HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 10 — Random variables and probability distributions

Jorge N. Tendeiro

Hiroshima University

Which lottery to pick?

Which lottery gives you the best chance of winning a prize?

Lottery 1

	Price A	Price B	Price C
Price (yen)	1000	500	100
Number of prizes	1	3	6

Lottery 2

	Price A	Price B	Price C
Price (yen)	1000	500	0
Number of prizes	2	2	6

Lottery 3

	Price A	Price B	Price C
Price (yen)	1000	400	100
Number of prizes	2	1	7

Today

- Random variables
- Probability distributions
- Expected value and variance of a random variable

Random variables and distributions

Trial, sample space (review)

Trial:

> *Any type of experiment involving uncertainty.*

Example:
Tossing a die once.

Sample space (Ω):

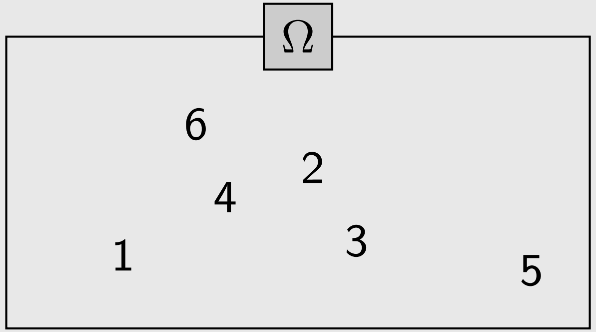
> *Set of all possible outcomes of a trial.*

Example:
The sample space of the trial consisting of tossing a die once is given by

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

or, using words,

$$\Omega = \{\text{getting } 1, \text{getting } 2, \cdots, \text{getting } 6\}$$



Event (review)

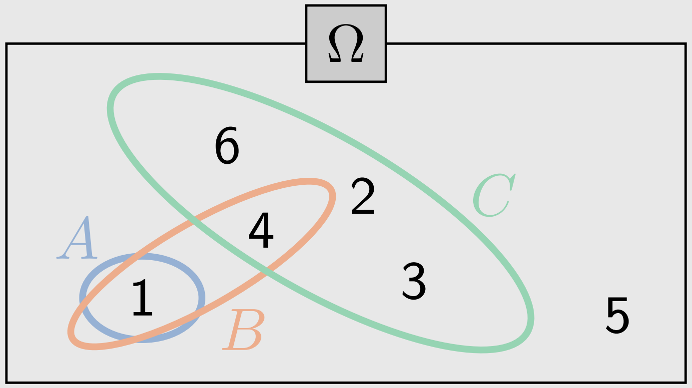
Event:

> *Any subset of the sample space* Ω.

Example:
Consider again the trial of tossing a die once.

Here are various possible events:

- Getting 1: $A = \{1\}$ or $\{\text{getting } 1\}$
- Getting 1 or 4: $B = \{1, 4\}$ or $\{\text{getting } 1, \text{getting } 4\}$
- Not getting 1 or 5: $C = \{2, 3, 4, 6\}$ or $\{\text{getting } 2, \text{getting } 3, \text{getting } 4, \text{getting } 6\}$



Random variables

It is easier to deal with sample spaces and events written numerically than in words.

A random variable is a function that allows converting any type of event into numbers.

Here are two simple examples.

Tossing a die:

$$X = \begin{cases} 1, & \text{if face 1 lands up} \\ 2, & \text{if face 2 lands up} \\ 3, & \text{if face 3 lands up} \\ 4, & \text{if face 4 lands up} \\ 5, & \text{if face 5 lands up} \\ 6, & \text{if face 6 lands up} \end{cases}$$

Winning/losing a game:

$$Y = \begin{cases} 1, & \text{in case of winning} \\ -1, & \text{in case of losing} \end{cases}$$

Here, 1 and -1 serve to distinguish between the two possible outcomes (*winning* or *losing*).

Choosing the values of a random variable wisely

Typically, we choose the values of a random variable that better reflect our research question.

Example.

Consider the trial consisting of throwing two coins at the same time.
In this case, the sample space is

$$\Omega = \{(\text{heads}, \text{heads}), (\text{heads}, \text{tails}), (\text{tails}, \text{heads}), (\text{tails}, \text{tails})\}\,.$$

From the following three random variables X, Y, and Z,

Random variable	(heads, heads)	(heads, tails)	(tails, heads)	(tails, tails)
X	4	3	2	1
Y	1	0	0	-1
Z	2	1	1	0

which random variable is the *most appropriate* in case you want to study how many coins landed heads up?

Clearly, the best random variable in this case is Z.

Random variables and probabilities

Given a random variable, we may define probabilities for each possible event.

Example.

Consider the trial consisting of throwing two coins at the same time.
Here is random variable Z = *the total number of heads*:

Random variable	(heads, heads)	(heads, tails)	(tails, heads)	(tails, tails)
Z	2	1	1	0

We can define the probability of Z being equal to 0, 1, or 2.

Random variables and probabilities

$$\Omega = \{(\text{heads}, \text{heads}), (\text{heads}, \text{tails}), (\text{tails}, \text{heads}), (\text{tails}, \text{tails})\}$$

For example, if both coins are fair then (*notation:* $\#(x)$ = number of elements in set x):

- $P(Z=0) = P(0 \text{ heads}) = \frac{\#(\{(\text{tails},\text{tails})\})}{\#(\Omega)} = \frac{1}{4}$
- $P(Z=1) = P(1 \text{ heads}) = \frac{\#(\{(\text{heads},\text{tails}),(\text{tails},\text{heads})\})}{\#(\Omega)} = \frac{2}{4} = \frac{1}{2}$
- $P(Z=2) = P(2 \text{ heads}) = \frac{\#(\{(\text{heads},\text{heads})\})}{\#(\Omega)} = \frac{1}{4}$

Therefore, we have the so-called probability distribution associated to random variable Z:

Z	0	1	2	TOTAL
Probability	$1/4$	$1/2$	$1/4$	1

Probability distributions

The idea of associating probabilities to events of a random variable is quite common.
In general, given a random variable X with possible values

$$x_1, x_2, \ldots, x_n,$$

we may define a probability distribution by choosing values

$$p_1, p_2, \ldots, p_n$$

such that

$$P(X = x_k) = p_k, \text{ for } k = 1, \ldots, n.$$

X	x_1	x_2	$\cdots$	x_n	TOTAL
$P(X = x_k)$	p_1	p_2	$\cdots$	p_n	1

The values p_k are probabilities (i.e., non-negative real numbers that sum to 1).

