

# Probability Theory for EOR

Introduction of the course motivation/organisation

Lingwei Kong

[l.kong@rug.nl](mailto:l.kong@rug.nl)

These slides adopt materials constructed by Dr. Tom Boot, who taught the course in previous years.

# Why?

# Uncertainty

- ▶ Econometrics = uncertainty
- ▶ We never know the exact outcome (economy one year from now?)
- ▶ Maybe we can determine the probability of different outcomes/events.

What does this mean: “*probability*” ?

## What we'll learn: 6 chapters

Book: *Introduction to Probability, Second Edition* by J. Hwang and J.K. Blitzstein.

- ▶ Chapter 1: A coherent way to assign probabilities to events.  
*This is bowling. There are rules.* (The Big Lebowski)
- ▶ Chapter 2: How probabilities change with new information.
- ▶ Chapter 3: From events to (discrete) numbers. Because we like numbers!
- ▶ Chapter 4: What numbers/outcomes do we expect?
- ▶ Chapter 5: From discrete to continuous outcomes.
- ▶ Chapter 6: Moment Generating Functions. (Surprisingly, MGF determines the distribution (under some regular conditions).)

## Organization of the course

# This course

- ▶ Lectures
  - Week 1-6: Chapter 1-6.
  - Week 7: overflow and/or repetition.
- ▶ On-campus Tutorials & Teaching Assistants:

---

Karlijn Zwiggelaar	k.a.h.zwiggelaar@student.rug.nl
Joost Doornbos	joostdoornbos01@gmail.com
Quinten Huisman	q.n.huisman@student.rug.nl
Sander van Beek	s.g.h.van.beek@student.rug.nl

---

- ▶ You!

# Highlights

Details in the course syllabus.

- ▶ Lectures
  - ▶ Part I: Recorded knowledge clips available every Monday at 9:00;
  - ▶ Part II: Live Wednesday at 10:00. Here we briefly review the materials and discuss the questions you have while watching the videos.
- ▶ Discussions:
  - ▶ On-campus tutorials;
  - ▶ Discussion board on Nestor.
- ▶ Note: The book contains all material relevant for the exam. Use it wisely!

# Tutorials

- ▶ On campus tutorials (discuss tutorial exercises once a week:  
Thursday or Friday based on your groups)
  - ▶ 45 minutes peer review (feedback)/discussions to produce your results (better to have done some works before the tutorials)
  - ▶ 45 minutes explanation of exercises.
- ▶ Group composition on Nestor.
- ▶ Earn you bonuses!

## Assessment

- ▶ Tutorial exercises (Bonuses)
- ▶ Assignments (20%, in pairs *within* tutorial groups, week 3/6).
- ▶ Closed-book on-campus exam (80%, individual). **One exercise, 3/10, you would have seen either from assignments, weekly tutorials or selected exercises.**
- ▶ Practice, practice, practice. We have other selected exercises. Understand instead of knowing by heart. Earn bonus by working hard and participating the tutorials.
- ▶ Keep up. **One chapter each week!**

Let's get to it!

# WHY USE PROGRAMMING?

- ▶ As a general rule: you don't understand something if you can't program it.
- ▶ Programming helps you check your answers.
- ▶ In exams, I see many answers that can never be true
  - ▶ probabilities outside of  $[0,1]$
  - ▶ instances when  $P(A \cap B) > P(A)$
  - ▶ squares that give negative answers
  - ▶ and many other exotics
- ▶ My guess is that the reason is that the material is abstract.

# WHY USE PROGRAMMING?

- ▶ A computer program makes things more real: it allows you to throw a die  $N = 1000$  times.
- ▶ Probability of seeing 2 eyes or less?
- ▶ Simply count how many of these 1000 throws show 2 eyes or less. Call this number  $N_1$ .
- ▶ Calculate  $N_1/N$ .
- ▶ Of course,  $N_1 \leq N$ , so  $N_1/N \leq 1$ .

# R

- ▶ All this coin flipping and dice throwing is a lot of work.
- ▶ It helps to have a computer.
- ▶ We will use R (Python is also quite similar), a freely available programming tool that is used a lot for statistical purposes.
- ▶ I recommend using RStudio from [rstudio.com](https://rstudio.com) (or [Google Colab](#) where you can also find interesting materials on Python coding). You may find useful information on R coding from this [website](#).
- ▶ Let's do a short tour on R and Python. Pick one you like. The textbook uses R though.

# List of useful identities and PDF/PMFs

1. Useful summation tricks ( $m \leq n$ )<sup>1</sup>:

- (a)  $\sum_{i=m}^n (f(i) + g(i)) = \sum_{i=m}^n f(i) + \sum_{i=m}^n g(i)$
- (b)  $\sum_{i=m}^n f(i) = \sum_{i=0}^n f(i) - \sum_{i=0}^{m-1} f(i)$
- (c)  $\sum_{i=m}^n cf(i) = c \sum_{i=m}^n f(i).$

2. Some other useful identities<sup>2</sup>:

- (a)  $\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2}$
- (b)  $\sum_{i=0}^{n-1} i^2 = \frac{n(n-1)(2n-1)}{6}$
- (c)  $\sum_{i=0}^{n-1} i^3 = \left(\frac{n(n-1)}{2}\right)^2$
- (d)  $\sum_{i=m}^n 1/(k(k+1)) = \sum_{i=m}^n (1/k - 1/(k+1)) = 1/m - 1/(n+1)$
- (e)  $\sum_{n=1}^{\infty} \frac{1}{n}$  diverges
- (f)  $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$  (Basel problem, main point is to realize that this is finite, while the sum on the previous line is not)
- (g)  $\sum_{x=0}^n z^x = \frac{1-z^{n+1}}{1-z}$  (finite geometric series)
- (h)  $\sum_{x=0}^{\infty} z^x = \frac{1}{1-z}$  when  $|z| < 1$  (infinite geometric series)
- (i)  $\sum_{x=1}^{\infty} xz^{x-1} = \frac{1}{(1-z)^2}$  when  $|z| < 1$  (easy to memorize as it is simply taking derivatives on both sides of  $\sum_{x=0}^{\infty} z^x = \frac{1}{1-z}$ ), which also implies  $\sum_{x=1}^{\infty} xz^x = \frac{z}{(1-z)^2}$  by multiplying  $x$  on both sides.

---

<sup>1</sup>1a and 1c simply show that the summation operator also satisfies the linearity. Think about the expectation operator we have discussed.

<sup>2</sup>1 and 2 are heavily used in Pois, Expo, MGF...

- (j)  $\sum_{x=1}^{\infty} x^2 z^{x-1} = \frac{1+z}{(1-z)^3}$  when  $|z| < 1$  (easy to memorize as it is simply taking derivatives on both sides of  $\sum_{x=1}^{\infty} x z^x = \frac{z}{(1-z)^2}$ ), which also implies  $\sum_{x=1}^{\infty} x^2 z^x = \frac{z(1+z)}{(1-z)^3}$  by multiplying  $x$  on both sides.

- (k) The Taylor expansion for  $e^x$ :

$$e^x = \lim_{l \rightarrow \infty} \sum_{n=0}^l x^n / n! = \sum_{n=0}^{\infty} x^n / n! = 1 + x + \frac{1}{2!} x^2 + \frac{1}{3!} x^3 + \dots \quad (1)$$

- (l) The definition of  $e^x$ :

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n \quad (2)$$

We use quite often the Euler's number or exponential function (Pois, Expo, Normal). Eq(2) is covered in Math I, while for Eq (1), it is also easy to memorize once you learn the PMF of Pois( $x$ ) is simply

$$p(n) = e^{-x} \frac{x^n}{n!}, n = 0, 1, \dots$$

and the sum of all PMFs should be one which implies

$$\sum_{n=0}^{\infty} e^{-x} \frac{x^n}{n!} = 1$$

and thus

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x.$$

### 3. Productions:

$$(a) \prod_{0 \leq i < j \leq m} a_i a_j = \prod_{i=0}^m \prod_{j=i+1}^m a_i a_j.$$

$$(b) \prod_{0 \leq i, j \leq m} a_i a_j = \prod_{i=0}^m \prod_{j=0}^m a_i a_j.$$

### 4. There are some conventions of when we can ignore brackets, we list some to avoid confusions:

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

$$(b) \mathbb{E}XY := \mathbb{E}(XY)$$

### 5. Some identities concerning the choice functions (we list more than sufficient here):

$$(a) \text{ The Bernoulli identity } (x+y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k$$

$$(b) \sum_{k=0}^n \binom{n}{k} = 2^n$$

$$(c) \binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$

$$(d) \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \quad \text{for all integers } n, k : 1 \leq k \leq n-1$$

$$(e) \sum_{m=0}^n \binom{m}{j} \binom{n-m}{k-j} = \binom{n+1}{k+1}$$

$$(f) \sum_{m=k}^n \binom{m}{k} = \binom{n+1}{k+1}$$

$$(g) \sum_{r=0}^m \binom{n+r}{r} = \binom{n+m+1}{m}$$

$$(h) \binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1}$$

$$(i) \binom{n-1}{k} - \binom{n-1}{k-1} = \frac{n-2k}{n} \binom{n}{k}$$

$$(j) \binom{n}{h} \binom{n-h}{k} = \binom{n}{k} \binom{n-k}{h}$$

6. Some tricks concerning integration

(a) **Integration by parts:**

Define  $F(x)$  such that  $\frac{dF(x)}{dx} = f(x)$ . Then

$$\int_a^b f(x)g(x)dx = F(x)g(x)|_a^b - \int_a^b F(x)\frac{dg(x)}{dx}dx$$

Example

$$\begin{aligned} I &= \int_0^\infty x^2 e^{-x} dx = -x^2 e^{-x}|_0^\infty - \int_0^\infty 2x(-e^{-x})dx \\ &= 2 \int_0^\infty x e^{-x} dx = 2 \left( -xe^{-x}|_0^\infty - \int_0^\infty (-e^{-x})dx \right) \\ &= 2 \int_0^\infty e^{-x} dx = 2 [-e^{-x}]_0^\infty = 2 \end{aligned}$$

(b) **Integration by substitution:**

Suppose you need to integrate a function of the form  $h(x) = f(g(x))\frac{dg(x)}{dx}$ . Then you can use that

$$\int_a^b h(x)dx = \int_{g(a)}^{g(b)} f(u)du$$

Why?

$$\int_a^b h(x)dx = \int_a^b \frac{F(g(x))}{dx}dx = F(g(b)) - F(g(a)) = \int_{g(a)}^{g(b)} f(u)du$$

Example. Take  $h(x) = 2x \exp(-x^2)$ . Then,

$$I = \int_0^2 2x \exp(-x^2)dx = \int_0^4 \exp(-u)du = 1 - e^{-4}$$

7. PDFs:

Name	Param.	PMF or PDF	Mean	Variance
Bernoulli	$p$	$P(X = 1) = p, P(X = 0) = q$	$p$	$pq$
Binomial	$n, p$	$\binom{n}{k} p^k q^{n-k}, \text{ for } k \in \{0, 1, \dots, n\}$	$np$	$npq$
FS	$p$	$pq^{k-1}, \text{ for } k \in \{1, 2, \dots\}$	$1/p$	$q/p^2$
Geom	$p$	$pq^k, \text{ for } k \in \{0, 1, 2, \dots\}$	$q/p$	$q/p^2$
NBinom	$r, p$	$\binom{r+n-1}{r-1} p^r q^n, n \in \{0, 1, 2, \dots\}$	$rq/p$	$rq/p^2$
HGeom	$w, b, n$	$\frac{\binom{w}{k} \binom{b}{n-k}}{\binom{w+b}{n}}, \text{ for } k \in \{0, 1, \dots, n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1}) n \frac{\mu}{n} (1 - \frac{\mu}{n})$
Poisson	$\lambda$	$\frac{e^{-\lambda} \lambda^k}{k!}, \text{ for } k \in \{0, 1, 2, \dots\}$	$\lambda$	$\lambda$
Uniform	$a < b$	$\frac{1}{b-a}, \text{ for } x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\mu, \sigma^2$	$\frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	$\mu$	$\sigma^2$
Log-Normal	$\mu, \sigma^2$	$\frac{1}{x \sigma \sqrt{2\pi}} e^{-(\log x - \mu)^2/(2\sigma^2)}, x > 0$	$\theta = e^{\mu + \sigma^2/2}$	$\theta^2(e^{\sigma^2} - 1)$
Expo	$\lambda$	$\lambda e^{-\lambda x}, \text{ for } x > 0$	$1/\lambda$	$1/\lambda^2$
Gamma	$a, \lambda$	$\Gamma(a)^{-1} (\lambda x)^a e^{-\lambda x} x^{-1}, \text{ for } x > 0$	$a/\lambda$	$a/\lambda^2$
Beta	$a, b$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \text{ for } 0 < x < 1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
Chi-Square	$n$	$\frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \text{ for } x > 0$	$n$	$2n$
Student- $t$	$n$	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} (1 + x^2/n)^{-(n+1)/2}$	0 if $n > 1$	$\frac{n}{n-2}$ if $n > 2$

## Layman's talk : what is probability

Probability is a function

●○○○○○

# Probability is a function

In our daily life, we sometimes would see:

- ▶ this stock price is more likely to increase
- ▶ the risk of default is getting higher
- ▶ chances are that tomorrow is going to rain

The aforementioned are predictions, and there are many other relevant words such as coincidence, lucky or randomness. How do we formally discuss these terms?

## Probability!

**A measure of uncertainty and random events.**

## Quantify the uncertainty

A natural thought would be assigning numbers to those random events, the larger the number the more likely it is going to happen.

- ▶
- ▶ This is similar to measuring one's height:
  - (a) pick one people from a group of people
  - (b) use a ruler to measure the height
  - (c) You assign each people a number to represent their heights; the larger number the taller

The above procedure is a function!

- ▶ Similarly, probability is also a function, relates random events to real numbers within  $[0,1]$ .

## Probability is a function

$$\begin{array}{ccc} \mathbb{P} : & A & \mapsto x \\ & \subseteq & \in \\ & S & [0, 1] \end{array}$$

where  $S$  denotes a collection of all possible outcomes, or a sample space, and  $A$  denotes an arbitrary (**but properly chosen**) subset of  $S$ .

For example, if we return to the first sentence in the slides we could denote  $S = A \cup B$  where  $A$  represents the event where the stock price increases while  $B$  represents the event it doesn't increase, and if we have  $\mathbb{P}(A) > 1/2$  then we can say it is more likely to increase.

## Probability is a special function

$$\mathbb{P}(\emptyset) = 0;$$

$$\mathbb{P}(S) = 1;$$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

where  $A_i \cap A_j = \emptyset, i \neq j$ . The right-hand side of the third equation (countable additivity) involves a summable/convergent infinite series.

Return to the example,  $S = A \cup B$  where A represents the event where the stock price increases while B represents the event it doesn't increase. Here, apparently we would have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

Probability is a function

oooooo●

Probability is a function with special properties.

Probability is a function

oooo

(a) Elements in the domain of probability

ooooo

(c) link events to real numbers within [0,1]

oooo

# Probability Theory for EOR

## Sample Space and Naive Definition of Probability

Probability is a function

●○○○

(a) Elements in the domain of probability  
○○○○○

(c) link events to real numbers within [0,1]  
○○○○

# Probability is a function

- ▶ To quantify the uncertainty and randomness, we want a number for a random event. The larger, the higher chance for such event to happen.
  - ▶ For example, when we throw a coin, what is a proper number you would have in mind for the event getting the head of the coin?
  - ▶ Fifty-fifty ( $1/2$ ).
  - ▶ What about the event of getting either head or the bottom?
  - ▶ Pretty sure 100% chances (1).
- ▶ Those are the informal/causal ways we use the idea of probability in daily life:

probability maps **random events** to **numbers** (non-negative numbers within 0 and 1).

# Probability is a function

- ▶ Essentially, the probability is a function. Like a ruler measure heights, it measures how likely one event is going to happen.
- ▶ Let  $A$  denote a random event containing some possible random outcomes.

$$\begin{array}{ccc} \mathbb{P}: & \{A_i \subseteq S, i \in I\} & \mapsto & [0, 1] \\ & A & \mapsto & x \end{array}$$

where  $S$  denotes a collection of all possible outcomes, or a sample space, and  $A$  denotes an arbitrary (but properly chosen) subset of  $S$ .

- ▶ To understand a function (maps elements from one set to elements in another set), we need to understand: (a) domain of the function (a collection of the subsets of  $S$ ,  $\{A_i \subseteq S, i \in I\}$ ); (b) codomain of the function ( $[0,1]$ )); (c) how the elements from the two sets are linked to each other.

