} q & x & 0 & 1 \\ \hline P(X=x) & 1-p=q & & p \end{array} \quad P(X=x) = p^x (1-p)^{1-x} \text{ for } x=0,1$$

- We call the probability of success  $p$  the parameter of the Bernoulli distribution.
- Each value of  $p$  identifies a specific Bernoulli distribution out of the family of Bernoulli r.v.'s where  $p$  is any value between 0 and 1 (inclusive).
- If a r.v.  $X$  is modeled by a Bernoulli distribution, then we write in shorthand

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

**Theorem 3.19.** Mean and SD of a Bernoulli r.v.

If  $X$  is a Bernoulli r.v. with probability of success  $p$ , then

$$\mu = \mathbb{E}[X] = p$$

$$\sigma^2 = \text{Var}(X) = p(1-p)$$

$$\sigma = \text{SD}(X) = \sqrt{p(1-p)}$$

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=1}^n x_i P(X=x_i) \\ &= 1(p) + 0(1-p) \\ &= p \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \sum_{i=1}^n (x_i - \mu)^2 P(X=x_i) = (1-p)^2 p + (0-p)^2 (1-p) \\ &= (1-p)^2 p + p^2 (1-p) = (1-p)p(1-p+p) = (1-p)p \end{aligned}$$

### 3.2.2. Binomial distribution.

Recall Example 3.17

- About 25% of people that test positive for Covid-19 are vaccinated for Covid-19.
- Define the r.v.  $X_i$  to be 1 if someone that tests positive is vaccinated and 0 if they are not vaccinated.  $X_1, X_2, X_3 \leftarrow$  each  $X_i \sim \text{Bern}(p=0.25)$
- Suppose 3 people have tested positive for Covid-19 (independently of each other).
- Let  $T$  denote the number of people that are vaccinated amongst the 3 that tested positive.

The random variable  $T$  above is an example of a Binomial random variable.

In general, a random variable  $X$  is **Binomial** if the following hold:

- (1) The trials are independent.
- (2) The number of trials,  $n$ , is fixed.
- (3) Each trial outcome can be classified as a *success or failure*.
- (4) The probability of a success,  $p$ , is the same for each trial.
- (5) The r.v.  $X$  is the total number of successes in the  $n$  trials.

**Definition 3.20. Distribution of a Binomial random variable.**

Let  $X$  be the total number of successes in  $n$  independent trials, each with probability  $p$  of a success.

Then probability of observing exactly  $k$  successes in  $n$  independent trials is

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k=0, 1, 2, \dots, n$$

"n choose k"

Use R for calculations!

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \text{ where } n! = n \cdot (n-1) \cdot (n-2) \cdots \cdot 2 \cdot 1$$

- The parameters of a binomial distribution are  $p$  and  $n$ .  $X \sim \text{Bin}(n, p)$
- If a r.v.  $X$  is modeled by a binomial distribution, then we write in shorthand

#ways to choose which  $k$  of the  $n$  trials results in a success.

$$\frac{\overbrace{X_1 + X_2 + X_3 + \dots + X_n}^{n \text{ trials}}}{n} = \frac{1}{n} + \frac{0}{(n-1)} + \frac{1}{(n-2)} + \dots + \frac{1}{(n-(k+1))}$$

$\frac{n!}{(n-k)!k!} = \binom{n}{k}$  ← total # ways to arrange  $n$  trials

**Theorem 3.21.** Mean and SD of a Binomial r.v. ←  
If  $X$  is a binomial r.v. with probability of success  $p$ , then n trials

$$\mu = \mathbb{E}[X] = np$$

$$\sigma^2 = \text{Var}(X) = np(1-p)$$

$$\sigma = \text{SD}(X) = \sqrt{np(1-p)}$$

Why?

$\bar{X}$  = total # of successes in  $n$  indep. trials

$$= \sum_{i=1}^n X_i \quad \text{where } X_i \sim \text{Bern}(p) \text{ (independent)}$$

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p = np$$

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n p(1-p) = np(1-p)$$

### Example 3.22. Vaccinated people testing positive for Covid-19 (revisited)

About 25% of people that test positive for Covid-19 are vaccinated for Covid-19.

Suppose 10 people have tested positive for Covid-19 (independently of each other).

Let  $X$  denote the number of people that are vaccinated amongst the 10 that tested positive.

$$X \sim \text{Bin}(n=10, p=0.25)$$

*Class discussion*

(1) What is the expected value of  $X$ ?

(2) What is the SD of  $X$ ?

(3) What is the probability that exactly 4 of the 10 people that tested positive are vaccinated?

$$X \sim \text{Bin}(n=10, p=0.25)$$

$$P(X=4) = \binom{10}{4} (0.25)^4 (0.75)^{10-4} = \frac{10!}{4! 6!} (0.25)^4 (0.75)^6$$

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

for  $k=0, 1, 2, \dots, n$

R calculation:

$$P(X=4) = \text{dbinom}(x=4, \underbrace{\text{size}=10}_{n}, \underbrace{\text{prob}=0.25}_p)$$

*distribution*

$$P(X=k) = \text{dbinom}(x=k, \text{size}=n, \text{prob}=p)$$

$$X \sim \text{Bin}(n=10, p=0.25)$$

15

(4) What is the probability that at most 3 of the 10 people that tested positive are vaccinated?

$$\mathbb{P}(X \leq 3) = \mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) + \mathbb{P}(X=3)$$

$$= \sum_{k=0}^3 \mathbb{P}(X=k)$$

$$= \sum_{k=0}^3 \binom{10}{k} \cdot 0.25^k \cdot 0.75^{10-k}$$

$$= \binom{10}{0} \cdot 0.25^0 \cdot 0.75^{10} + \binom{10}{1} \cdot 0.25^1 \cdot 0.75^9 + \binom{10}{2} \cdot 0.25^2 \cdot 0.75^8 + \binom{10}{3} \cdot 0.25^3 \cdot 0.75^7$$



R calculation

$$\mathbb{P}(X \leq 3) = \text{pbinom}(q=3, size=10, prob=0.25) = 0.7759$$

*, lower.tail = TRUE*

(5) What is the probability that at least 5 of the 10 people that tested positive are more than 5  $\rightarrow \mathbb{P}(X > 5)$

$$\mathbb{P}(X \geq 5) =$$

0 1 2 3 4 5 6 7 8 9 10

$$\mathbb{P}(X=5) + \mathbb{P}(X=6) + \mathbb{P}(X=7) + \mathbb{P}(X=8) + \mathbb{P}(X=9) + \mathbb{P}(X=10)$$

$$= \sum_{k=5}^{10} \mathbb{P}(X=k) = \sum_{k=5}^{10} \binom{10}{k} \cdot 0.25^k \cdot 0.75^{10-k}$$

$$= \binom{10}{5} \cdot 0.25^5 \cdot 0.75^{10-5} + \binom{10}{6} \cdot 0.25^6 \cdot 0.75^{10-6} + \dots + \binom{10}{10} \cdot 0.25^{10} \cdot 0.75^{10-10}$$

R calculation

*pb*inom

$$\text{Option 1} \quad \mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4)$$

$$= 1 - \text{pbinom}(q=4, size=10, prob=.25, \text{lower.tail}=TRUE)$$

default

$$= 0.0781$$

TRUE & FALSE  
are case sensitive  
 $\rightarrow$  use ALL CAPS

$$\text{Option 2} \quad \mathbb{P}(X \geq 5) = \mathbb{P}(X > 4)$$

$$= \text{pbinom}(q=4, size=10, prob=.25, \text{lower.tail}=FALSE)$$

$$= 0.0781$$