The set of all probabilities of a random variable define the corresponding probability distribution.
Conversely, a probability distribution determines the set of all probabilities of a random variable.

Probability distributions

X	x_1	x_2	$\cdots$	x_n	TOTAL
$P(X = x_k)$	p_1	p_2	$\cdots$	p_n	1

Probability distribution are **extremely important** to study uncertain phenomena!
E.g.: Weather prediction, winner of an election, chance of malfunction, etc.

By knowing a probability distribution, we can understand how events occur, probabilistically.

Exercise (1)

Consider the following lottery:

Lottery 1

	Price A	**Price B**	**Price C**
Price (yen)	1000	500	100
Number of prizes	1	3	6

This lottery includes 10 (=1+3+6) lots in total.
Each lot offers either prize A, prize B, or prize C.
Suppose you randomly pick one lot.

Denoting $X =$ *prize money in yen*, what is the corresponding probability distribution?
Fill in the table below.

X	1000	500	100	TOTAL
Probability				1

Exercise (1) — ANSWER

Consider the following lottery:

Lottery 1

	Price A	**Price B**	**Price C**
Price (yen)	1000	500	100
Number of prizes	1	3	6

This lottery includes 10 (=1+3+6) lots in total.
Each lot offers either prize A, prize B, or prize C.
Suppose you randomly pick one lot.

Denoting $X =$ *prize money in yen*, what is the corresponding probability distribution?
Fill in the table below.

X	1000	500	100	TOTAL
Probability	1/10	3/10	6/10	1

Discrete versus *continuous* random variables

Classification of random variables

Random variables can be broadly classified as being either discrete or continuous.

Discrete random variables:

> *Random variables with values that are discrete.*
> *This means that it is **not** always possible to conceive of an intermediate value between any two values.*

Continuous random variables:

> *Random variables with values that are continuous.*
> *This means that it is always possible to conceive of an intermediate value between any two values.*

Defining and computing probabilities works differently for discrete and continuous random variables.

Examples of discrete random variables

- $W =$ result of tossing a die.

  Then $W = 1, 2, 3, 4, 5, 6.$
  W is a discrete random variable since its values are discrete (finite in this case).
  Also, we cannot conceive of an intermediate value between "1" and "2", for example.

- $X = 1$ if winning a game, -1 if losing a game.

  Then $X = -1, 1.$
  Can you explain why X is a discrete random variable?

- $Y =$ prize money from a lottery (1000 for Prize A, 500 for Prize B, 100 for Prize C).

  Then $Y = 100, 500, 1000.$
  Can you explain why Y is a discrete random variable?

- $Z =$ number of traffic accidents in a year.

  Then $Z = 0, 1, 2, 3, \ldots$
  Can you explain why Z is a discrete random variable?

Discrete distributions

A discrete probability distribution, or simply a discrete distribution, is the probability distribution followed by a discrete random variable.

If random variable X can assume values $x_1, x_2, \ldots$, then a discrete distribution for X is given by a set of probabilities $p_1, p_2, \ldots$, such that

$$P(X = x_k) = p_k, \text{ for } k = 1, 2, \ldots$$

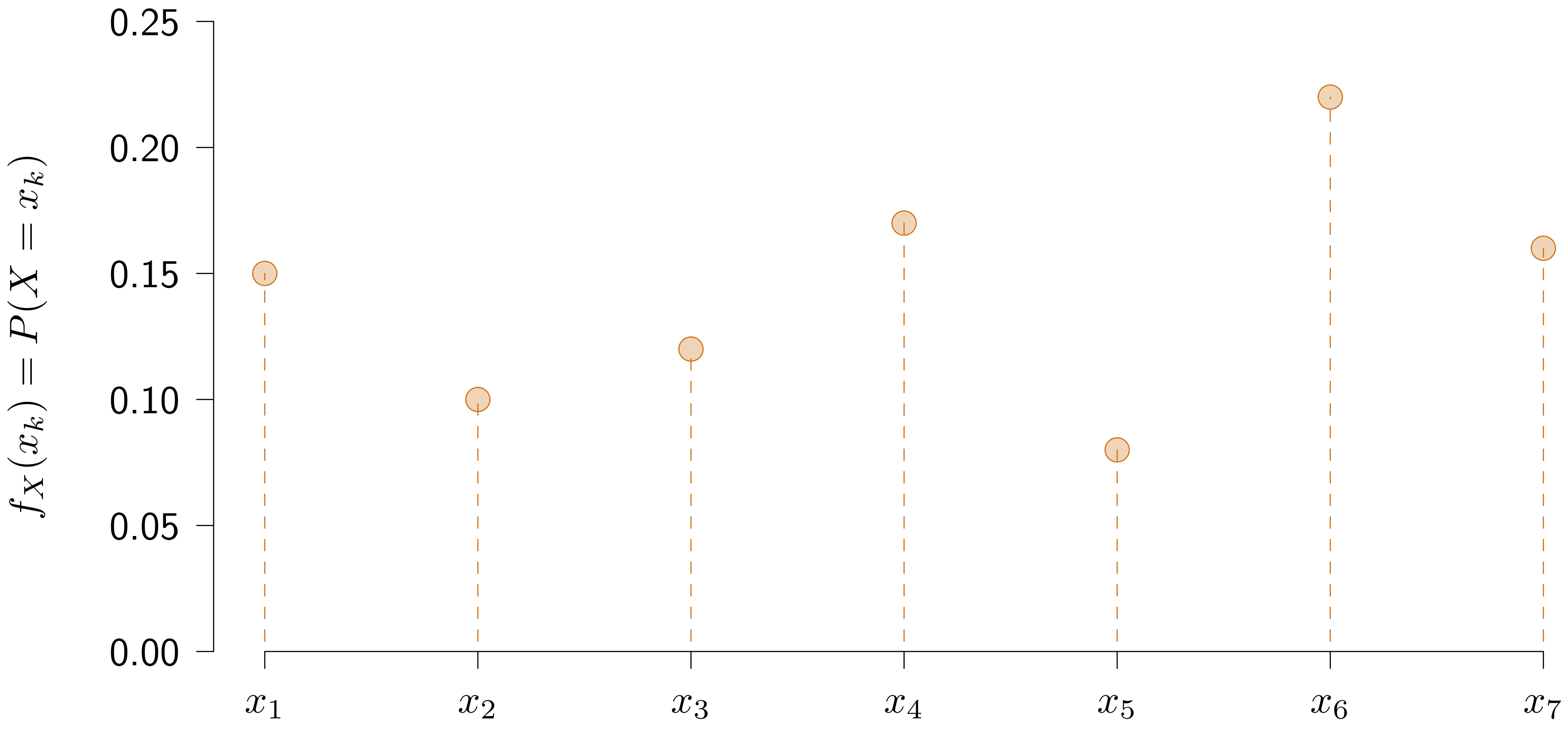
Note that $0 \leq p_k \leq 1$ for $k = 1, 2, \ldots$, and $\sum_k p_k = 1$.