# Probability is a function

- ▶ Let  $A$  denote a random event containing some possible random outcomes.

$$\begin{array}{ccc} \mathbb{P}: & \{A_i \subseteq S, i \in I\} & \mapsto & [0, 1] \\ & A & \mapsto & x \end{array}$$

where  $S$  denotes a collection of all possible outcomes, or a sample space, and  $A$  denotes an arbitrary (but properly chosen) subset of  $S$ .

- ▶ To understand a function (maps elements from one set to elements in another set), we need to understand: (a) domain of the function (a collection of the subsets of  $S$ ,  $\{A_i \subseteq S, i \in I\}$ ); (b) codomain of the function ( $[0,1]$ ); (c) how the elements from the two sets are linked to each other.
- ▶ To understand a function (maps elements from one set to elements in another set), we need to understand: (a) domain of the function (a collection of the subsets of  $S$ ,  $\{A_i \subseteq S, i \in I\}$ ); (b) codomain of the function ( $[0,1]$ ); (c) how the elements from the two sets are linked to each other.
- ▶ To understand a function (maps elements from one set to elements in another set), we need to understand: (a) domain of the function (a collection of the subsets of  $S$ ,  $\{A_i \subseteq S, i \in I\}$ ); (b) codomain of the function ( $[0,1]$ ); (c) how the elements from the two sets are linked to each other.
- ▶ To understand a function (maps elements from one set to elements in another set), we need to understand: (a) domain of the function (a collection of the subsets of  $S$ ,  $\{A_i \subseteq S, i \in I\}$ ); (b) codomain of the function ( $[0,1]$ ); (c) how the elements from the two sets are linked to each other.

Probability is a function

○○○○

(a) Elements in the domain of probability

●○○○○

(c) link events to real numbers within [0,1]

○○○○

## (a) Elements in the domain of probability

# Sets of possible outcomes are elements in the domain of one probability function

## Definition (Event, sample space)

The **sample space**  $S$  is the set of all possible **outcomes** of the **experiment**. An **event**  $A$  is a **subset** of the sample space  $S$ , and we say that  $A$  **occurred** if the actual outcome is in  $A$ .

Note that:

- (1)  $S \subseteq S$ , and we always include  $S$  as one event (the largest one as it contains all possible outcomes);
- (2) if  $A$  and  $B$  are two events, then their intersections, unions are also included as events, so are their complements;
- (3)  $\emptyset$  is also an event (zero probability though);
- (4) if  $A_i$ 's are events, so is  $\cup_{i=1}^{\infty} A_i$ .

Probability is a function

○○○○

(a) Elements in the domain of probability

○○●○○

(c) link events to real numbers within [0,1]

○○○○

## Example I: coin tossing for one time

- ▶  $S = \{\text{Head, Bottom}\};$
- ▶  $A = \{\text{Head}\}, B = \{\text{Bottom}\};$
- ▶  $A \cup B = S;$
- ▶  $A \cap B = \emptyset;$
- ▶  $A^c = B.$

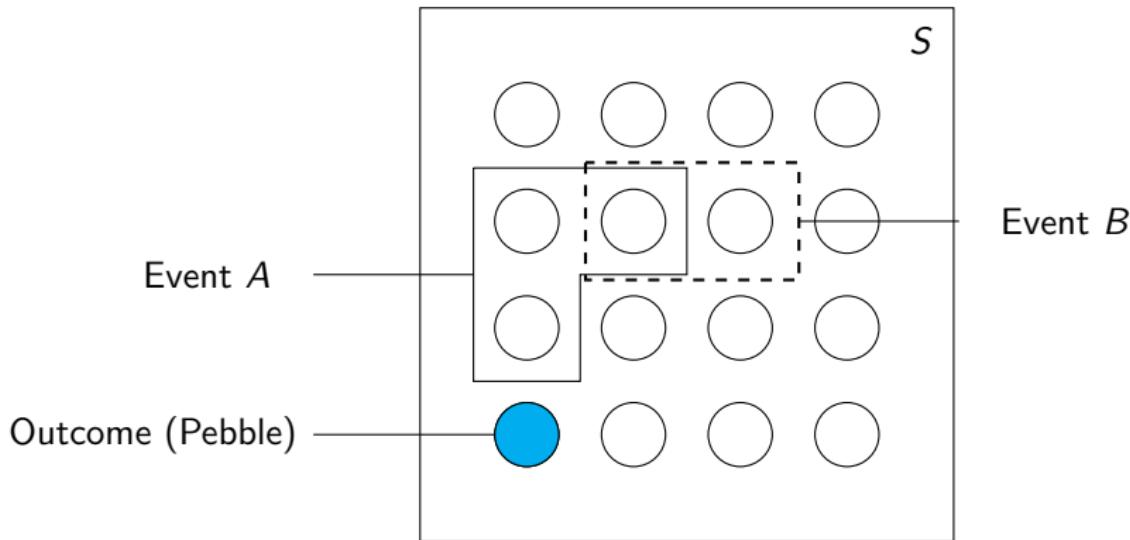
Probability is a function  
○○○○

(a) Elements in the domain of probability  
○○○●○

(c) link events to real numbers within [0,1]  
○○○○

## Example II: pebble space

(all the 16 pebbles are equally likely to be chosen).



# DIY: Laws

- ▶ Commutative laws

- ▶  $A \cap B = B \cap A$
- ▶  $A \cup B = B \cup A$

- ▶ Associative laws

- ▶  $(A \cup B) \cup C = A \cup (B \cup C)$
- ▶  $(A \cap B) \cap C = A \cap (B \cap C)$

- ▶ Distributive laws

- ▶  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
- ▶  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

Probability is a function

oooo

(a) Elements in the domain of probability

ooooo

(c) link events to real numbers within [0,1]

●ooo

(c) link events to real numbers within [0,1]

## First attempt: naive definition of probability

- ▶ Events are sets of possible outcomes.
- ▶ Let's consider a special case: finite outcomes, all the outcomes are equally likely to happen.
- ▶ The more outcomes one event contains, the more likely it happens.

### Definition (Naive definition of probability)

Denote  $A$  be an event for an experiment with a **finite\*** sample spaces  $S$  and each outcome is **equally likely\*** to happen:

$$P_{\text{Naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes in } S} = \frac{|A|}{|S|}.$$

Probability is a function  
○○○○

(a) Elements in the domain of probability  
○○○○○

(c) link events to real numbers within [0,1]  
○○●○

## Example I: coin tossing for one time

(head and bottom are equally likely to happen)

- ▶  $S = \{\text{Head, Bottom}\};$
- ▶  $A = \{\text{Head}\}, B = \{\text{Bottom}\};$

$$\mathbb{P}_{\text{Naive}}(A) = 1/2, \mathbb{P}_{\text{Naive}}(B) = 1/2, \mathbb{P}_{\text{Naive}}(A \cup B) = 1.$$

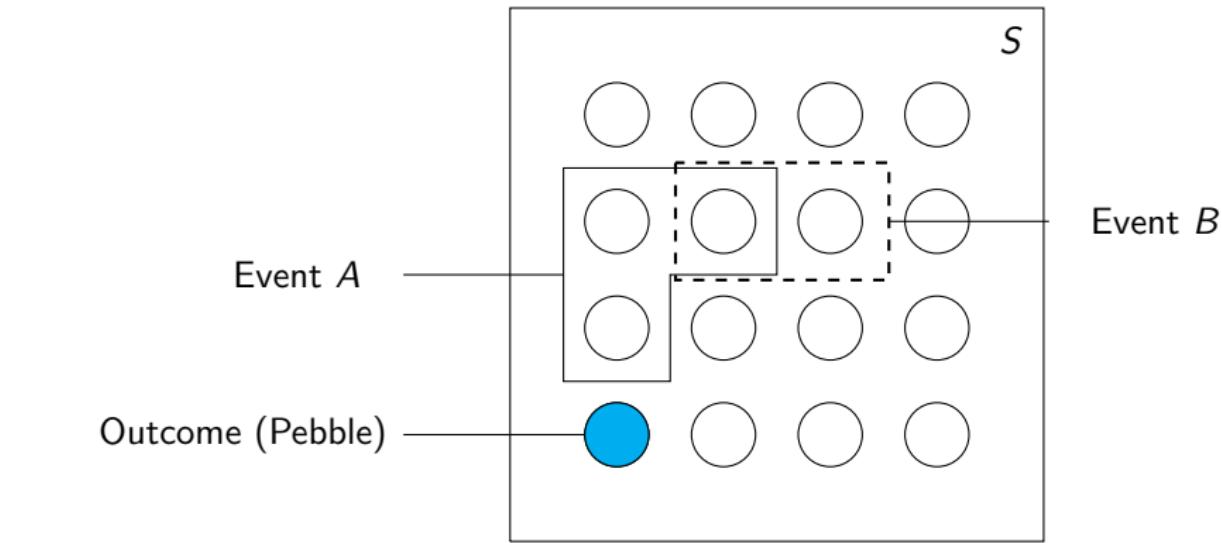
Probability is a function  
○○○○

(a) Elements in the domain of probability  
○○○○○

(c) link events to real numbers within [0,1]  
○○○●

## Example II: pebble space

(all the 16 pebbles are equally likely to be chosen).



$$\begin{aligned}\mathbb{P}_{\text{Naive}}(A) &= 3/16, \mathbb{P}_{\text{Naive}}(B) = 1/8, \\ \mathbb{P}_{\text{Naive}}(A \cup B) &= 1/4, \mathbb{P}_{\text{Naive}}(A \cap B) = 1/16.\end{aligned}$$

Problem I: hands in poker

oooooooooo

Problem II: full house with dice

ooooo

Differences in the two problems

oo

## Probability Theory for EOR

How to count the number of possible outcomes in one event

Problem I: hands in poker  
oooooooooo

Problem II: full house with dice  
ooooo

Differences in the two problems  
oo

Gottfried Wilhelm Leibniz (1646-1716)

*Music is the pleasure that human mind experiences from counting without being aware that it is counting.*

Problem I: hands in poker  
oooooooooo

Problem II: full house with dice  
ooooo

Differences in the two problems  
oo

## Definition (Naive definition of probability)

Denote  $A$  be an event for an experiment with a **finite\*** sample spaces  $S$  and each outcome is **equally likely\*** to happen:

$$P_{\text{Naive}}(A) = \frac{\text{number of outcomes favorable to/in } A}{\text{number of all possible outcomes in } S} = \frac{|A|}{|S|}.$$

It is designed for common but very limited cases (**finite equally likely outcomes**), yet, it is already complicated in certain cases.

Not always easy to count the number of outcomes in one event.

We need **binomial coefficient**:  $\binom{n}{k} := \frac{n!}{k!(n-k)!}$ .

Problem I: hands in poker

●○○○○○○○○

Problem II: full house with dice

○○○○○

Differences in the two problems

○○

## Problem I: hands in poker

Problem I: hands in poker

○●○○○○○○

Problem II: full house with dice

○○○○○

Differences in the two problems

○○

## Example

A 5-card hand is dealt from a standard well-shuffled 52-card deck.

The hand is called a *full house* in poker if it consists of three cards from same rank and two cards of another rank, e.g. three 7's and two 10's (in any order).

What is the probability of a full house?

We need to know  $|S|$  and  $|A|$ , where  $S$  denotes the set of all possible outcomes, and  $A$  denotes the set of outcomes that are a full house.

# Number of all outcomes $|S|$

## Multiplication rule (sampling without replacement)

- ▶ We have  $n$  cards.
- ▶ We select  $k$  cards *one at a time*.
- ▶ We keep the selected card (without replacement).

How many outcomes ( $N$ ) when order matters?

- ▶ Step 1:  $n$  possible outcomes. Keep the card!
- ▶ Step 2:  $n - 1$  possible outcomes. Keep the card!
- ⋮
- ▶ Step  $k$ :  $n - k + 1$  possible outcomes.

MR:  $N = n \cdot (n - 1) \cdot \dots \cdot (n - k + 1)$  ordered outcomes.

Problem I: hands in poker  
○○○●○○○○

Problem II: full house with dice  
○○○○○

Differences in the two problems  
○○

# Number of all outcomes $|S|$

## Definition

### Factorial

We define  $n!$  (say:  $n$  factorial) as

$$\begin{aligned} n! &= n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 1 \\ 0! &= 1 \end{aligned}$$

$$N = n \cdot (n - 1) \cdot \dots \cdot (n - k + 1) = \frac{n!}{(n - k)!}$$

# Number of all outcomes $|S|$

## Definition

### Factorial

We define  $n!$  (say:  $n$  factorial) as

$$n! = n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 1$$
$$0! = 1$$

$$N = n \cdot (n - 1) \cdot \dots \cdot (n - k + 1) = \frac{n!}{(n - k)!}$$

Alternative interpretation of how we arrive at this number of ordered outcomes:

- We can order the cards in  $n!$  ways. Permutation of  $n$  items.
- Don't care about the ordering of the  $(n - k)$  not selected cards. The not-selected cards can be ordered in  $(n - k)!$  ways.

But to calculate  $|S|$ , we still need to adjust for over-counting!

**We don't care for the orders of selected cards either!** Divide by  $k!$ .

Problem I: hands in poker  
○○○○○●○○○

Problem II: full house with dice  
○○○○○

Differences in the two problems  
○○

## Number of all outcomes $|S|$

We adjust for overcounting:  $|S| = N/k! = \frac{n!}{k!(n-k)!} = \binom{n}{k}$

The **binomial coefficient** gives the number of possible outcomes of picking  $k$  items out of  $n$  items (where orders do not matter).

In our case:  $n = 52, k = 5$ .

Also, we know these outcome are equally likely, which is why we can use the naive probability.

Problem I: hands in poker  
oooooooo●oo

Problem II: full house with dice  
ooooo

Differences in the two problems  
oo

## Number of outcomes in A, $|A|$

- ▶ Choose the rank that we have the three cards of. **13**. Keep the rank
- ▶ How many different ways of choosing three cards of a given rank (orders does not matter): we have to **choose 3 out of 4 suits**  $\binom{4}{3}$ .
- ▶ Choose the rank that we have the two cards of from the remaining ranks. **12**.
- ▶ How many different ways of choosing two cards of a given rank (orders does not matter): we have to **choose 2 out of 4 suits**  $\binom{4}{2}$ .
- ▶  $|A| = 13\binom{4}{3}12\binom{4}{2}$

Problem I: hands in poker

○○○○○○○●○

Problem II: full house with dice

○○○○○

Differences in the two problems

○○

## What is the probability of a full house?

$$\mathbb{P}(\text{Full house}) = |A|/|S| = 13 \binom{4}{3} 12 \binom{4}{2} / \binom{52}{5}$$

Problem I: hands in poker

oooooooo●

Problem II: full house with dice

ooooo

Differences in the two problems

oo

Check examples of *Poker probability* on Wikipedia and practice the counting strategy.

It is fun.

Problem I: hands in poker

○○○○○○○○

Problem II: full house with dice

●○○○○

Differences in the two problems

○○

## Problem II: full house with dice

Problem I: hands in poker  
○○○○○○○○

Problem II: full house with dice  
○●○○○

Differences in the two problems  
○○

## Example

Suppose we roll 5 identical dice. What is the probability of a full house (three dice of a same number and two remaining dice of a same but different number, e.g., three 3's and two 6's)?

Note that in this case, **order** matters.

Problem I: hands in poker  
○○○○○○○○

Problem II: full house with dice  
○○●○○

Differences in the two problems  
○○

## Number of all outcomes $|S|$

### Multiplication rule (sampling with replacement)

- ▶ We have  $k$  dice.
- ▶ each would have  $n$  outcomes.

How many outcomes ( $N$ ) when order matters?

- ▶ Dice 1:  $n$  possible outcomes.
- ▶ Dice 2:  $n$  possible outcomes.
- ⋮
- ▶ Dice  $k$ :  $n$  possible outcomes.

MR:  $|S| = n^k$ . (23333 different from 33332)

Problem I: hands in poker  
○○○○○○○○

Problem II: full house with dice  
○○○●○

Differences in the two problems  
○○

## Number of outcomes in A, $|A|$ (ordered)

- ▶ First we select the number,  $a$ , that we have three dice of (a triplet). **6.**
- ▶ How many different ways of choosing three dice out 5 to have the number  $a$ : **choose 3 out of 5**  $\binom{5}{3}$ .
- ▶ Select the number,  $b (\neq a)$ , that we have two dice of (a pair). **5.**
- ▶ The remaining two dice must have the number  $b$ .
- ▶ Note that here we would have the number of *Ordered* outcomes:  
 $|A| = 6 \binom{5}{3} 5$ .