X	x_1	x_2	$\cdots$	TOTAL
$P(X = x_k)$	p_1	p_2	$\cdots$	1

$P(X = x_k)$ is called a probability mass function of X.
It is often denoted as $f_X(x_k)$ or simply f_X.

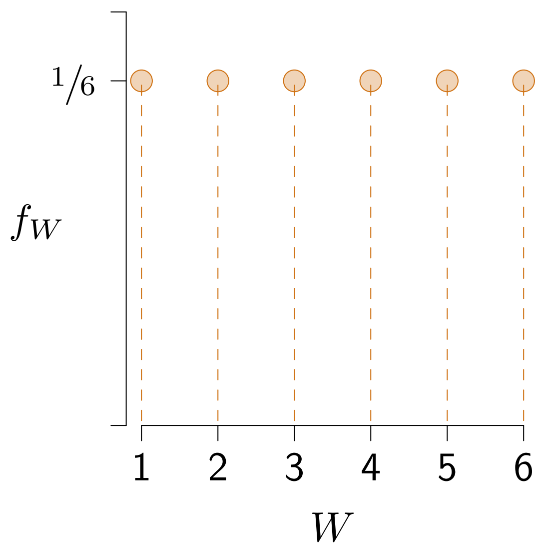
$$f_X(x_k) = \begin{cases} p_1, & \text{if } x = x_1 \\ p_2, & \text{if } x = x_2 \\ \vdots & \end{cases}$$

Probability mass function — Example



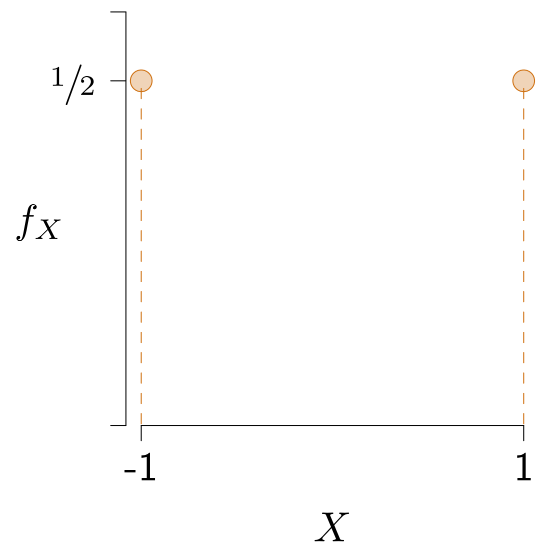
Probability mass function — Examples

W = result of tossing a fair die.

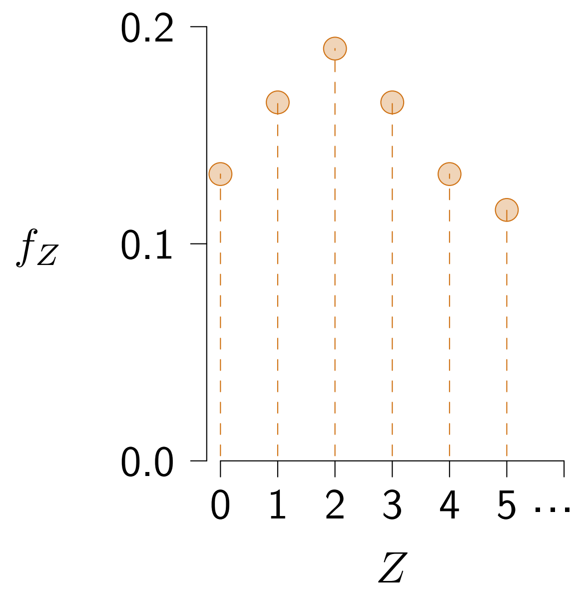


$X = 1$ if winning a game, -1 if losing a game.

Assume that *winning* is as likely as *losing*.



Z = number of traffic accidents in a year.



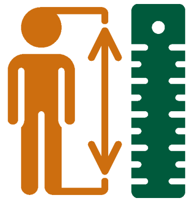
Examples of continuous random variables

- $Z =$ amount of sugar contained in cola bought at a vending machine

  Note that $Z \geq 0$.

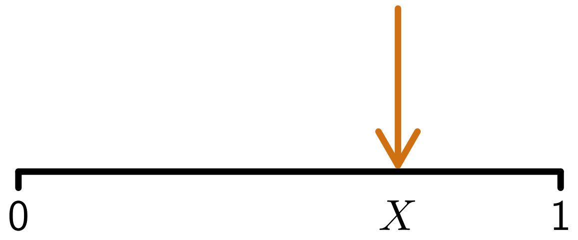
- $Y =$ Height in cm.