Problem I: hands in poker  
○○○○○○○○

Problem II: full house with dice  
○○○○●

Differences in the two problems  
○○

## What is the probability of a full house?

$$\mathbb{P}(\text{Full house}) = |A|/|S| = 6 \binom{5}{3} 5/6^5$$

Problem I: hands in poker

○○○○○○○○○○

Problem II: full house with dice

○○○○○

Differences in the two problems

●○

## Differences in the two problems

- ▶ Without replacement or with replacement. (Or, whether the outcome number changes or not).
- ▶ **Orders.** In Problem I, there is the term "in any order"; while in Problem II there is not such term.  
We do so in order to make sure each outcome would be equally likely. **Order matters when replacement is present.**
- ▶ The way we count may not be unique.  
For example, you can also consider orders in Problem I (then  $|S|$  and  $|A|$  are both  $5!$  times larger than the ones without considering the orders, think about why), you will find the probability remains the same.
- ▶ Practice more.

If you wonder why **order matters when replacement is present**. Think about the simple experiment: roll two dice and think about the following two events (a) two dice are both ones and (b) one die is 2 and another one is 3. Are they equally likely?

# Probability Theory for EOR

Story proof of binomial coefficient equalities

The **binomial coefficient**  $\binom{n}{k}$  counts the number of ways to form an unordered collection of  $k$  items chosen from a collection of  $n$  distinct items.

- We know that the binomial coefficient is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

We set it to be zero if  $n < k$ .

- Many mathematical identities involve the binomial coefficient.
- Sometimes, these identities can be easily proven algebraically.

e.g.,

$$(a + b)^n = \sum_{j=0}^n \binom{n}{j} a^j b^{n-j}$$

## Combinatorial identities: algebra

$$n \binom{n-1}{k-1} = k \binom{n}{k}$$

$$n \frac{(n-1)!}{(k-1)! \cdot (n-k)!} = \frac{n!}{\left(\frac{1}{k}\right) \cdot k! \cdot (n-k)!} = k \binom{n}{k}$$

- ▶ Algebra is ok, but
  - ▶ it is easy to make mistakes.
  - ▶ it can be boring.
- ▶ We can also make up a story (think about the underlying meaning of these coefficients) to prove identities involving the binomial coefficients.

## Combinatorial identities: story proof, I

$$n \binom{n-1}{k-1} = k \binom{n}{k}$$

Ways of counting the same number is not unique.

**Prove the equality by showing both sides are counting the same number!**

**Consider a group of  $n$  people. We need to select a  $k$ -people group consisting of a leader and  $k - 1$  followers.**

Left hand side: There are  $n$  choices for the leader, and then  $\binom{n-1}{k-1}$  ways to select the followers.

Right hand side: first select the whole group:  $\binom{n}{k}$  options. Then select the leader from the  $k$  selected persons:  $k$  options.

## Combinatorial identities: story proof, II

$$\binom{n}{k} \binom{n-k}{m} = \binom{n}{m} \binom{n-m}{k}$$

**Count the ways of getting two separate groups of size  $k$  and  $m$  respectively with given  $n$  people.**

Left hand side: First draw the group of size  $k$  and then from the remaining persons draw a group of size  $m$ .

Right hand side: First the group of size  $m$  and then from the remaining persons a group of size  $k$ .

## Example: Vandermonde's identity

$$\binom{m+n}{k} = \sum_{j=0}^k \binom{m}{j} \binom{n}{k-j}$$

Left hand side: You have  $m$  apples and  $n$  pears and want to select  $k$  pieces of fruit.

Right hand side: first select  $j$  apples and then the remaining  $k - j$  to be pears. You can do this for all  $j$ , so you sum over  $j$ .

# Can you do this one?

$$\sum_{k=0}^n k^2 \binom{n}{k}^2 = n^2 \binom{2(n-1)}{n-1}.$$

Hint:

$$\sum_{k=0}^n k^2 \binom{n}{k}^2 = \sum_{k=1}^n \binom{n}{k} k \binom{n}{n-k} k$$

**Form a  $(n+1)$ -people group from two groups of size  $n$  each, and each of the group need to select a leader.**

Left: select  $k$  people from group one, and select the first leader from the  $k$  people, then select the remaining  $n-k+1$  people from the group two by first selecting  $n-k$  followers and then the second leader from the remaining  $k$  people left in group two;

Right: first select the two leaders from two groups of size  $n$  each, and then select the  $n-1$  followers from the remaining  $2(n-1)$  people.

# Probability Theory for EOR

## Non-naive definition of probability

## General definition of probability

## Definition ( Naive definition of probability)

Denote  $A$  be an event for an experiment with a **finite\*** sample spaces  $S$  and each outcome is **equally likely\*** to happen:

$$P_{\text{Naive}}(A) = \frac{\text{number of outcomes favorable to/in } A}{\text{number of outcomes in } S} = \frac{|A|}{|S|}.$$

Very limited cases with (**finite equally likely outcomes**).

- ▶ A probability measure assigns probabilities to events.
- ▶ But how in general cases?
- ▶ After centuries of thinking, we have agreed on two

## *Axioms of Probability*

## Definition (General definition of probability)

A **probability function (measure)**  $P$  maps an event , a (well-constructed) subset  $A$  of the sample space  $S$  ( $A \subseteq S$ ), to the probability of the event  $P(A)$ , a real number within  $[0, 1]$ . It should satisfy the *Axioms of Probability*

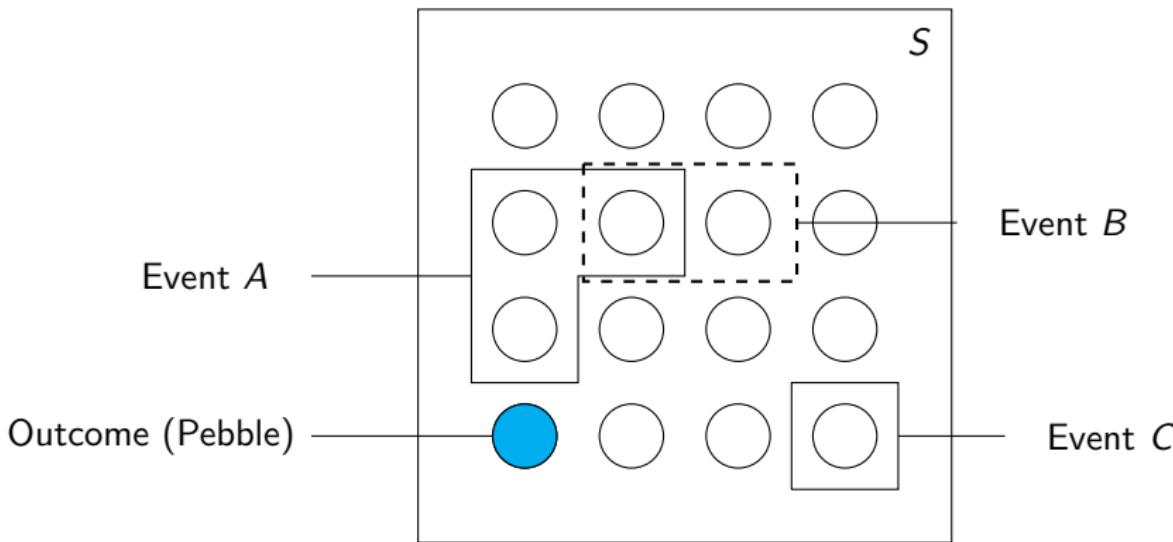
1.  $P(\emptyset) = 0$ , and  $P(S) = 1$ .
2. If  $A_1, A_2, \dots$  are disjoint (i.e. mutually exclusive,  $A_i \cap A_j = \emptyset, i \neq j$ ) events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- **Note:** reporting probabilities outside  $[0, 1]$  without the disclaimer that this cannot be the right answer will be considered a capital blunder and renders your complete answer invalid (even if that answer is 99% correct).
- The second axiom (countable additivity) involves a summation of a infinite series. It is more intuitive if you only choose finitely many non-empty  $A_i$ 's:  $P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$ .

## Example

The second axiom (countable additivity) involves a summation of a infinite series. It is more intuitive if you only choose finitely many non-empty  $A_i$ 's:  $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$ .



$$P(A) = 3/16, P(B) = 1/8, P(C) = 1/16, P(A \cup B) = 1/4, P(A \cup C) = 1/4.$$

## Quiz: which one is potentially a probability function?

Suppose we have a domain of probability:  $\{A, A^c, \emptyset, S\}$  and functions  $P_1, P_2, P_3, P_4$ .

- $P_1(\emptyset) = 1$ .
- $P_2(S) = 0$ .
- $P_3(A) = 0, P_3(\emptyset) = 0, P_3(A^c) = 1, P_3(S) = 1$ .
- $P_4(A) = 0, P_4(\emptyset) = 0, P_4(A^c) = 1/2, P_4(S) = 1$ .

Only  $P_3$ !

## General definition embeddes the naive one

- ▶  $P_{\text{Naive}}(\emptyset) = |\emptyset|/|S| = 0.$
- ▶  $P_{\text{Naive}}(S) = |S|/|S| = 1.$
- ▶  $P_{\text{Naive}}(\cup_{i=1}^m A_i) = |\cup_{i=1}^m A_i| / |S| = \sum_{i=1}^m |A_i| / |S| = \sum_{i=1}^m P_{\text{Naive}}(A_i)$

Now we can directly work with the general definition, and the naive one is simply a special case of the general definition.

The properties derived based on the general definition also holds for the naive one.

General definition of probability  
○○○○○

Properties of probability: I  
●○

Properties of probability: II  
○○

An exercise  
○○

# Properties of probability: I

A **probability function (measure)**  $P$  maps an event to a real number within  $[0, 1]$ . It should satisfy the *Axioms of Probability*

1.  $P(\emptyset) = 0$ , and  $P(S) = 1$ .
2. If  $A_1, A_2, \dots$  are disjoint (i.e. mutually exclusive) events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

## Theorem

For any events  $A$  and  $B$ , we would have

- $P(A^c) = 1 - P(A)$ .
- If  $A \subseteq B$ ,  $P(A) \leq P(B)$ .
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

General definition of probability  
○○○○○

Properties of probability: I  
○○

Properties of probability: II  
●○

An exercise  
○○

## Properties of probability: II

**A probability function (measure)**  $P$  maps an event , a (well-constructed) subset  $A$  of the sample space  $S$  ( $A \subseteq S$ ), to a real number within  $[0, 1]$ . It should satisfy the *Axioms of Probability*

1.  $P(\emptyset) = 0$ , and  $P(S) = 1$ .
2. If  $A_1, A_2, \dots$  are disjoint (i.e. mutually exclusive) events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

## Theorem

(*Inclusion-exclusion principle: two events*)

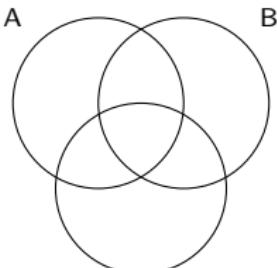
For any events  $A$  and  $B$ , we would have  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

## Theorem

(*Inclusion-exclusion principle: three events*)

For any events  $A$ ,  $B$  and  $C$ , we would have

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$



General definition of probability  
○○○○○

Properties of probability: I  
○○

Properties of probability: II  
○○

An exercise  
●○

## An exercise

## Example: void in at least one suit

**Exercise** What is the probability that a 13-card hand is void in at least one suit?

Conditional probability is also a probability (function)

○○○○○

Independence means products.

○○

## Layman's talk : what is conditional probability

Conditional probability is also a probability (function)

●○○○○

Independence means products.

○○

Conditional probability is also a probability  
(function)

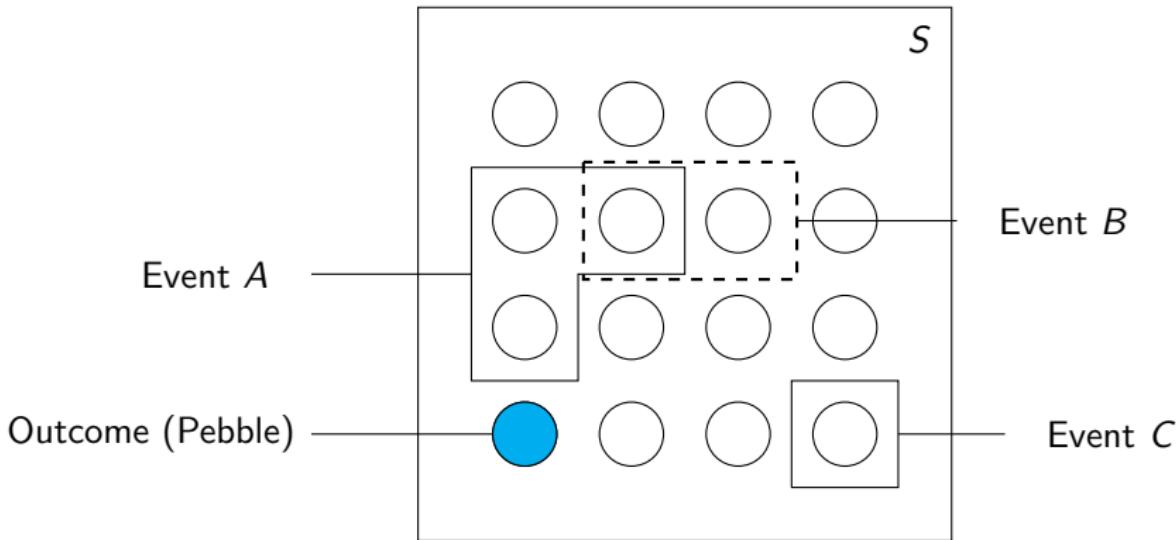
Suppose we have already had a probability function maps a collection of events to numbers within  $[0, 1]$ .

How do we quantify the uncertainty if we know an event  $B$  ( $P(B) > 0$ ) already happen?

Conditional probability (also a probability) maps again events to numbers within  $[0, 1]$ .

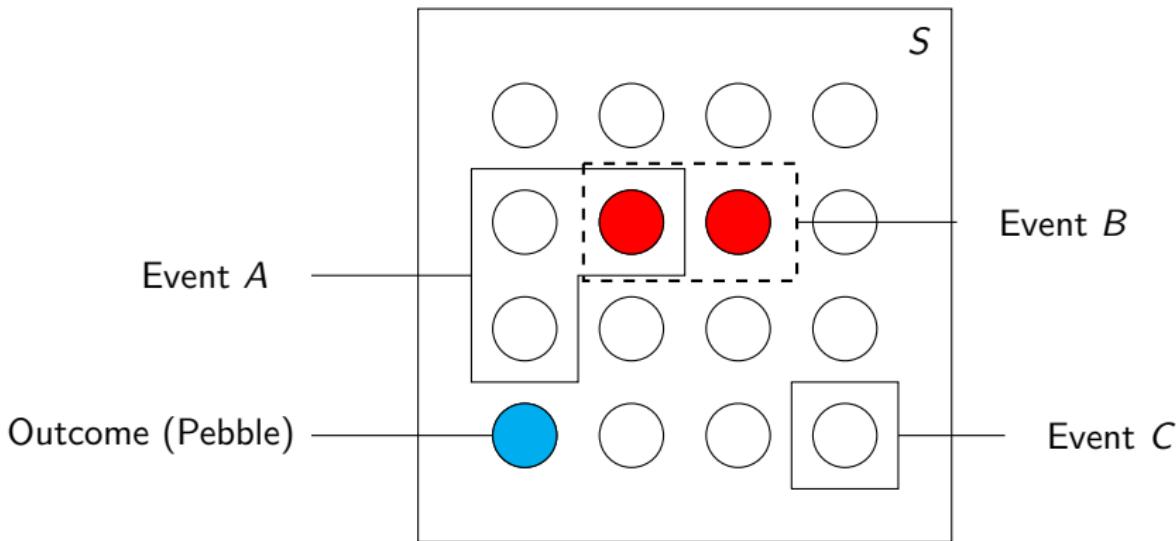
## Example

Each pebble is equally likely to be chosen.



## Example

Each pebble is equally likely to be chosen.



If we know  $B$  occurred, what can you say about event  $A$  and  $C$ ?  
 $P(A|B) = 1/2, P(C|B) = 0, P(B|B) = 1$  ( $B$  is like our new sample space.)

$$\begin{aligned}\mathbb{P}(\cdot|B) : \quad & \{A_i \subseteq S, i \in I\} & \mapsto & [0, 1] \\ & A & \mapsto & P(A \cap B)/P(B)\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\cdot|B) : \quad & \{A_i \subseteq S, i \in I\} & \mapsto & [0, 1] \\ & A & \mapsto & P(A \cap B)/P(B)\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\cdot|B) : \quad & \{A_i \subseteq S, i \in I\} & \mapsto & [0, 1] \\ & A & \mapsto & P(A \cap B)/P(B) (*)\end{aligned}$$

Measure how likely A occurred given the B already occurred ( $P(B) > 0$ ).  
 (Informally, the percentage of outcomes in B that also in A.)

# How the conditional probability/probability are linked to each other?

$$P(A|B) = P(A \cap B)/P(B) (*)$$

- ▶ Bayes' rule:



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

- ▶ LOTP:



$$P(A) = \sum_i P(A|B_i)P(B_i); \cup_i B_i = S, B_i \cap B_j = \emptyset, (i \neq j).$$

Conditional probability is also a probability (function)

○○○○○

Independence means products.

●○

Independence means products.

Think about throwing two fair coins separately, knowing the results of the first coin does not provide any information on the second one.

With  $P(B) > 0$ :

$$P(A|B) = P(A).$$

Eqivalently,

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

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

**Independence means products.**

## Probability Theory for EOR

Conditional probability (conditional on a non-zero probability event)

# Definition and Intuition

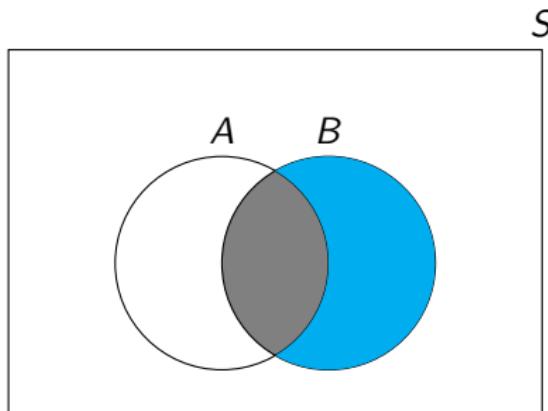
# Definition

## Definition

**Conditional probability of A given B ( $P(B) > 0$ ):**

if  $A$  and  $B$  are events with  $P(B) > 0$ , then the conditional probability of  $A$  given  $B$ , denoted by  $P(A|B)$ , is defined as  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ .

$$\begin{array}{lll} \mathbb{P}(\cdot|B) : & \{A_i \subseteq S, i \in I\} & \mapsto [0, 1] \\ & A & \mapsto P(A \cap B)/P(B) (*) \end{array}$$



$$P(B|B) = ? \quad P(A|B) = ?$$

# Conditional probability is a probability

**We have a new probability!**

- ▶  $P(S|B) = 1;$
- ▶  $P(\emptyset|B) = 0;$
- ▶  $P(\cup_i A_i|B) = \sum_i P(A_i|B), A_i \cap A_j = \emptyset (\forall, i \neq j).$

# In Latin

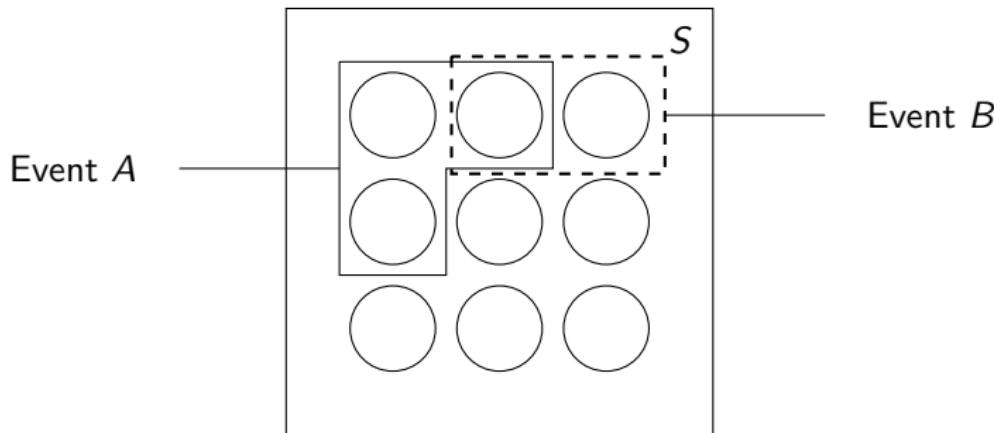
## Some terminology

- ▶  $P(A)$  is sometimes called the *prior* of  $A$  (or unconditional/marginal probability of event  $A$ ).
- ▶ When information  $B$  enters, we *update* the prior.
- ▶ The updated probability  $P(A|B)$  is called the *posterior* of  $A$  given  $B$  (or conditional probability of event  $A$  conditional on event  $B$ ).

# Example

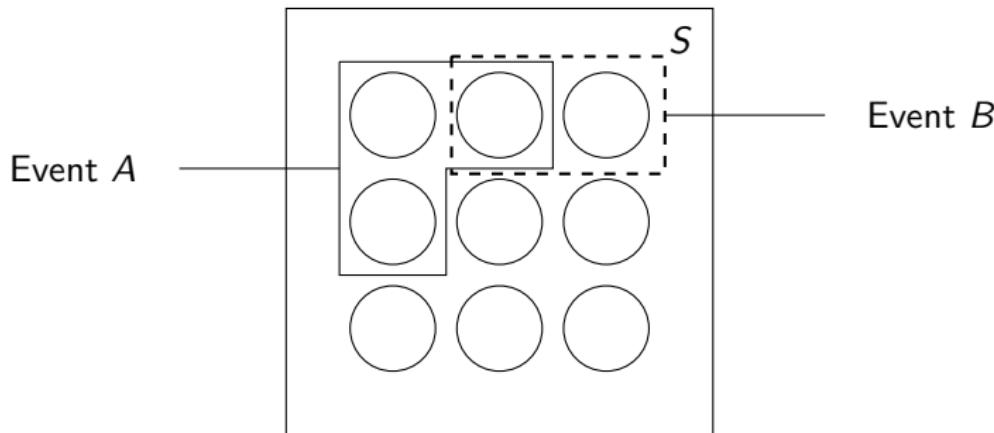
## Example: visualizing conditional probabilities

If we know that  $B$  happens, what is the probability that  $A$  happens?



## Example: visualizing conditional probabilities

If we know that  $A$  happens, what is the probability that  $B$  happens?



# Important lesson

(In general)

$$P(A|B) \neq P(B|A)$$

$$P(A|B) \neq P(A|A)$$

# Probability Theory for EOR

Bayes' rule and the law of total probability (LOTP)

# Bayes' rule

# Conditional Probability (link one to another)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

## Theorem

### Bayes'rule

Suppose  $P(A), P(B) > 0$ .

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Unless  $P(A) = P(B) > 0$ ,  $P(A|B) \neq P(B|A)$ !

## A bit of a digression via a simple example

Now there are two schools in statistics: Frequentists and Bayesian.

**Let  $F$  be the event that a six-sided die is fair.**

There are cases you encounter unfair dies, and  $F$  can be regarded as one random event as well. We want to know whether  $F$  is true or not (the hypothesis we want to test).

**Let  $E$  be the event of throwing 7 times a 2.**

Then you can consider  $E$  as the data we observe.

► **Frequentist:**  $P(E|F)$ . Data given hypothesis.

► **Bayesian:**  $P(F|E)$ . Hypothesis given data.

By Bayes' rule, this requires  $P(F)$ : a prior belief on whether or not the die is fair, and might need some other priors as well.

Bayes' rule  
○○○

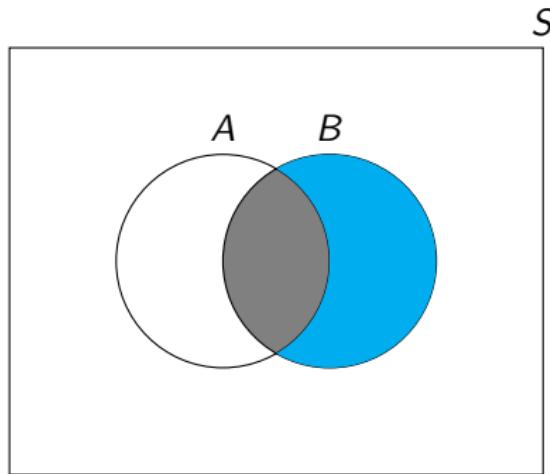
The law of total probability  
●○○○

Exercise  
○○○○

## The law of total probability

# Law Of Total Probability (LOTP)

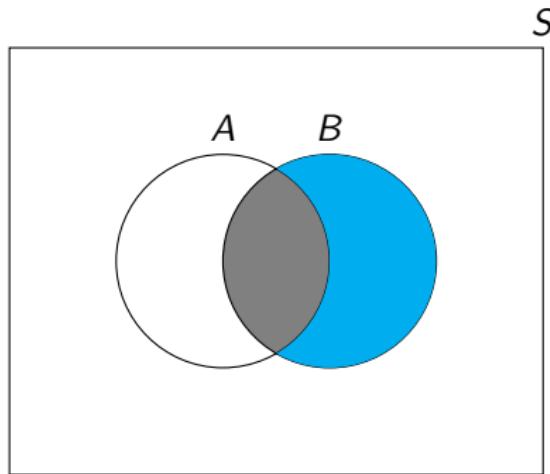
$$P(B) = P(B \cap A) + P(B \cap A^C)$$



$$P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$$

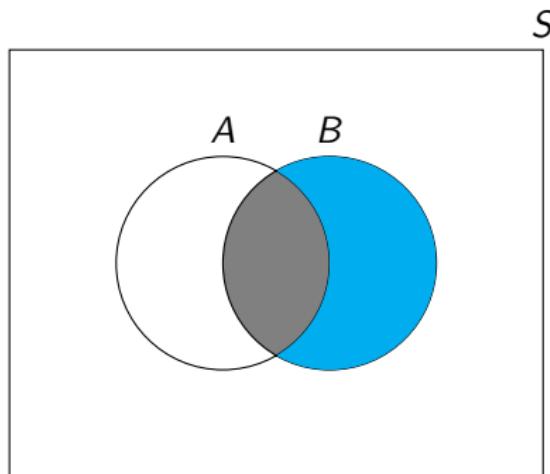
# Law Of Total Probability (LOTP)

$$P(B) = P(B \cap A) + P(B \cap A^C)$$



$$P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$$

## Law Of Total Probability (LOTP)



If we have a partition of  $S$ : disjoint  $A_i$ 's such that  $\cup_i A_i = S$  and  $P(A_i) > 0$ , then

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

Bayes' rule  
○○○

The law of total probability  
○○○○

Exercise  
●○○○

# Exercise

## Problem: medical tests

You decide to undergo a Corona test.

- ▶ The test is positive for 99 out of 100 people who have the virus (sensitivity)
- ▶ This test is negative for 99 out of 100 people who don't (specificity)

The current incidence of Corona is 1/1000.

You test positive, what is the probability you have the disease?

## Problem: medical tests

**Bayes' rule + LOTP:**

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}. \end{aligned}$$

## Problem: medical tests

$V$ : {carrying the Corona virus}.

$T_+$ : {testing positive}.

- ▶  $P(T_+|V) = 0.99$ .
- ▶  $P(T_+^C|V^C) = 0.99$ .
- ▶  $P(V) = p = 0.1\%$ .

$$P(V|T_+) = \frac{P(T_+|V)P(V)}{P(T_+|V)P(V)+P(T_+|V^C)P(V^C)} = \frac{0.99p}{0.99p+0.01(1-p)} = \frac{0.99p}{0.98p+0.01}.$$

Therefore, we know

$$P(V|T_+) = \frac{99}{98 + 1/p}.$$

$$\approx \mathbf{0.0901639344} \approx 9\% > 0.1\%. \text{ If } p=30\%, \frac{99}{98+1/p} \approx 0.976973684!$$

## Probability Theory for EOR

### Random variables and their distributions

## Independence/Conditional Independence (two random events)

●○○

Independence/Conditional Independence (two random events)

## Independence means products

- Sometimes, 'additional' information  $B$  does *not* change the probability of an event  $A$ .
- In cases with  $P(B) > 0$ ,  $P(A|B) = P(A)$ .  
What does this imply? If  $P(A) > 0$ ,  $P(B|A) = P(B)$ .

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

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

$$(if P(A) > 0) \Rightarrow P(B|A) = \frac{P(A \cap B)}{P(A)} = P(B)$$

- We say that the events  $A$  and  $B$  are independent.

**$A$  and  $B$  are independent if**

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

Quiz:  $P(\{\text{fist coin head}\} \text{ and } \{\text{second coin head}\}) = 1/4 = P(\{\text{fist coin head}\}) P(\{\text{second coin head}\})$ .

## Conditional independence means products

- $A$  and  $B$  are independent if

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

- Conditional independence!:  $A$  and  $B$  are independent **conditional on/given**  $C$  if

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

- Knowing that  $B$  (or  $A$ ) also occurred does not provide any extra information on the probability of  $A$  (or  $B$ ) once we already learn that  $C$  occurred:

$$P(A|C, B) = \frac{P(A \cap B|C)}{P(B|C)} = \frac{P(A|C)P(B|C)}{P(B|C)} = P(A|C)$$

$$P(B|C, A) = \frac{P(A \cap B|C)}{P(A|C)} = \frac{P(A|C)P(B|C)}{P(A|C)} = P(B|C)$$

$$P(A|C, B) := P(A|C \cap B) = \frac{P(A \cap (C \cap B))}{P(C \cap B)} = \frac{P((A \cap B) \cap C)}{P(B \cap C)} = \frac{P((A \cap B) \cap C) / P(C)}{P(B \cap C) / P(C)} = \frac{P(A \cap B|C)}{P(B|C)}.$$

When we write  $P(A|B)$  down, we always assume  $P(B) > 0$ . So here  $P(B \cap C) > 0$  as well.

Conditioning as a problem-solving tool  
○○○

Pitfalls and Paradoxes: I  
○○○○○

Pitfalls and Paradoxes: II  
○○○

Pitfalls and Paradoxes: III  
○○○○○

# Probability Theory for EOR

## Conditioning is a useful tool

Conditioning as a problem-solving tool

●○○

Pitfalls and Paradoxes: I

○○○○○

Pitfalls and Paradoxes: II

○○○

Pitfalls and Paradoxes: III

○○○○○

# Conditioning as a problem-solving tool

## Example: a simple game

You play a game where you start with 1 dollar. You flip a three sided fair die.

1. Throw 1: you lose one dollar.
2. Throw 2: you keep your dollar.
3. Throw 3: you win one dollar.

What is the probability that this is a never-ending game (*NEG*)?

## Solution via conditioning: conditional on the information you know

- ▶ The complement of ( $NEG$ ) is the probability that you lose all your money.
- ▶ Define  $D$  as  $NEG^C$ .
- ▶ At the start of the game, you have one dollar. Options
  - ▶ Lose the dollar, then  $D$  occurs
  - ▶ Keep the dollar, then the exact same game repeats
  - ▶ Win a dollar, take a moment!

$$\begin{aligned} P(D) = & P(D|\{\text{lose the dollar}\})P(\{\text{lose the dollar}\}) \\ & + P(D|\{\text{keep the dollar}\})P(\{\text{keep the dollar}\}) \\ & + P(D|\{\text{win the dollar}\})P(\{\text{win the dollar}\}). \end{aligned}$$

$$P(D) = \frac{1}{3} + \frac{1}{3}P(D) + \frac{1}{3}P(D)^2.$$

Solving for  $P(D)$  yields  $P(D) = 1$ , so  $P(NEG) = 0$ .

Sometimes, it is easier to work with conditional probability.

Conditioning as a problem-solving tool  
○○○

Pitfalls and Paradoxes: I  
●○○○○○

Pitfalls and Paradoxes: II  
○○○

Pitfalls and Paradoxes: III  
○○○○○

# Pitfalls and Paradoxes: I

## Problem: the effect of studying econometrics on wages

You want to know the effect of studying econometrics on wages.

- ▶ MSc in Econometrics ( $E$ ,  $\{\text{MSc in Econometrics}\}$ );
- ▶ Two types of wages:  
high ( $H$ ,  $\{\text{earn high salary}\}$ ) and low ( $H^C$ ,  $\{\text{low salary}\}$ ).

Research shows:  $P(H) = \frac{6}{100}$ ,  $P(H|E) = \frac{60}{100}$ ,  $P(E) = \frac{5}{100}$ .

- ▶ So, should you study econometrics to get a high wage?

## Problem: econometrics leads to a high wage?

You want to know the effect of studying econometrics on wages.

- ▶ MSc in Econometrics ( $E$ ,  $\{\text{MSc in Econometrics}\}$ );
- ▶ Two types of wages:  
high ( $H$ ,  $\{\text{earn high salary}\}$ ) and low ( $H^C$ ,  $\{\text{low salary}\}$ ).

Research shows:  $P(H) = \frac{6}{100}$ ,  $P(H|E) = \frac{60}{100}$ ,  $P(E) = \frac{5}{100}$ .

Plot twist: there is an additional event (“lurking in the background”):  
you like mathematics ( $M$ ) or not ( $M^C$ ).