  Note that $Y \geq 0$.

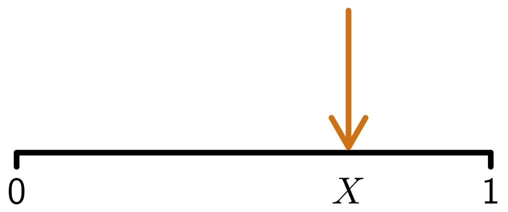


- $X =$ Random number between 0 and 1.

  Note that $0 \leq X \leq 1$.



Probability of a continuous random variable



Q: What is the probability of a randomly generated number between 0 and 1 being equal to, say, 0.5?
A: Strangely enough, the probability is... 0!

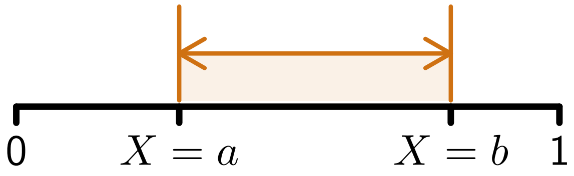
The problem is that, between 0 and 1, there is an infinite, *uncountable*, amount of numbers.
It is **impossible** that each such number has a non-zero probability and still expect that the sum of all probabilities is equal to 1.

Hence, for continuous random variables, any specific value has probability equal to 0:

$$P(X = x) = 0, \text{for any value } x.$$

Probability of a continuous random variable

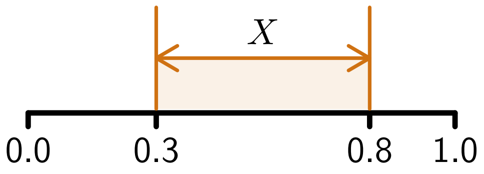
Instead of thinking about the probability at a value, we consider the probability of a range of values.



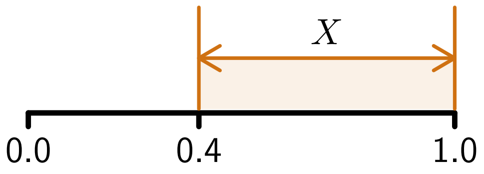
$$P(a \le X \le b) = \frac{[a,b] \text{ interval width}}{[0,1] \text{ interval width}} = \frac{b-a}{1-0} = b - a.$$

So, for example:

- $P(0.3 \le X \le 0.8) = 0.8 - 0.3 = 0.5$



- $P(X \ge 0.4) = 1 - 0.4 = 0.6$

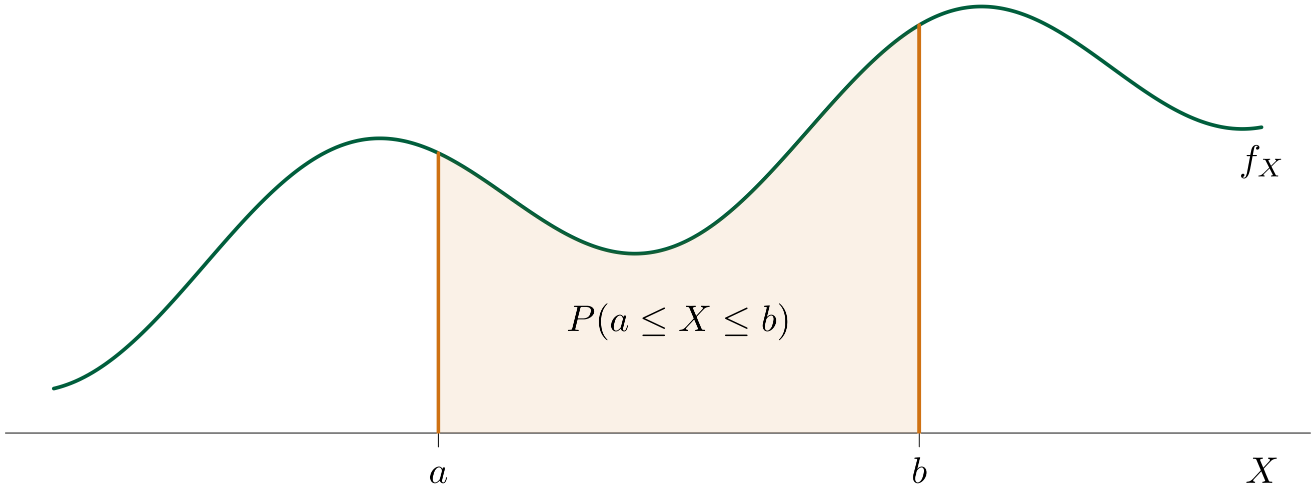


Probability of a continuous random variable

In general, for any continuous random variable X,

$$P(a \leq X \leq b) = \text{area under function } f_X.$$

Function f_X is known as the probability density function of continuous random variable X.



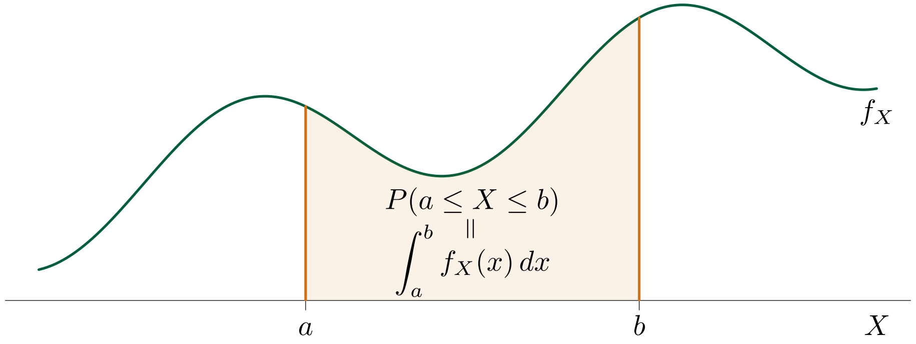
Probability of a continuous random variable

Technically, the probability of a continuous random variable X between a and b is given by

$$P(a \leq X \leq b) = \text{area under function } f_X = \int_a^b f_X(x)\,dx,$$

where f_X is the probability density function of X and $\int_a^b$ denotes the integral of f_X in the interval (a, b).