Conditioning as a problem-solving tool

○○○

Pitfalls and Paradoxes: I

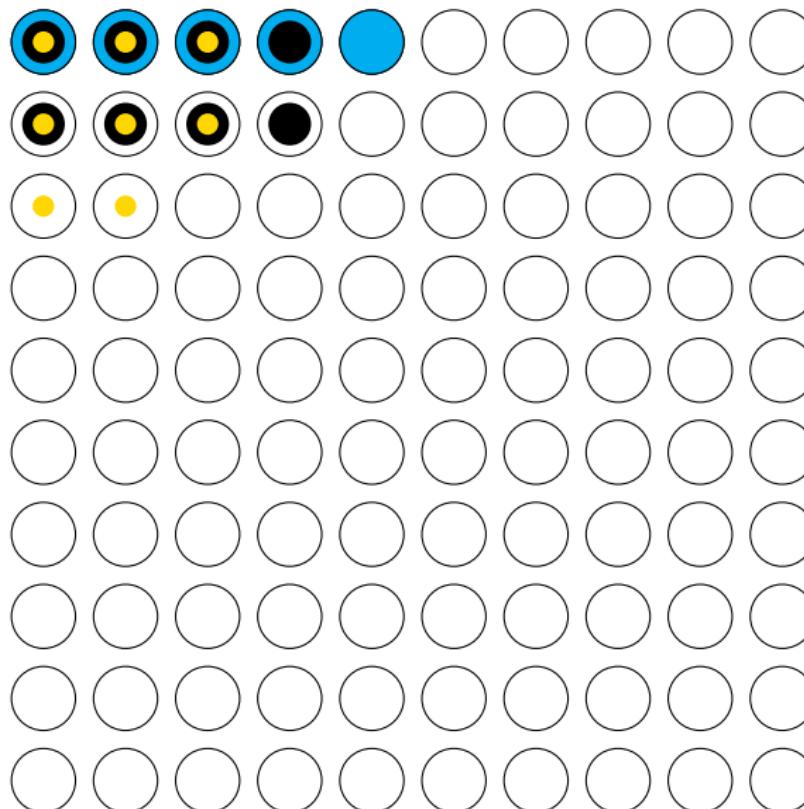
○○○●○○

Pitfalls and Paradoxes: II

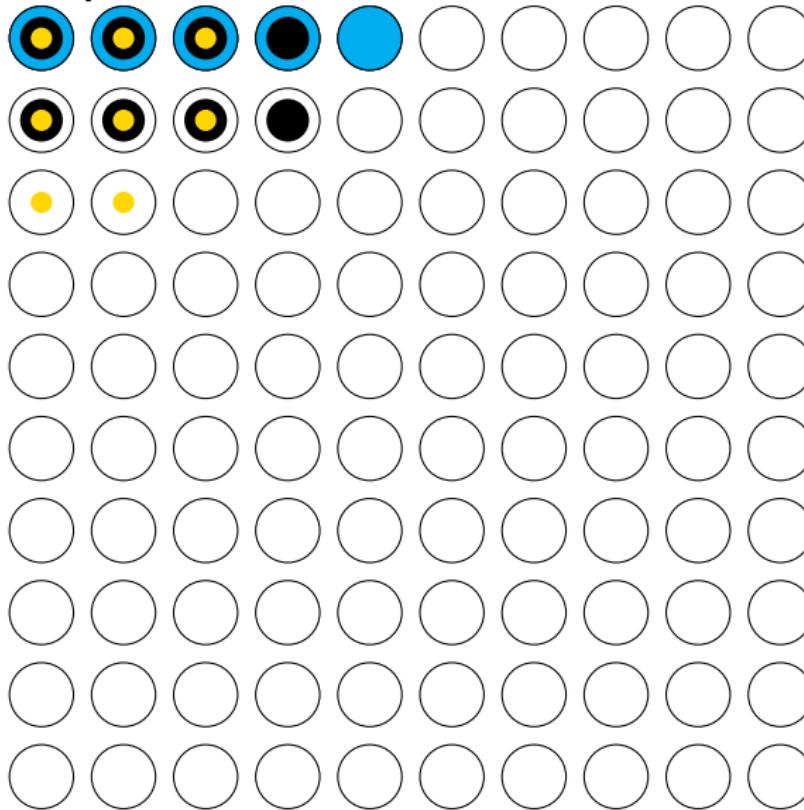
○○○

Pitfalls and Paradoxes: III

○○○○○



**Sample: 100 students: E, M, H.**

**Sample: 100 students**

$$P(H) = \frac{8}{100}$$

$$P(H|E) = \frac{3}{5}$$

$$P(H|M) = \frac{6}{8}$$

$$P(H|M, E) = \frac{3}{4}$$

$$P(H|M^C) = \frac{2}{92}$$

$$P(H|M^C, E) = 0$$

## In numbers

- ▶ High income:  $P(H) = \frac{8}{100}.$
- ▶ High income |  $E$ :  $P(H|E) = \frac{3}{5}$
- ▶ High income |  $M$ :  $P(H|M) = \frac{3}{4}.$
- ▶ High income |  $E$  and  $M$   $P(H|M, E) = \frac{3}{4}.$
- ▶ High income | not  $M$ :  $P(H|M^C) = \frac{2}{92}$
- ▶ High income |  $E$ , but not  $M$ ,  $P(H|M^C, E) = 0$

Lessons:

- ▶  $H$  is not independent of  $E$ .
- ▶  **$H$  is independent of  $E$  given  $M$ . Conditional independence!**
- ▶  $H$  is not independent of  $E$  given  $M^C$ .
  
- ▶ *What fits is best.*
  - ▶ If  $M$ , it doesn't matter whether you  $E$ .
  - ▶ If  $M^C$ , you are better off  $E^C$ .

Conditioning as a problem-solving tool  
○○○

Pitfalls and Paradoxes: I  
○○○○○

Pitfalls and Paradoxes: II  
●○○

Pitfalls and Paradoxes: III  
○○○○○

## Pitfalls and Paradoxes: II

## Prosecutor's fallacy

*In 1998, Sally Clark was tried for murder after two of her sons died shortly after birth. During the trial, an expert witness for the prosecution testified that the probability of a newborn dying of sudden infant death syndrome (SIDS) was 1/8500, so the probability of two deaths due to SIDS in one family was  $(1/8500)^2$ , or about one in 73 million. Therefore, he continued, the probability of Clark's innocence was one in 73 million.*

What do you think of this line of reasoning?

# $P(A|B) \neq P(B|A)$ !

- The expert witness says

$$P(D_1|I) \approx \frac{1}{8500}, \quad P(D_1 \cap D_2|I) \approx \frac{1}{8500^2}$$

- Then he states that *therefore*

$$P(I|D_1 \cap D_2) \approx \frac{1}{8500^2}$$

- Bayes says no!

$$P(I|D_1 \cap D_2) = \frac{P(D_1 \cap D_2|I)P(I)}{P(D_1 \cap D_2|I)P(I) + P(D_1 \cap D_2|I^C)P(I^C)}$$

If *Presumption of innocence* (namely,  $P(I^C) \approx 0$ ), then  
 $P(I|D_1 \cap D_2) \approx 1$ .

Conditioning as a problem-solving tool  
○○○

Pitfalls and Paradoxes: I  
○○○○○

Pitfalls and Paradoxes: II  
○○○

Pitfalls and Paradoxes: III  
●○○○○

## Pitfalls and Paradoxes: III

## Simpson's paradox

You have a job that consists of two tasks. You know that at both tasks, you are outperforming your colleague Jim, who is sleeping half of the time. At one point, your boss calls you in and says:

*"Bert, you're not doing well. Even Jim, who's sleeping half of the time, scores better than you. We have to let you go."*

What happened?

Job Performance	You (Bert)		Jim	
	Task 1	Task 2	Task 1	Task 2
Fails	40	1	10	18
Successes	80	22	15	100

Denote by

- ▶  $S$  the event that a task is a success,
- ▶  $T_i$  be the event that the task is task  $i$ , with  $i = \{1, 2\}$ ,
- ▶  $B$  the event that you are performing the task,  
 $B^C$  is the event that Jim performs the task.

We see that the probability that you succeed at Task 1 and Task 2 is higher than for Jim, i.e.

$$P(S|T_1, B) \approx \frac{80}{120} > \frac{15}{25} \approx P(S|T_1, B^C)$$

$$P(S|T_2, B) \approx \frac{22}{23} > \frac{100}{118} \approx P(S|T_2, B^C)$$

# Why you lost your job

Job Performance	You (Bert)		Jim	
	Task 1	Task 2	Task 1	Task 2
Fails	40	1	10	18
Successes	80	22	15	100

Aggregating across the two tasks however

$$P(S|B) \approx \frac{102}{143} < \frac{115}{143} \approx P(S|B^C)$$

Using the LOTP

$$P(S|B) = P(S|T_1, B)P(T_1|B) + P(S|T_2, B)P(T_2|B)$$

$$P(S|B^C) = P(S|T_1, B^C)P(T_1|B^C) + P(S|T_2, B^C)P(T_2|B^C)$$

- ▶ Be careful with conditional probabilities.
  - I. Econometrics → high wage: omitted conditioning variable.  
(Microeconomics: endogeneity issue.)  
Careful, conditional probability v.s. causality.
  - II. Mistaking  $A$  conditional on  $B$  with  $B$  conditional on  $A$
  - III. Averaging over events that are not comparable.

Layman's talk : what is random variable

Random variables are functions

●○○○○

Random variables are functions

## Introduce random variables via a simple example

Which one is easier to work with in a coin-flipping game?

- ▶  $A_{1i} = \{\text{ith throw gets a Head}\}$ ,  $A_{2i} = \{\text{ith throw gets a Bottom}\}$ ;
- ▶  $X_i = 1$  if  $i$ th throw gets a Head, and  $X_i = 0$  if  $i$ th throw gets a Bottom.

With the second notation, we can do more things in easier ways. E.g., think about what would happen if we let  $n$  increase in the following formula

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

this average is more likely to be closer and closer to  $\frac{1}{2}$  as  $n \rightarrow \infty$ .

**Now we can work with numbers!**

## Introduce random variables via a simple example

$$X_i(s) = \begin{cases} 1; & s \in A_{1i} = \{ \textit{i-th throw gets a Head} \} \\ 0; & s \in A_{2i} = \{ \textit{i-th throw gets a Bottom} \} \end{cases}$$

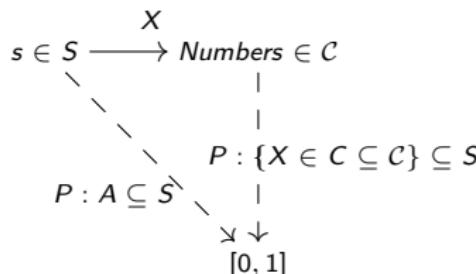
Random variables are functions!

# Introduce random variables via a simple example

$$X_i(s) = \begin{cases} 1; & s \in A_{1i} = \{ \text{ith throw gets a Head} \} \\ 0; & s \in A_{2i} = \{ \text{ith throw gets a Bottom} \} \end{cases}$$

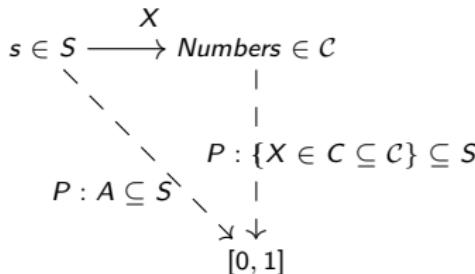
- ▶ Let  $\{X_i = 1\}$  denote the subset of the sample space  $S$  such that for  $s \in \{X_i = 1\}$ ,  $X_i(s) = 1$ .
- ▶  $P(X_i = 1) = P(A_{1i}) = 1/2$ .

**$X$  maps elements from  $S$  to numbers! Now we can work with functions/numbers!**



- ▶  $X$  essentially is a function, and when we manipulate random variables, it would be similar to manipulating functions and there are so many transformations we can work with: summation, division, max, min...

$X$  maps elements from  $S$  to numbers! Now we can work with functions/numbers!



$X$  essentially is a function, and when we manipulate random variables, it would be similar to manipulating functions and there are so many transformations we can work with: summation, division, max, min...

- ▶ We can now use  $\{X \in \mathcal{C}\}$  with  $\mathcal{C} \subseteq \mathcal{C}$  to represent different random events, and properties derived for random events would be inherited.
- ▶ Given the introduction of these additional structures, we have a larger space to explore: conditional probability/independence based on random variables, random variables with specific probability functions...

**Random variables are functions, there are selected functions with nice properties, and so are some selected random variables.**

# Probability Theory for EOR

## Random variables (discrete)

Random variables are functions

●○○○○○○○

Discrete random variable

○○○

# Random variables are functions

## Definition (Random variables (r.v.'s) (preliminary))

A random variable (r.v.) is a function maps elements in sample space  $S$  to numbers in  $\mathcal{C}$ , e.g.,  $\mathcal{C} \subseteq \mathbb{R}$ .

**X maps elements from  $S$  to numbers! Now we can work with functions/numbers!**

Usually, we consider **real-valued r.v.'s**, i.e.,

$$X : S \mapsto \mathbb{R}$$

- ▶  $X$  essentially is a function, and when we manipulate random variables, it would be similar to manipulating functions and there are so many transformations we can work with: summation, division, max, min...
- ▶ **There are many functions from  $S$  to  $\mathbb{R}$  (so there are also many random variables defined with the same  $S$ ), not all functions can be regarded as random variables, but all random variables are functions.** Detailed discussion goes beyond the course scope.
- ▶ We can now use  $\{X \in C\}$  with  $C \subseteq \mathcal{C}$  to represent different random events, and properties derived for random events would be inherited. **This would connect random variables with probability functions** as probability functions have random events as elements in its domain.
- ▶ Given the introduction of these additional structures, we have a much larger place for our mind adventures: conditional probability/independence based on random variables, random variables with specific probability functions... **Random variables are functions, there are selected functions with nice properties, and so are some selected random variables.**

Random variables are functions

○○●○○○○

Discrete random variable

○○○

There are many functions from a sample space to a set of numbers, and we start with those map to sets of a sequence of numbers  $a_1, a_2, a_3, \dots$ .

## Simple example: throw a coin 10 times

- ▶ The outcomes are strings like  $HHTHTHHTHT$ .
- ▶ Notations:
  - ▶ define the event  $H_i$ : we see  $i$  heads.
  - ▶ define the event  $T_j$ : we see  $j$  tails.
  - ▶  $H_i = T_{n-i}$ .
- ▶ **Cumbersome!**

## Simple example: throw a coin 10 times

- ▶ Let  $X$  be the number of heads and  $Y$  be the number of tails.
- ▶ Improves notation: no subindices as with  $H_i$ ,  $T_j$  or  $H_i = T_{n-i}$ .
- ▶ Improves calculation: we see  $Y = n - X$ .

But, what are these  $X$  and  $Y$ ?

## Simple example: throw a coin 10 times

Throw a coin  $n$  times. Define  $X$  to be the number of heads.

- $X$  takes an outcome  $s$  in the sample space  $S$ , and maps that to a number on the real line.

$$X(HHTHTHHTHT) = 6,$$

$$X(HHTHTHHTTH) = 6,$$

$$X(TTTTTTTTTT) = 0,$$

⋮

- $X$  is a *function* that maps each  $s \in S \rightarrow \mathbb{R}$ .
- Because the outcome is random,  $X$  is called a *random variable* (r.v.). But it is simply a function.

## Simple example: throw a coin 10 times

- ▶  $X$  needs a set of outcomes. An r.v. is defined *on* a sample space  $S$ .
- ▶ We can define *multiple* r.v.'s on  $S$ :
  - ▶  $X$ : number of heads;
  - ▶  $Y$ : number of tails;
  - ▶  $Z$ : only equal to one if the last throw is a head.
- ▶  $X$  assigns a number to each  $s \in S$ .

## Simple example: throw a coin 10 times

- ▶ Furthermore, from  $X$  we could define events, e.g., let  $\{X \in \{6\}\}$  (or  $\{X = 6\}$ ) denote the set of outcomes  $s \in S$  such that  $X(s) = 6$ , and it would be the event that we get 6 heads in 10 throws.
- ▶  $P(X = 6) = P(\{HHTHTHHHTT\} \cup \{HHTHTHHHTT\} \cup \dots) = \frac{\binom{10}{6}}{2^{10}} \approx 0.205078$ .
- ▶ With the defined events, it is easy to see that

$$P(X \in \{0, 1, 2, \dots, 10\}) = 1$$

- ▶ We have a **discrete random variable!**

Random variables are functions

○○○○○○○○

Discrete random variable

●○○

## Discrete random variable

## Definition (Discrete random variable)

A random variable  $X$  is a **discrete** random variable if there exists a set  $C$  with at most countably many numbers (so you can index all the elements within  $C$  using a set of natural numbers):

$$P(\{X \in C\}) = 1, \text{ or equivalently, } P(X \in C) = 1.$$

The **support** of a **discrete** random variable is the set of all numbers,  $x$ 's, such that  $P(X = x) > 0$ :

$$\text{supp}(X) = \{x \in C : P(X = x) > 0\}.$$

## Previous simple example revisits: throw a coin 10 times

- ▶ Previously:  $X$  takes an outcome  $s$  in the sample space  $S$ , and maps that to a number on the real line.

$$X(HHTHTHHTHT) = 6,$$

$$X(HHTHTHHTTH) = 6,$$

$$X(TTTTTTTTTT) = 0,$$

⋮

$X$  is a **real-valued** r.v.!