(In simple terms: 'integral' = 'area'.)



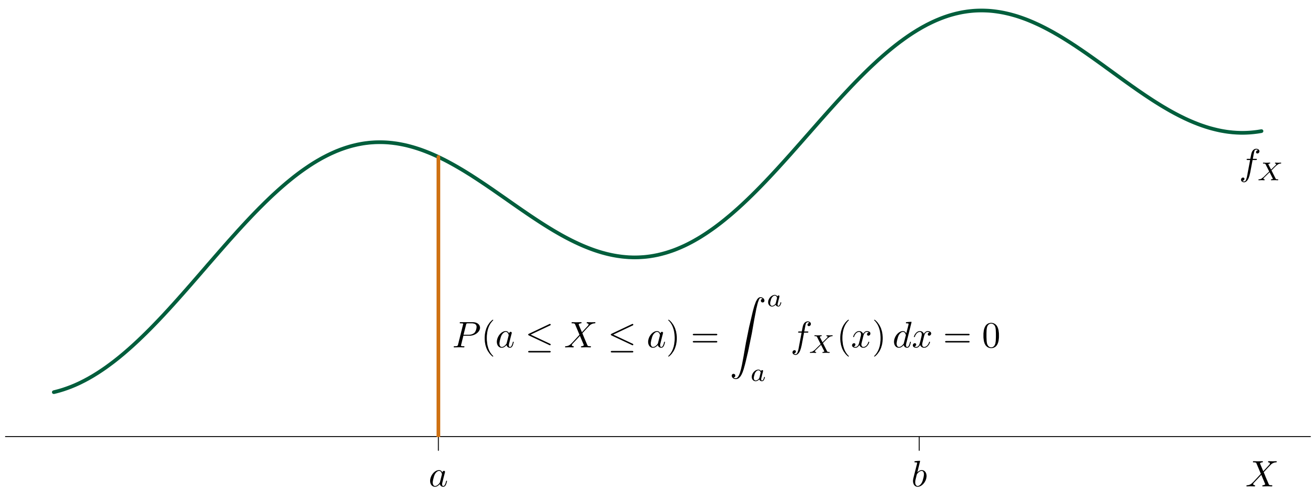
The probabilities $P(a \leq X \leq b)$ determine the probability density function f_X.
Conversely, a probability density function f_X determines the set of all probabilities $P(a \leq X \leq b)$.

Probability of a continuous random variable

Note that

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)\,dx = 0,$$

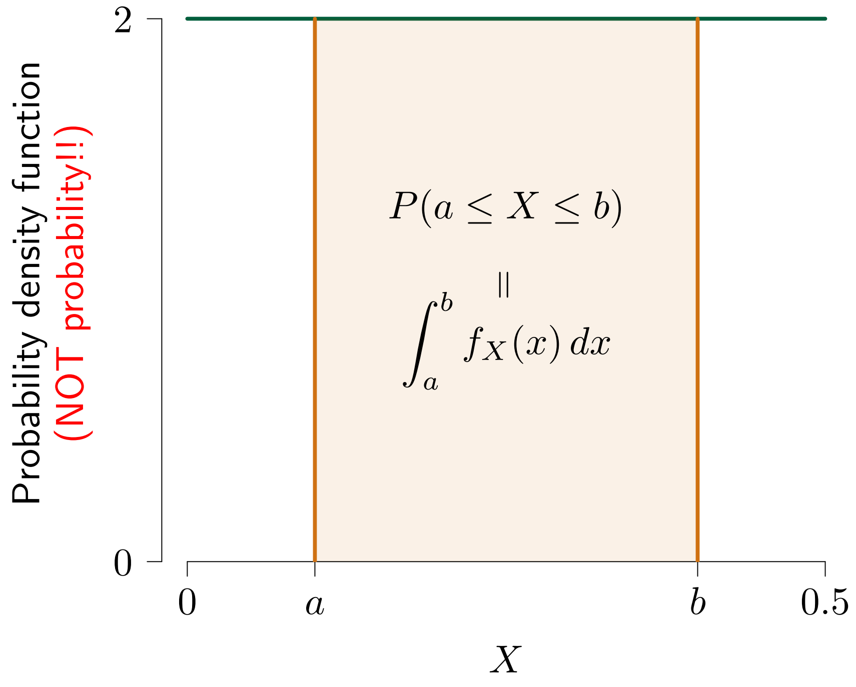
since the area of a line is equal to 0.



Probability density function — Example

Consider the following probability density function:

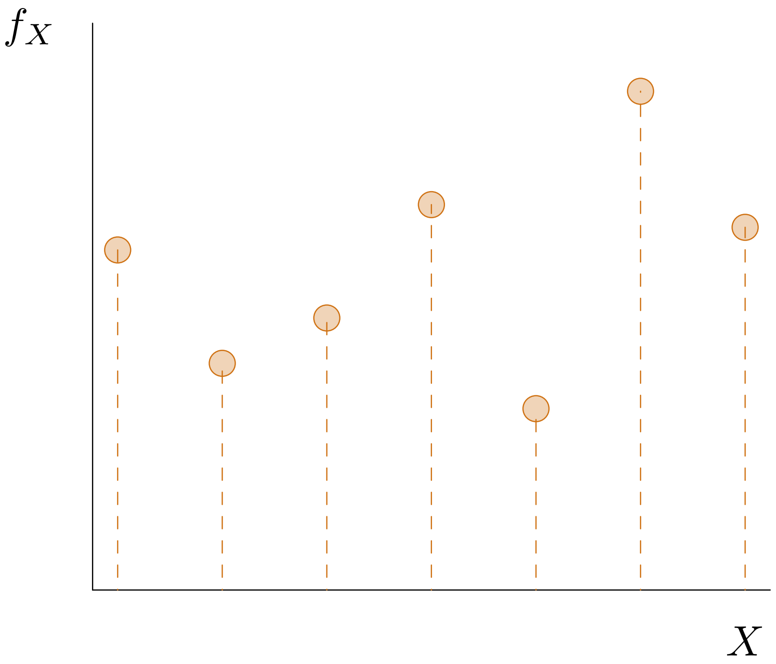
$$f_X(x) = 2, \text{ for } 0 \leq x \leq 0.5$$

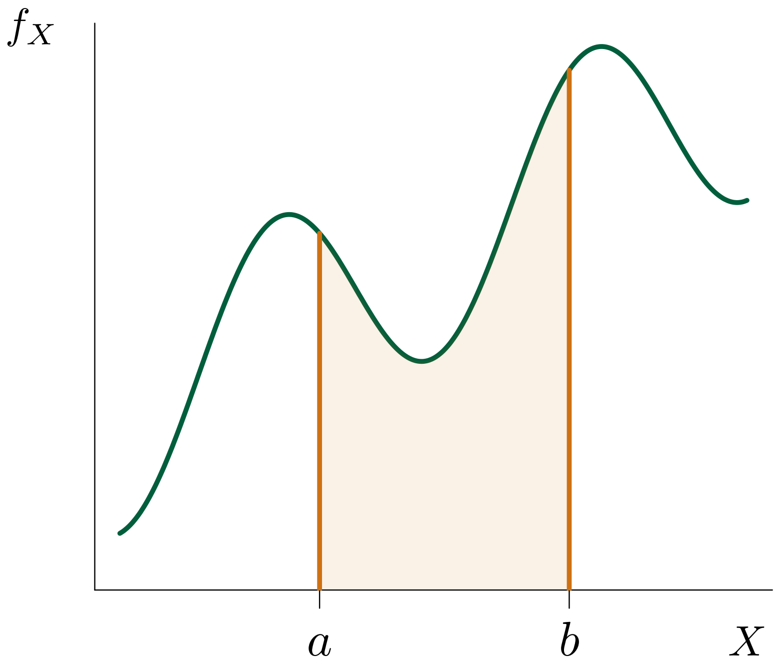


The probability of X between a and b is

$$P(a \leq X \leq b) = \int_a^b f_X \, dx = 2(b - a).$$

Discrete and continuous random variables — Summary





- Probability mass function $f_X(x_k)$
- $P(X = x_k) = f_X(x_k)$

- Probability density function $f_X(x)$
- $P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$

Expected value and variance of a discrete random variable

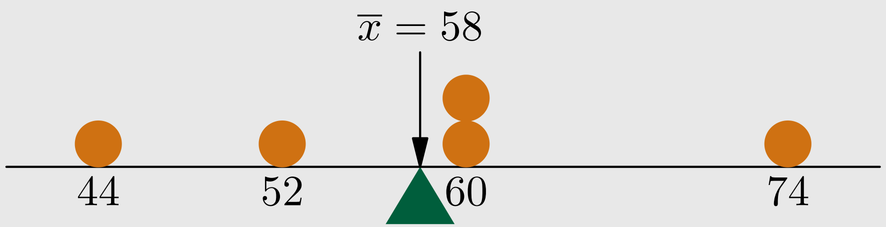
Descriptive statistics: Mean (*location*), variance (*spread*)

Individual	1	2	3	4	5
Weight (kg)	60	52	44	74	60

Mean (AKA *average, expected value*):

$$\overline{x} = \frac{60 + 52 + 44 + 74 + 60}{5} = 58$$

The mean is the data's center of mass:



Variance:

$$s_x^2 = \frac{(60-58)^2 + (52-58)^2 + \cdots + (60-58)^2}{5}$$
$$= \frac{2^2 + 6^2 + \cdots + 2^2}{5} = 99.2$$

The variance describes the spread of the data around the mean.

> *We can also define the mean (average, expected value) and the variance of random variables.*

Discrete random variable — Expected value

Definition

> *The expected value of discrete random variable X with probability mass function $f_X(x_k) = p_k$ for $k = 1, 2, \ldots$ is given by*
>
> $$\mathbb{E}[X] = \sum_k x_k p_k = x_1 p_1 + x_2 p_2 + \cdots$$

X	$P(X = x_k)$	Product
x_1	p_1	$x_1 p_1$
x_2	p_2	$x_2 p_2$
$\vdots$	$\vdots$	$\vdots$
		Sum $= \mathbb{E}[X]$

Expected value of discrete random variable — Example

Consider the following lottery:

Lottery 1

	Price A	**Price B**	**Price C**
Price (yen)	1000	500	100
Number of prizes	1	3	6

Suppose you could *repeatedly* and *randomly* pick a lot.
What is the expected prize money?

Expected value of discrete random variable — Example

Lottery 1

	Price A	Price B	Price C
Price (yen)	1000	500	100
Number of prizes	1	3	6

Answer:

Denoting $X =$ *prize money in yen*, we need to compute the expected value of X:

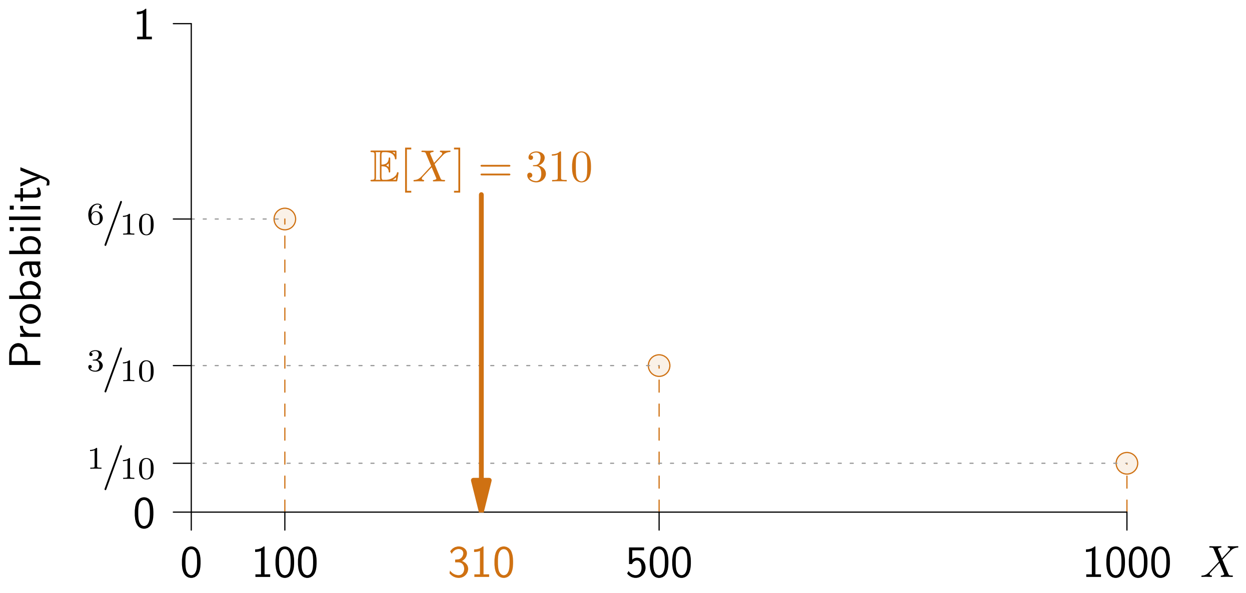
$$\mathbb{E}[X] = \sum_k x_k p_k = 1000 \times \frac{1}{10} + 500 \times \frac{3}{10} + 100 \times \frac{6}{10} = 310.$$

X	1000	500	100	
Probability	$1/10$	$3/10$	$6/10$	
Product	100	150	60	$\mathbb{E}[X] = 310$

Thus, in the long run, you expect to earn 310 yen per lot.

Expected value of discrete random variable — Example

X	1000	500	100	TOTAL
Probability	$1/10$	$3/10$	$6/10$	1



The expected value of X, $\mathbb{E}[X]$, represents the center of gravity of the probability distribution.

Expected value of discrete random variable — Example

Lottery 1

	Price A	**Price B**	**Price C**
Price (yen)	1000	500	100
Number of prizes	1	3	6

Denoting X = *prize money in yen*, we learned that $\mathbb{E}[X] = 310$ yen.

> *Suppose that the price of one lot is 400 yen.*
> *In the long run, is it profitable for me to buy one lot?*

Answer:

Since

$$\mathbb{E}[X] = 310 < 400 = \text{price of one lot}$$

buying one lot does **not** seem to pay off.
In the long run, we expect to **lose** 90 yen on average per lot.

Discrete random variable — Variance

Definition

> *The variance of discrete random variable X with probability mass function $f_X(x_k) = p_k$ for $k = 1, 2, \ldots$ is given by*
>
> $$\begin{aligned} V[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_k (x_k - \mathbb{E}[X])^2 p_k \\ &= (x_1 - \mathbb{E}[X])^2 p_1 + (x_2 - \mathbb{E}[X])^2 p_2 + \cdots \end{aligned}$$

X	$P(X = x_k)$		
x_1	p_1	$x_1 p_1$	$(x_1 - \mathbb{E}[X])^2 p_1$
x_2	p_2	$x_2 p_2$	$(x_2 - \mathbb{E}[X])^2 p_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$\mathbb{E}[X]$	$V[X]$

Variance of discrete random variable — Example

Lottery 1	Price A	Price B	Price C
Price (yen)	1000	500	100
Number of prizes	1	3	6

Let X = *prize money in yen*.
We learned before that $\mathbb{E}[X] = 310$ yen.

The variance of X is then equal to

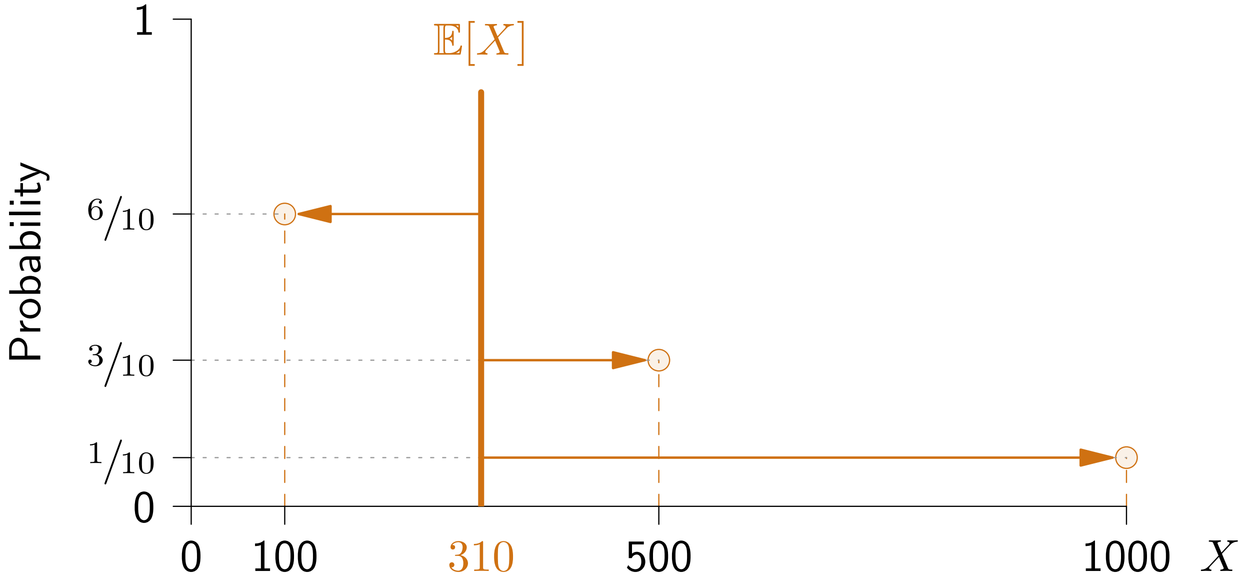
$$
\begin{aligned}
V[X] &= \sum_k (x_k - \mathbb{E}[X])^2 p_k \\
&= (1000 - 310)^2 \times \frac{1}{10} + (500 - 310)^2 \times \frac{3}{10} + (100 - 310)^2 \times \frac{6}{10} \\
&= 84900
\end{aligned}
$$

The units of a variance are squared.
Thus, we can say that the variance of X is equal to 84900 *squared yen*.