- ▶  $C = \{x_1, x_2, \dots\} = \{x_i, i \in \mathbb{N}_+\}$  such that  $x_i = i - 1$ :

$$P(X \in C) = 1.$$

So,  $X$  is a **discrete** r.v.!

(The choice of such  $C$  is not unique.)

- ▶

$$\text{supp}(X) = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

# Probability Theory for EOR

## Random variables and their distributions

Distributions: probability values of events associated with random variables



Distributions: probability values of events associated  
with random variables

**Real-valued r.v.  $X$  maps elements from  $S$  to real numbers!**

Now we can work with functions/numbers:

$$X : S \mapsto \mathbb{R}$$

- ▶ We can now use  $\{X \in C\}$  with some subsets  $C \subseteq \mathbb{R}$  to represent different random events (not all subsets  $C$  generate well-defined events, and usually we focus on different intervals: e.g.,  $(-\infty, a]$ ,  $[a, b] \cap (a, e) \dots$ ).

**The distribution of a given r.v. needs to specify all probability values of all those events generated by the r.v.:**

$$P : \{\{X \in C\} : \text{some subsets } C \subseteq \mathbb{R}\} \mapsto [0, 1].$$

For example, a distribution needs to be able to answer  $P(X \in [a, b] \cup [c, d])$ ,  $P(X = e)$ , ...

- ▶ Though there are so many associated events, **we only need to know probability values of some selected events** and the rest probability values can be inferred from these values.

## Definition (CDF of real-valued r.v.)

The **cumulative distribution function (CDF)** of an **real-valued r.v.**,  $X : S \rightarrow \mathbb{R}$ , is the function, usually denoted by  $F_X$ , which maps numbers to numbers within  $[0,1]$ .

$$F_X(x) = P(X \leq x).$$

- We only need to know probability values of the following types of events:  $\{X \leq x\}, x \in \mathbb{R}$ .

### Properties:

- **Increasing function from zero to one:**

$$0 \leq F_X(x_1) \leq F_X(x_2) \leq 1, \forall -\infty < x_1 \leq x_2 < \infty.$$

$$(a): \lim_{x \rightarrow -\infty} F_X(x) = 0; \text{ think about } \emptyset!$$

$$(b): \lim_{x \rightarrow \infty} F_X(x) = 1; \text{ think about } S!$$

- **Right-continuous:**  $F_X(a) = \lim_{x \downarrow a} F(x)$ .

## Simple example revisits: throw a coin 10 times

- ▶ Let  $X$  be the r.v. that returns the number of heads.
- ▶ If we know  $F_X$ , it is easy to calculate, e.g.,  
 $P(s \in S : X(s) = 6) = P(X = 6) =$
- ▶  $= F_X(6) - F_X(5)!$
- ▶ **However, for this special case, there is easier way as  $X$  at most takes 11 values (values in the support of  $X$ ), we only need to know**

$$P(X = i), i = 0, 1, \dots, 10.$$

## From CDF to PMF

### Definition (CDF of real-valued r.v.)

The **cumulative distribution function (CDF)** of an **real-valued** r.v.,  $X : S \rightarrow \mathbb{R}$ , is the function, usually denoted by  $F_X$ :

$$F_X(x) = P(X \leq x).$$

### Definition (PMF of discrete r.v.)

The **probability mass function (PMF)** of a **discrete** r.v.,  $X : S \mapsto \{x_1, x_2, \dots\}$ , is the function:

$$p_X(x) = P(X = x); x \in \text{Supp}(X).$$

## Definition (PMF of discrete r.v.)

The **probability mass function (PMF)** of a **discrete r.v.**,  $X : S \mapsto \{x_1, x_2, \dots\}$ , is the function:

$$p_X(x) = P(X = x).$$

- We only need to know probability values of the following types of events:  $\{X = x\}, x \in \text{supp}(X)$ .

### Properties:

- **Non-negativity:**  $p_X(x) > 0$  if  $x \in \text{Supp}(X)$ ;  $p_X(x) = 0$  otherwise.
- **Sum to 1.** Suppose  $X$  has support  $x_1, x_2, \dots$ :  $\sum_{j=1}^{\infty} p_X(x_j) = 1$ .

Distributions: probability values of events associated with random variables

○○○○○●○

**Discrete r.v.'s and their distributions**

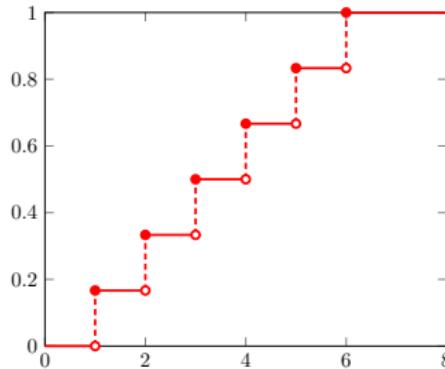
The distribution tell us almost all we need to know about a random variable.

## Another simple example: Throw a fair six-sided die

What is the PMF for the random variable  $X$  denotes the number of eyes of the outcome if we throw a six-sided fair die for one time?

$$P(X = i) = 1/6, i = 1, 2, 3, 4, 5, 6.$$

From the PMF, we can also derive CDF, let's draw a graph:



Throw a fair six-sided die twice. What is the PMF of,  $Y$ , the sum of the

# Probability Theory for EOR

## Some special discrete random variables

Some **discrete random variables** are associated very special/ubiquitous **distributions (PMFs)**, they get their own names!

## Definition

The **probability mass function (PMF)** of a **discrete r.v.**,  $X : S \mapsto \{x_1, x_2, \dots\}$ , is the function:

$$p_X(x) = P(X = x).$$

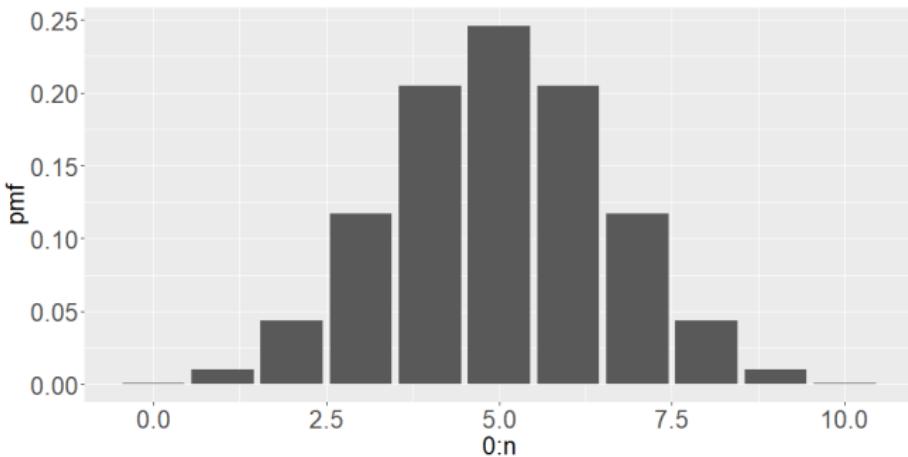
## Binomial: $\text{Bin}(n, p)$

- ▶ Throw a special coin  $n$  times (with probability  $p$  getting head in each flip), let  $X$  denotes the number of heads in  $n$  throws.
- ▶  $X \sim \text{Bin}(n, p)$ .
- ▶ The PMF of  $X$  is

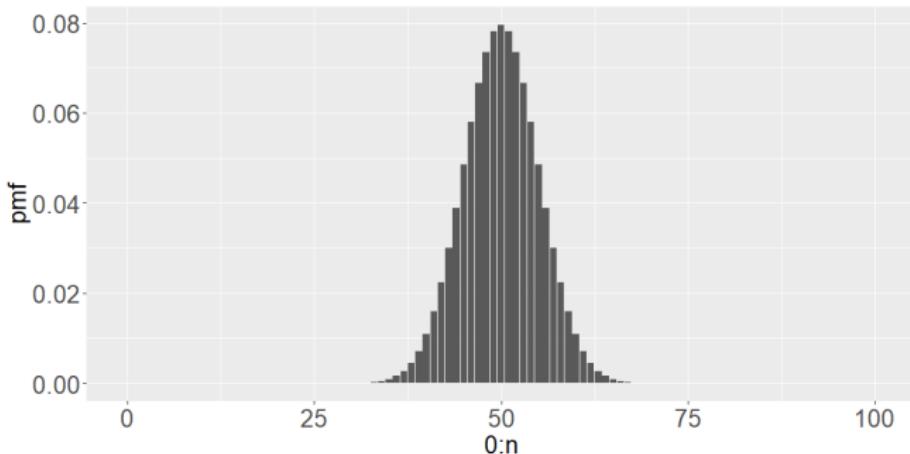
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}; k = 0, 1, \dots, n.$$

- ▶  $\text{Bin}(1, p)$  is also called Bernolli distribution  $\text{Bern}(p)$ .

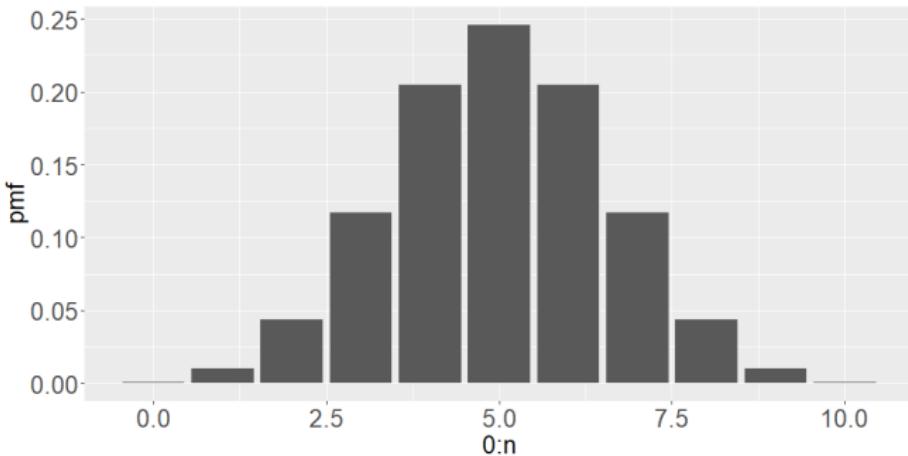
## Some Binomial PMFs in R ( $n = 10$ , $p = 0.5$ )



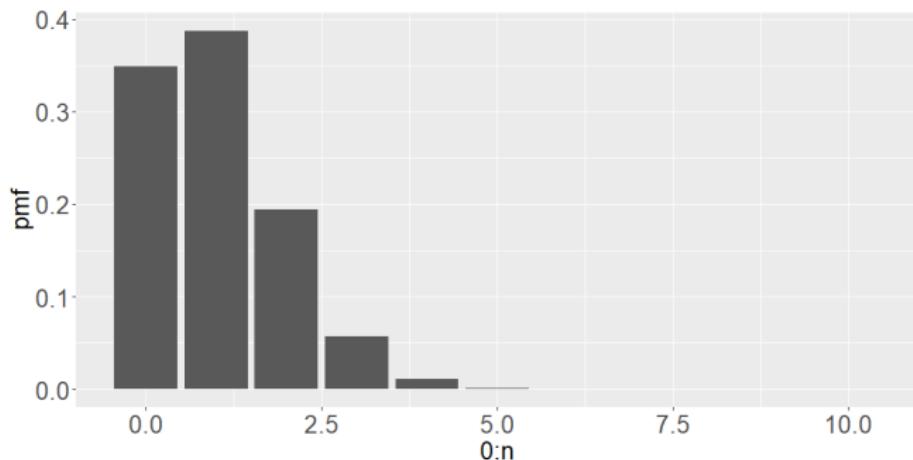
## Some Binomial PMFs in R ( $n = 100$ , $p = 0.5$ )



## Some Binomial PMFs in R ( $n = 10$ , $p = 0.5$ )



## Some Binomial PMFs in R ( $n = 10$ , $p = 0.1$ )



## Discrete uniform: $\text{Dunif}(C)$

- ▶  $C$  contains  $n$  different numbers ( $|C| = n$ ), Throw a  $n$ -sided fair die and each side is marked with a different number in  $C$ . Denote  $X$  the number of one die roll, and  $\{X = x\}$  for any  $x \in C$  are equally likely to occur.
- ▶  $X \sim \text{Dunif}(C)$ .
- ▶ The PMF of  $X$  is

$$P(X = x) = 1/n; x \in C.$$

- ▶ What is  $P(X \in \{x_1, x_2\})$ ,  $x_1 \neq x_2, x_1, x_2 \in C$ ?  
 $2/n$ .
- ▶  $\text{Bin}(1, 0.5)$  also a  $\text{Dunif}(\{0, 1\})$ .

## Hypogeometric: $\text{HGeom}(w, b, n)$

- ▶ Mark  $n$  different balls with \*'s from a urn of  $w$  white balls and  $b$  **black** balls (each ball is equally likely to be marked), and  $X$  is the number of balls being both **white** and marked with \*'s.
- ▶  $X \sim \text{HGeom}(w, b, n)$ .
- ▶ The PMF of  $X$  is

$$P(X = k) = \frac{\binom{w}{k} \binom{b}{n-k}}{\binom{w+b}{n}}; \max\{n - b, 0\} \leq k \leq \min\{n, w\}.$$

- ▶  $X \sim \text{HGeom}(w, b, n)$  and  $Y \sim \text{HGeom}(n, w + b - n, w)$  have the identical distribution:

$$\frac{\binom{w}{k} \binom{b}{n-k}}{\binom{w+b}{n}} = \frac{\binom{n}{k} \binom{w+b-n}{w-k}}{\binom{w+b}{w}}$$

Mark  $w$  different balls with **white** from a urn of  $n$  balls with \*'s and  $w + b - n$  balls without \*'s (each ball is equally likely to be marked), and  $Y$  is the number of balls being both **white** and marked with \*'s.

## Probability Theory for EOR

Independence of random variables (two random variables)

**Real-valued r.v.'s  $X, Y$ :**

$$X, Y : S \mapsto \mathbb{R}$$

- ▶ It is natural to think that two r.v.'s are independent if knowing the events generated by one random variable does not change the distribution of another one.
- ▶ We know:

- (I) **the probability values of events of type  $\{X \leq x\}, x \in \mathbb{R}$  would determine the whole distribution of real-valued r.v.  $X$  ;**
- (II) **the independence of two random events  $P(A \cap B) = P(A)P(B)$ ;**
- (III) **look at events  $\{X \leq x\}$  and  $\{Y \leq y\}, x, y \in \mathbb{R}$  :**

$$P(\{X \leq x\} \cap \{Y \leq y\}) = P(\{X \leq x\})P(\{Y \leq y\}).$$

**Sometimes, we use  $P(X \leq x, Y \leq y)$  to denote  $P(\{X \leq x\} \cap \{Y \leq y\})$ .**

Real-valued r.v.'s  $X, Y$  are independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y); \forall x, y \in \mathbb{R}.$$

- **Independence means products.**
- There are many events of the type  $\{X \leq x\}$  associated with real-valued r.v.'s, and independence would imply products involve all these events.

## Simple example: Throw a coin twice

Denote  $(X_1, X_2)$  as the outcome of the two flips and  $X_i = 1, i = 1, 2$  if  $i$ th throw get a head other wise zero.

- $S = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ . Finite equally likely outcomes.
- Easy to check that:

$$P(X_1 \leq x, X_2 \leq y) = P(X_1 \leq x)P(X_2 \leq y); \forall x, y \in \mathbb{R}.$$

- And indeed, e.g.,

$$P(X_1 \leq x | X_2 \leq y) = P(X_1 \leq x), \forall x \in \mathbb{R}, y \geq 0.$$

**Real-valued** r.v.'s  $X, Y$  are independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y); \forall x, y \in \mathbb{R}.$$

**Discrete** r.v.'s  $X, Y$  are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y); \forall x \in \text{supp}(X), y \in \text{supp}(Y).$$

## Simple example revisits: Throw a coin twice

Denote  $(X_1, X_2)$  as the outcome of the two flips and  $X_i = 1, i = 1, 2$  if  $i$ th throw get a head other wise zero.

- $S = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ . Finite equally likely outcomes.
- Easy to check that:

$$P(X_1 = x, X_2 = y) = P(X_1 = x)P(X_2 = y); \forall x, y \in \{0, 1\}.$$

- And indeed, e.g.,

$$P(X_1 \leq x | X_2 \leq y) = P(X_1 \leq x), \forall x \in \mathbb{R}, y \geq 0.$$

**Real-valued** r.v.'s  $X, Y$  are independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y); \forall x, y \in \mathbb{R}.$$

**Discrete** r.v.'s  $X, Y$  are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y); \forall x \in \text{supp}(X), y \in \text{supp}(Y).$$

**Real-valued** r.v.'s  $X, Y$  are **conditionally** independent given a discrete r.v.  $Z$  if

$$P(X \leq x, Y \leq y | Z = z) = P(X \leq x | Z = z)P(Y \leq y | Z = z); \\ \forall x, y \in \mathbb{R}, z \in \text{supp}(Z).$$

**Discrete** r.v.'s  $X, Y$  are are **conditionally** independent given a discrete r.v.  $Z$  if

$$P(X = x, Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z); \\ \forall x \in \text{supp}(X), y \in \text{supp}(Y), z \in \text{supp}(Z).$$

Similar to independence among events, these concepts can also be generalized to  $n$  r.v.'s.

Conditional independence  $\not\Rightarrow$  independence

## Example: mystery opponent

Conditional independence  $\not\rightarrow$  independence

- You play against one of two identical twins for two rounds, flip a coin to determine which game you play:

**Game A.** Against the first twin ( $A$ ) you are evenly matched for two rounds.

**Game B.** Against the other you win with probability  $3/4$  for two rounds.

- $X$  and  $Y$ : outcome of rounds 1 and 2. No hot hands.
- $Z$  equal to one if you play against ( $A$ ) otherwise  $Z$  equals 0.
- Conditional on  $Z = 1$ ,  $X$  and  $Y$  are i.i.d.  $\text{Bern}(1/2)$ .
- Conditional on  $Z = 0$ ,  $X$  and  $Y$  are i.i.d.  $\text{Bern}(3/4)$ .
- Without  $Z$ ,  $P(Y = 1|X = 1) > P(Y = 1)$ .

$$P(Y = 1) = 1/2 \times 1/2 + 1/2 \times 3/4 = 5/8 = 0.625.$$

$$P(Y = 1|X = 1) = \frac{P(Y = 1, X = 1)}{P(Y = 1)} = \frac{1/2 \times (1/2)^2 + 1/2 \times (3/4)^2}{5/8} = 13/20 = 0.65.$$

Note here: by definition we should have  $P(Y = 1|X = 1) = \frac{P(Y=1, X=1)}{P(X=1)}$ , In this example, this number (probability value of a given event),  $P(\textcolor{red}{X} = 1)$ , happens to be equal to the probability value  $P(Y = 1)$ , these two numbers are equal here.

## Example: coin flip

Conditional independence  $\not\leftarrow$  independence

- ▶ Cook up an example!
- ▶ Flip two coins:  $X_i = I_{A_i}, i = 1, 2, A_i = \{i\text{th flip is a head}\}$ .
- ▶  $X_1, X_2$  are independent.
- ▶ Denote  $Z = X_1 + X_2$  then  $X_1, X_2$  are not conditionally independent given  $Z$ .

# Probability Theory for EOR

## Functions of random variables

# Functions of random variables

Random variables are functions.

**$X$  maps elements from  $S$  to numbers! Now we can work with functions/numbers!**: e.g.,

$$X : S \mapsto \mathbb{R}$$

There are transformations on **functions** (composite functions), e.g., from  $f, g$  to get **new functions**:

$$(f)^2, (f)^3, f + 1, fg, f + g, \max\{f, g, 0\} \dots$$

We can manipulate **random variables** in a similar way:  
 $(X)^2, (X)^3, X + 1, XY, X + Y, \max\{X, Y, 0\} \dots$

## Functions of a real-valued r.v.'s

For an r.v.  $X : S \mapsto \mathbb{R}$  with sample space  $S$ , and a proper function  $g : \mathbb{R} \rightarrow \mathbb{R}$ ,  $g(X)$  is the r.v. that maps  $s$  to  $g(X(s)) \in \mathbb{R}$  for all  $s \in S : S \mapsto \mathbb{R}$ .

**Example:**



$X \sim \text{Bern}(p)$ , then  $Y = 2(X + 1)$  is also an r.v. with PMF:

$$P(Y = 2) = 1 - p; P(Y = 4) = p.$$

For a discrete r.v.  $X$ , the PMF of  $Y = g(X)$ :

$$P(Y = y) = \sum_{x: g(x)=y} P(X = x); y \in \{g(x) : x \in \text{supp}(X)\}.$$

## Functions of multi real-valued r.v.'s

For multiple r.v.'s  $X_i : S \mapsto \mathbb{R}, i = 1, 2, 3 \dots, n$  with sample space  $S$ , and a proper function  $g : \mathbb{R}^n \mapsto \mathbb{R}$ ,  $g(X_1, X_2, \dots, X_n)$  is the r.v. that maps  $s$  to  $g(X_1(s), X_2(s), \dots, X_n(s)) \in \mathbb{R}$  for all  $s \in S : S \mapsto \mathbb{R}$ .

**Example:**

$X_1, X_2 \sim \text{Bern}(0.5)$  and they are independent, then  $Y = X_1 + X_2$  is also an r.v. with PMF:

$$P(Y = 0) = 1/4; P(Y = 1) = 1/2; P(Y = 2) = 1/4.$$

## Practice and some new results

### Throw a fair six-sided die

What is the PMF for the random variable  $X$  denotes the number of eyes of the outcome if we throw a six-sided fair die for one time?

$$P(X = i) = 1/6, i = 1, 2, 3, 4, 5, 6.$$

**Throw a fair six-sided die twice.** What is the PMF of,  $Y$ , the sum of the two **independent** die rolls?

$$P(Y = i), i = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.$$

Denote  $X_1$  the number of eyes of the outcome in the first throw, and  $X_2$  the number of eyes of the outcome in the second throw ( $Y = X_1 + X_2$ ):

$$\begin{aligned} P(Y = 3) &= \sum_{x_1, x_2 : x_1 + x_2 = 3} P(X_1 = x_1, X_2 = x_2) \\ &= \sum_{x_1, x_2 : x_1 + x_2 = 3} P(X_1 = x_1)P(X_2 = x_2) \\ &= P(X_1 = 1)P(X_2 = 2) + P(X_1 = 2)P(X_2 = 1) \\ &= 1/6 \times 1/6 + 1/6 \times 1/6 = 1/18. \end{aligned}$$

**Throw a fair six-sided die three times.** Denote by  $X, Y, Z$  the eyes by the first, second and third die. Suppose  $X, Y$  and  $Z$  are **independent identically distributed (i.i.d.)**. Calculate  $P(\max(X, Y, Z) > 5)$ :

## Results: Binomial and Hypergeometric

- $X \sim \text{Bin}(n, p)$ ,  $Y \sim \text{Bin}(m, p)$ .  $X$  and  $Y$  are independent:
  - $Z = X + Y$ ,  $Z \sim \text{Bin}(n + m, p)$ .
  - The conditional distribution of  $X$  given  $Z = r$  is  $\text{HGeom}(n, m, r)$ :

$$P(X = k | Z = r) = \frac{\binom{n}{k} \binom{m}{r-k}}{\binom{n+m}{r}}$$

- $X \sim \text{HGeom}(w, b, n)$ , if  $N = w + b \rightarrow \infty$  and  $p = \frac{w}{w+b}$ :

The PMF of  $X$  converges to the  $\text{Bin}(n, p)$  PMF:

$$P(X = k) = \binom{n}{k} \frac{\prod_{i=0}^{k-1} (p - \frac{i}{N}) \prod_{j=0}^{n-k-1} (q - \frac{j}{N})}{\prod_{l=1}^{n-1} (1 - \frac{l}{N})} \rightarrow \binom{n}{k} p^k q^{n-k}$$

Expectation summarizes "average" value:, a function mapping well-behaved random variables to numbers  
ooooo

## Layman's talk : what is expectation

Expectation summarizes "average" value:, a function mapping well-behaved random variables to numbers

●○○○○

Expectation summarizes "average" value:  
a function mapping well-behaved random variables  
to numbers

- ▶ Distributions assign (probability values) numbers to random events associated with random variables.  
*Many values associated with many events...*
- ▶ Sometimes, one would prefer to know what happens "on average".  
*E.g., if heads for 1 bottoms for 0, we know from our daily experience, the average would be 0.5.*
- ▶ **Expectation** formalizes the above idea and assigns a number (if exists) for one random variable.

○○●○○

**Example:**



$X \sim \text{Bern}(p)$  with PMF:

$$P(X = 0) = 1 - p; P(X = 1) = p.$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = p.$$

**Example:**



$X \sim \text{Bern}(p)$ , then  $Y = 2 + 2X$  is also an r.v. with PMF:

$$P(Y = 2) = 1 - p; P(Y = 4) = p.$$

$$E(X) = 0 \times P(X = 0) + 1 \times P(X = 1) = p.$$

$$E(Y) = 2 \times P(Y = 2) + 4 \times P(Y = 4) = 2 + 2p.$$

**For a discrete r.v.  $X$ :**

$$E(X) = \sum_{x \in \text{supp}(X)} xP(X = x).$$

**For a discrete r.v.  $X$ :**

$$E(X) = \sum_{x \in \text{supp}(X)} xP(X = x).$$

- ▶ A weighted average (with weights being the associated probability values, these weights sum up to 1) of possible values taken by  $X$ .
- ▶ For  $E(X)$  to be well-defined, we need the condition that this summation  $\sum_{x \in \text{supp}(X)} xP(X = x)$  is well-defined.

- ▶ Expectation essentially maps a subset of the collection of all random variables to numbers.  
It is a function!

$$E : \{X : \text{well-behaved r.v.'s}\} \mapsto \mathcal{C}$$

$E(X + Y) = E(X) + E(Y)$  and  $E(aX) = aE(X)$  for real-valued r.v.'s  $X, Y$  and  $a \in \mathbb{R}$ .

- ▶ **Expectation goes far beyond "average"** by working with different transformation of random variables.

E.g., it can recover probability values via the indicator random variable  $I_{\{X=x\}}$ :

$$P(X = x) = EI_{\{X=x\}}.$$

# Probability Theory for EOR

## Expectation of discrete random variables

Expected values of r.v.'s (if exist) are numbers.

## Expectation of discrete r.v.'s

## Definition (Expectation of discrete r.v.'s)

The expectation/expected value/first moment/mean/average of a **discrete** random variable  $X$  for  $x_1, x_2, x_3, \dots$  (if exists) is defined by

$$E[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i) \quad (= \sum_{i=1}^{\infty} x_i \Delta F_X(x_i))$$

(we order all values from the support of  $X$  by the increasing order:  $x_1 \leq x_2 \leq x_3, \dots$ .

We may choose  $\Delta F_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ , let  $x_0 < x_1$ ,  $x_0 \notin \text{supp}(X)$  and thus  $F(x_0) = 0$ .

We may choose  $\Delta F_X(x) = F(x) - \lim_{y \uparrow x} F(y)$ , as another possible expression.

Essentially both expressions give:  $\Delta F_X(x_i) = P(X = x_i)$ .

If  $|\text{supp}(X)|$  is finite with elements  $x_1, x_2, x_3, \dots, x_n$ , then

$$E[X] = \sum_{i=1}^n x_i \mathbb{P}(X = x_i) \quad (= \sum_{i=1}^n x_i \Delta F_X(x_i))$$

- ▶ Expectation may not exist for some random variables, e.g., the summation  $\sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i)$  may diverge.
- ▶ Same distributions, same expectations (if exist).

## Example

A box of 10 apples has 3 bad apples. We choose 3 apples at random, without replacement. What is the number of bad apples that you expect?

Denote by  $X$  the number of bad apples. Then

$$P(X = 0) = \frac{\binom{7}{3} \binom{3}{0}}{\binom{10}{3}}, P(X = 1) = \frac{\binom{7}{2} \binom{3}{1}}{\binom{10}{3}}$$

$$P(X = 2) = \frac{\binom{7}{1} \binom{3}{2}}{\binom{10}{3}}, P(X = 3) = \frac{\binom{7}{0} \binom{3}{3}}{\binom{10}{3}}$$

$$\mathbb{E}[X] = P(X = 0) \cdot 0 + P(X = 1) \cdot 1 + P(X = 2) \cdot 2 + P(X = 3) \cdot 3 = 3 \times \frac{3}{10}.$$

# Probability Theory for EOR

Properties of expectation  
(proofs with discrete random variables)

*Expectation* maps a subset of the collection of random variables to numbers (expectations/expected values of associated random variables).

If we regard *expectation* as a special function from random variables to numbers, what are its properties?

Properties of expectation (proofs with discrete r.v.'s)

**Linearity.** Assume all expectations are well defined.

For any **real-valued** r.v.'s  $X, Y$  and any constant  $c$ ,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$$\mathbb{E}(cX) = c\mathbb{E}(X)$$

A linear function!

**Proof I with discrete r.v.'s:**

$$\begin{aligned}
\mathbb{E}(X + Y) &= \sum_c cP(X + Y = c) = \sum_c c \left( \sum_{a \in \text{supp}(X)} P(X = a, Y = c - a) \right) \\
&= \sum_{a \in \text{supp}(X)} \sum_c cP(X = a, Y = c - a) = \sum_{a \in \text{supp}(X)} \sum_c ((a) + (c - a))P(X = a, Y = c - a) \\
&= \sum_{a \in \text{supp}(X)} \sum_{b \in \text{supp}(Y)} (a + b)P(X = a, Y = b) \\
&= \sum_{a \in \text{supp}(X)} \sum_{b \in \text{supp}(Y)} aP(X = a, Y = b) + \sum_{a \in \text{supp}(X)} \sum_{b \in \text{supp}(Y)} bP(X = a, Y = b) \\
&= \sum_{a \in \text{supp}(X)} a \sum_{b \in \text{supp}(Y)} P(X = a, Y = b) + \sum_{b \in \text{supp}(Y)} b \sum_{a \in \text{supp}(X)} P(X = a, Y = b) \\
&= \sum_{a \in \text{supp}(X)} aP(X = a) + \sum_{b \in \text{supp}(Y)} bP(Y = b) \\
&= E(X) + E(Y)
\end{aligned}$$

**Proof II with discrete r.v.'s and finite outcomes**  $|S| < \infty$ :

$$\begin{aligned}
\mathbb{E}(X + Y) &= \sum_c c P(X + Y = c) \\
&= \sum_c c \sum_{s \in S: (X+Y)(s)=c} P(\{s\}) \\
&= \sum_c \sum_{s \in S: (X+Y)(s)=c} c P(\{s\}) \\
&= \sum_c \sum_{s \in S: (X+Y)(s)=c} (X(s) + Y(s)) P(\{s\}) \\
&= \sum_{s \in S} (X(s) + Y(s)) P(\{s\}) \\
&= \left( \sum_{s \in S} X(s) P(\{s\}) \right) + \left( \sum_{s \in S} Y(s) P(\{s\}) \right) \\
&= \left( \sum_{a \in \text{supp}(X)} \sum_{s \in S: X(s)=a} X(s) P(\{s\}) \right) + \left( \sum_{b \in \text{supp}(Y)} \sum_{s \in S: Y(s)=b} Y(s) P(\{s\}) \right) \\
&= \left( \sum_{a \in \text{supp}(X)} a P(X = a) \right) + \left( \sum_{b \in \text{supp}(Y)} b P(Y = b) \right) \\
&= E(X) + E(Y)
\end{aligned}$$

## Example

Throw a coin for one time ( $X$  denotes the number of heads), what is  $E(X)$ ?  $1/2$ .

Throw a coin for two time ( $Y$  denotes the number of heads), what is  $E(Y)$ ?  $1$ .

Throw a coin for  $n$  time ( $Z_n$  denotes the number of heads), what is  $E(Z_n)$ ?  $n/2$ .

**Monotonicity.** Assume all expectations are well defined.

For any **real-valued** r.v.'s  $X, Y$  such that  $X \geq Y$  with probability one ( $P(X \geq Y) = 1$ ), then

$$\mathbb{E}(X) \geq \mathbb{E}(Y)$$

with equality holding iff  $\mathbb{P}(X = Y) = 1$ .

A linear monotone function!

The results can be extended to more general cases, e.g., any **real-valued** r.v.'s  $\tilde{X}, \tilde{Y}$  have identical distributions with  $X, Y$  respectively such that  $X \geq Y$  with probability one, then  $E(\tilde{X}) \geq E(\tilde{Y})$ .

**Proof with discrete r.v.'s:**

Note that  $Z = X - Y$  would be non-negative with probability one, such that  $Z(s) \geq 0$  for all  $s \in S$ , and thus  $E(Z)$  is a weighted sum of non-negative values ( $\geq 0$ ), and by linearity

$$E(X) - E(Y) = E(Z) \geq 0.$$

If  $E(X) = E(Y)$ , then  $E(Z) = \sum_{z_i \in \text{supp}(Z)} z_i P(Z = z_i) = 0$ . Since  $z_i \geq 0$ ,  $P(Z = z_i) > 0$ , we know  $z_i = 0$  (otherwise  $E(Z) > 0$ ).

## Example

Throw a fair coin for two time.