Variance of discrete random variable — Example

The variance describes the scatter of the values of a random variable arounds its expected value:

- The smaller $V[X]$, the closer the values of X are to $\mathbb{E}[X]$.
- The larger $V[X]$, the further away the values of X are to $\mathbb{E}[X]$.



Sample mean and variance versus mean and variance of a random variable

Sample mean, sample variance:

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

$$s_x^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

Mean and variance of random variable X:

$$\mathbb{E}[X] = \sum_k x_k p_k$$

$$V[X] = \sum_k (x_k - \mathbb{E}[X])^2 p_k$$

The sample mean and sample variance...

- ... represent characteristics of a set of observed data
(namely, its location and spread, respectively).
- ... **do** require observed data.

The expected value and variance of a random variable...

- ... represent characteristics of the variable's probability distribution
(namely, its location and spread, respectively).
- ... **do not** require observed data.
But, a probability distribution is required.

Which lottery to pick?

Which lottery gives you the best chance of winning a prize?

Lottery 1

	Price A	Price B	Price C
Price (yen)	1000	500	100
Number of prizes	1	3	6

Lottery 2

	Price A	Price B	Price C
Price (yen)	1000	500	0
Number of prizes	2	2	6

Lottery 3

	Price A	Price B	Price C
Price (yen)	1000	400	100
Number of prizes	2	1	7

> *Let's judge it by calculating the expected value and variance of the random variable 'prize money in yen' for the probability distribution corresponding to each lottery.*

Which lottery to pick?

Lottery 1

	Price A	Price B	Price C
Price (yen)	1000	500	100
Number of prizes	1	3	6

Lottery 2

	Price A	Price B	Price C
Price (yen)	1000	500	0
Number of prizes	2	2	6

Lottery 3

	Price A	Price B	Price C
Price (yen)	1000	400	100
Number of prizes	2	1	7

Lottery 1

X	1000	500	100	TOTAL
Probability	$1/10$	$3/10$	$6/10$	1

$$\mathbb{E}[X] = 310, \quad V[X] = 84900$$

Lottery 2

Y	1000	500	0	TOTAL
Probability	$2/10$	$2/10$	$6/10$	1

$$\mathbb{E}[Y] = ??, \quad V[Y] = ??$$

Lottery 3

Z	1000	400	100	TOTAL
Probability	$2/10$	$1/10$	$7/10$	1

$$\mathbb{E}[Z] = ??, \quad V[Z] = ??$$

Which lottery to pick? — Consider the expected values

$$\mathbb{E}[Y] = 1000 \times \frac{2}{10} + 500 \times \frac{2}{10} + 0 \times \frac{6}{10}$$
$$= 300$$

$$\mathbb{E}[Z] = 1000 \times \frac{2}{10} + 400 \times \frac{1}{10} + 100 \times \frac{7}{10}$$
$$= 310$$

Lottery 1

X	1000	500	100	TOTAL
Probability	1/10	3/10	6/10	1

$$\mathbb{E}[X] = 310, \quad V[X] = 84900$$

Lottery 2

Y	1000	500	0	TOTAL
Probability	2/10	2/10	6/10	1

$$\mathbb{E}[Y] = 300, \quad V[Y] = ??$$

Lottery 3

Z	1000	400	100	TOTAL
Probability	2/10	1/10	7/10	1

$$\mathbb{E}[Z] = 310, \quad V[Z] = ??$$

Which lottery to pick? — Consider the variances

$$V[Y] = (1000 - 300)^2 \times \frac{2}{10} + \cdots + (0 - 300)^2 \times \frac{6}{10}$$
$$= 106000$$

$$V[Z] = (1000 - 310)^2 \times \frac{2}{10} + \cdots + (100 - 310)^2 \times \frac{7}{10}$$
$$= 126900$$

Lottery 1

X	1000	500	100	TOTAL
Probability	1/10	3/10	6/10	1

$$\mathbb{E}[X] = 310, \quad V[X] = 84900$$

Lottery 2

Y	1000	500	0	TOTAL
Probability	2/10	2/10	6/10	1

$$\mathbb{E}[Y] = 300, \quad V[Y] = 106000$$

Lottery 3

Z	1000	400	100	TOTAL
Probability	2/10	1/10	7/10	1

$$\mathbb{E}[Z] = 310, \quad V[Z] = 126900$$

Which lottery to pick? — Conclusion

Lottery	$\mathbb{E}[X]$	$V[X]$
1	310	84900
2	300	106000
3	310	126900

- If you want to risk to win a large prize (largest $V[X]$):

> *Lottery 3*

- If you want to play safe and aim for a more stable prize amount (smallest $V[X]$):

> *Lottery 1*

- Between lotteries 1 and 3, lottery 3 is riskier.

Continuous random variable — Expected value, variance

Expected value:

> *The expected value (mean) of a continuous random variable X with probability density function $f_X(x)$ is*
>
> $$\mathbb{E}[X] = \int x f_X(x)\, dx.$$

Variance:

> *The variance of a continuous random variable X with probability density function $f_X(x)$ is*
>
> $$V[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int (x - \mathbb{E}[X])^2 f_X(x)\, dx.$$

These formulas are the 'integral-analogues' of the formulas for discrete random variables.
This is for your information only; we will not be using these formulas in this course.
But do keep in mind: The computation of expected values and variances differs between discrete and continuous random variables.

Summary

- Random variable:
  Function that allows converting any type of event into a number.

- *Discrete* versus *continuous* random variable:
  Determined by the type of values that the random variable can assume.

- Probability distribution:
  Correspondence between the values of a random variable and their probabilities.
  A probability distribution is a tool to understand uncertain phenomena.

- Expected value and variance of a random variable:
  Quantities representing characteristics of the random variable's probability distribution.
  This **is different from the sample mean and sample variance from observed data**.