$X_1$  denotes the number of head of the first throw,  $X_2$  denotes the number of head of the second throw,  $Z = X_1 + X_2$ .  $P(Z \geq X_1) = P(Z - X_1 \geq 0) = P(X_2 \geq 0) = 1$ .

$E(Z) \geq E(X_1)$  (**via monotonicity directly**).

$Y$  denotes the number of tails. Note that there is outcome  $(s)$  such that  $Y(s) < X_1(s)$  ( $s$  : two heads,  $P(\{s\}) = 1/4$ ).  $Y$  and  $Z$  have the identical distribution, we still have

$E(Y) = 1 \geq E(X_1) = 0.5$  (**not via monotonicity directly, but through the fact that  $E(Y) = E(Z)$  and the fact that  $E(Z) \geq E(X_1)$  via monotonicity**).

**Distributions would determine expectations:**  $E(Y) = E(Z)$ .

# Probability Theory for EOR

Expectations of functions of random variables  
(with discrete random variables examples)

|

*Expectation* maps a subset of the collection of random variables to numbers (expectations/expected values of associated random variables).

We know some functions of random variables are still random variables, and we look into **expectations of functions of random variables and how they are linked with (the expectations of) the original random variables.**

## Law of the unconscious statistician (LOTUS)

**Distributions determine expectations.**

The **distribution** of  $X$  determines the **distribution** of  $g(X)$ .

The **distribution** of  $X$  determines the **expectation** of  $g(X)$ .

**LOTUS:**

The **distribution** of  $X$   
determines  
the **expectation** of  $g(X)$ .

## Example with discrete random r.v.'s.

If  $X$  is a discrete random variable with support  $\text{supp}(X) = \{x_1, x_2, x_3, \dots\}$ , and  $g(x)$  is a function from  $R \rightarrow R$  such that  $g(X)$  is a discrete r.v., then

$$\mathbb{E}[g(X)] = \sum_{x \in \text{supp}(X)} g(x)P(X = x) \quad (= \sum_{x \in \text{supp}(X)} g(x)\Delta F_X(x))$$

with  $\Delta F_X(x) = F(x) - \lim_{y \uparrow x} F(y)$ .

**Proof with discrete r.v.'s:**

$$\begin{aligned}
 E(g(X)) &= \sum_c c P(g(X) = c) \\
 &= \sum_c c \left( \sum_{a \in \text{supp}(X): g(a)=c} P(X = a) \right) \\
 &= \sum_c \sum_{a \in \text{supp}(X): g(a)=c} c P(X = a) \\
 &= \sum_c \sum_{a \in \text{supp}(X): g(a)=c} g(a) P(X = a) \\
 &= \sum_{a \in \text{supp}(X)} g(a) P(X = a)
 \end{aligned}$$

**Example:**

- ▶ Suppose a random variable  $X$  has outcomes  $\{0, 1, 2, 3\}$  with probabilities given by the probability mass function  $p_X(x)$ .
- ▶ Now consider  $X^2$ . This has outcomes  $\{0, 1, 4, 9\}$ .
- ▶ The probability of seeing 9 from  $X^2$  is the same as seeing 3 from  $X$ ;  
...

So

$$E[X^2] = \sum_{k=0}^3 k^2 P(X^2 = k^2) = \sum_{k=0}^3 k^2 p_X(k)$$

## Another example:

- ▶ Suppose a random variable  $X$  has outcomes  $\{-2, -1, 1, 2\}$  with probabilities given by the probability mass function  $p_X(x)$ .
- ▶ Now consider  $X^2$ . This has outcomes  $\{1, 4\}$ , occurring with the PMF:

$$p_{X^2}(1) = P(X^2 = 1) = P(X = 1) + P(X = -1) = p_X(1) + p_X(-1)$$

$$p_{X^2}(4) = P(X^2 = 4) = P(X = 2) + P(X = -2) = p_X(2) + p_X(-2).$$

- ▶ For the expected value we have

$$\begin{aligned} E[X^2] &= \sum_{k=1,4} kp_{X^2}(k) = \sum_{j=\{-2,-1,1,2\}} j^2 p_X(j) \\ &= \sum_{j=\{-1,1\}} j^2 p_X(j) + \sum_{j=\{-2,2\}} j^2 p_X(j) \end{aligned}$$

# Variance

## Definition

The **variance** of an **real-valued** r.v.  $X$  (if exists) is

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

and the **standard deviation** is

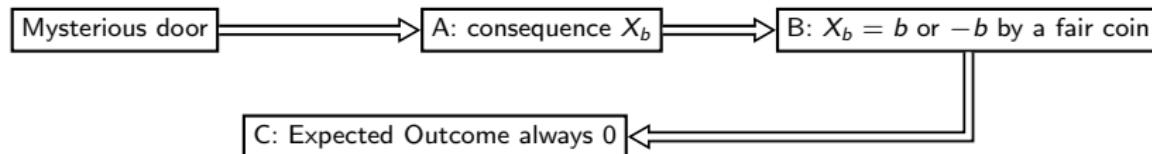
$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

- ▶ Variance of  $X$  is essentially the expectation of a function of  $X$ ,  $g(X)$ , with  $g(x) = (x - \mu)^2$ ,  $\mu = E(X)$ .
- ▶ The distribution of  $X$  determines the variance of  $X$ .

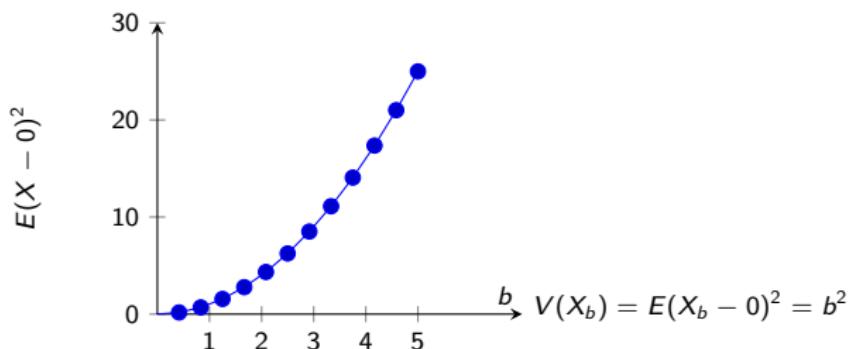
For any **real-valued** r.v.  $X$ ,

$$\begin{aligned}\text{Var}(X) &= E(X - EX)^2 \\&= E(X^2 - 2XE(X) + (EX)^2) \\&= E(X^2) - E(2XE(X)) + E((EX)^2) \\&= E(X^2) - 2E(X)E(X) + (EX)^2 \\&= E(X)^2 - (EX)^2\end{aligned}$$

## Example:



However, it feels different when  $b$  varies from **5** to **0**. How to describe the difference?



It describes the "spread" of the random variable from its "average".

### Some properties of variance:

- ▶ For scalars  $a, c \in \mathbb{R}$ , for **any** r.v.  $X$  (if variance well-defined):

$$\text{Var}(a + cX) = \text{Var}(cX) = c^2 \text{Var}(X).$$

- ▶ For **independent** r.v.'s,  $X_1, X_2, \dots, X_n$  (if exists):

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

Different from the expectation, as for expectation for **any** r.v.'s  $X_1, X_2, \dots, X_n$  (if exists):

$$E(a + cX_1) = a + cE(X_1).$$

$$E \left( \sum_{i=1}^n X_i \right) = \sum_{i=1}^n E(X_i).$$

**Example:**  $X_1, X_2 \sim \text{i.i.d. Bern}(p)$ .

- ▶ Calculate  $E(X_1), E(1 + 2X_1), E(X_1 + X_2), \text{Var}(X_1), \text{Var}(1 + 2X_1), \text{Var}(2X_1), \text{Var}(X_1 + X_2)$ :
- ▶  $E(X_1) = p, E(1 + 2X_1) = 1 + 2p, E(X_1 + X_2) = 2p, \text{Var}(X_1) = p(1 - p), \text{Var}(1 + 2X_1) = 4p(1 - p), \text{Var}(2X_1) = 4p(1 - p), \text{Var}(X_1 + X_2) = 2p(1 - p)$ .

# Probability Theory for EOR

Expectations of functions of random variables  
(with discrete random variables examples)  
II (indicator function)

*Expectation* maps a subset of the collection of random variables to numbers (expectations/expected values of associated random variables).

We know some functions of random variables are still random variables, and we look into

**expectations of functions of random variables and how they are linked with the original random variables.**

A special function: indicator function



## A special function: indicator function

A special function: indicator function

○●○○○○○○○○○○

What is indicator function

$$I_A : S \mapsto \{0, 1\}; A \subset S.$$



$$I_A(s) = \begin{cases} 1, & s \in A \\ 0, & s \in A^c (= S \setminus A) \end{cases}$$

- $(I_A)^k = I_A$  for any  $k \in \mathbb{N}_+$ .
- $I_{A^c} = 1 - I_A$ .
- $I_{A \cap B} = I_A I_B$ .
- $I_{A \cup B} = I_A + I_B - I_A I_B \leq I_A + I_B$ .

$$I_{A \cup B} = 1 - I_{(A \cup B)^c} = 1 - I_{A^c \cap B^c} = 1 - I_{A^c} I_{B^c} = 1 - (1 - I_A)(1 - I_B) = I_A + I_B - I_A I_B.$$

Similar to inclusion-exclusion principle with two events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

A special function: indicator function

○○●○○○○○○○○

Fundamental bridge

$I_A$  is a function, also a random variable, if we consider  $A$  a random event. It can also be a function of r.v.'s, if  $A$  is a random event generated by r.v.'s.

$I_A$  follows  $\text{Bern}(P(A))$ .

**Fundamental bridge:**  $P(A) = EI_A$ .

**Distributions determine expectations, and now expectations recover probabilities of associated random events!**

**Fundamental bridge:**  $P(A) = EI_A$ .

**Distributions determine expectations, and now expectations recover probabilities of associated random events!**

- ▶ Let's look into events associated with arbitrary real-valued r.v.  $X$  of type  $\{X \leq x\}$ .
- ▶  $I_{\{X \leq x\}}$  is a function of  $X$  and also a random variable: takes 1 if  $X \leq x$  otherwise 0.
- ▶ Expectations recover the distribution of any **real-valued** r.v.  $X$ : once all the expectations of  $I_{\{X \leq x\}}, x \in \mathbb{R}$  are known.

A special function: indicator function

○○○○●○○○○○○

Important results derived with indicators

Some important results derived via **indicators** and the **fundamental bridge**.

A special function: indicator function

○○○○●○○○○○

Important results derived with indicators

## I. Inclusion-exclusion principle

A special function: indicator function

○○○○○●○○○○○

Important results derived with indicators

► I.1 Two random events  $A, B$ .

From  $I_{A \cup B} = I_A + I_B - I_A I_B \leq I_A + I_B$ , we know

$$EI_{A \cup B} = EI_A + EI_B - EI_A EI_B \leq EI_A + EI_B.$$

Therefore, from the fundamental bridge

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B).$$

► I.2 n random events  $A_1, \dots, A_n$ .

Note that

$$\begin{aligned} 1 - I_{\bigcup_{i=1}^n A_i} &= I_{\bigcap_{i=1}^n A_i^c} = (1 - I_{A_1}) \cdots (1 - I_{A_n}) \\ &= 1 - \left( \sum_{i=1}^n I_{A_i} - \sum_{i < j} I_{A_i} I_{A_j} + \sum_{i < j < k} I_{A_i} I_{A_j} I_{A_k} - \dots + (-1)^{n+1} \prod_{i=1}^n I_{A_i} \right) \\ &= 1 - \left( \sum_{i=1}^n I_{A_i} - \sum_{i < j} I_{A_i \cap A_j} + \sum_{i < j < k} I_{A_i \cap A_j \cap A_k} - \dots + (-1)^{n+1} I_{\bigcap_{i=1}^n A_i} \right) \end{aligned}$$

Then from the fundamental bridge,  $P(\bigcup_{i=1}^n A_i) =$

$$\sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} P(\bigcap_{i=1}^n A_i).$$

Similarly, we know  $P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$  since  $I_{\bigcup_{i=1}^n A_i} \leq \sum_{i=1}^n I_{A_i}$ .

A special function: indicator function

○○○○○○○●○○○○

Important results derived with indicators

II. Calculating  $E(X)$  of any **non-negative integer-valued r.v.**  $X$  with its survival function.

A special function: indicator function

○○○○○○○○●○○○

Important results derived with indicators

Let  $X$  be a **non-negative integer-valued** r.v. then if its expectation exists

$$\mathbb{E}X = \sum_{i=0}^{\infty} G_X(i)$$

where  $G_X$  is the survival function of  $X$  such that  $G_X(x) = 1 - F_X(x)$ .

A special function: indicator function

○○○○○○○○●○○

Important results derived with indicators

► Proof:

Decompose  $X$  as a function of multiple indicators.

Note that  $X = I_{\{X \geq 1\}} + \cdots + I_{\{X \geq n\}} + \cdots = \sum_{i \in \mathbb{N}_+} I_{\{X \geq i\}}$ .

The above holds true since  $X$  and  $\sum_{i \in \mathbb{N}_+} I_{\{X \geq i\}}$  are the same function from  $S$  to the set of non-negative integers.

Linearity of expectation, fundamental bridge, and the fact that  $\{X \geq i\} = \{X > i - 1\}$ ,  $i \in \mathbb{N}_+$  give

$$EX = \sum_{i \in \mathbb{N}_+} \mathbb{E}I_{\{X \geq i\}} = \sum_{i=1}^{\infty} P(X \geq i) = \sum_{i=0}^{\infty} P(X > i) = \sum_{i=0}^{\infty} G_X(i).$$

A special function: indicator function

○○○○○○○○○●○

Important results derived with indicators

### III. Indicators are a useful tool to simplify questions.

A special function: indicator function

○○○○○○○○○●

Important results derived with indicators

There are  $n$  visitors and  $n$  balls with their names on respectively. If each visitor picks a ball randomly then keeps the ball, denote  $X$  the number of visitors picking the one with their own names. What is  $E(X)$ ?

- ▶ Denote  $A_i$  the event that  $i$ th visitor gets the correct ball.
- ▶  $X = \sum_{i=1}^n I_{A_i}$ .
- ▶ By fundamental bridge,  $EX = \sum_{i=1}^n P(A_i)$ .
- ▶ By symmetry,  $P(A_i) = 1/n$ .
- ▶  $E(X) = 1$ .

# Probability Theory for EOR

Some special discrete random variables

II

Some **discrete random variables** are associated very special/ubiquitous **distributions (PMFs)**, they get their own names!

## Definition

The **probability mass function (PMF)** of a **discrete r.v.**,  $X : S \mapsto \{x_1, x_2, \dots\}$ , is the function:

$$p_X(x) = P(X = x).$$

## Negative Binomial: $\text{NBin}(r, p)$

- If we keep throwing a special coin (with probability  $p$  getting head in each flip) until we get  $r$  heads, let  $X$  denote the number of tails/bottomes/failures.
- $X \sim \text{NBin}(r, p)$ .
- The PMF of  $X$  is

$$P(X = n) = \binom{n+r-1}{r-1} p^r (1-p)^n; n = 0, 1, \dots$$

- $\text{NBin}(1, p)$  is also called Geometric distribution  $\text{Geom}(p)$  (the number of bottoms before the first head).

## Negative Binomial: $\text{NBin}(r, p)$

If  $Z \sim \text{NBin}(r, p)$ , what is  $E(Z), \text{Var}(Z)$ ?

- ▶ Denote  $Y_1$  the number of bottoms before the first head,  $Y_2$  the number of bottoms after the first head and before the second head,  
...  
▶  $Y_i$  follows  $\text{Geom}(p)$  (i.i.d.);  $X = \sum_{i=1}^r Y_i$ .
- ▶ Then  $X \sim \text{NBin}(r, p)$ .
- ▶  $EZ = EX = \sum_{i=1}^r EY_i = r \times (1 - p)/p$ , since

$$EY_i = \sum_{k=0}^{\infty} k(1 - p)^k p = p(1 - p) \sum_{k=0}^{\infty} k(1 - p)^{k-1} = (1 - p)/p.$$

- ▶  $\text{Var}(Z) = \text{Var}(X) = \sum_{i=1}^r \text{Var}(Y_i) = r \times (1 - p)/p^2$ , since

$$EY_i^2 = \sum_{k=0}^{\infty} k^2(1 - p)^k p = (1 - p)(2 - p)/p^2.$$

$$\text{Var}(Y_i) = EY_i^2 - (EY_i)^2 = (1 - p)/p^2.$$

## Negative Hypergeometric: NHGeom( $w, b, r$ )

- ▶ Keep marking balls with \*'s from a urn of  $w$  white balls and  $b$  **black** balls (each ball is equally likely to be marked) until there are  $r$  white balls with marks, and let  $X$  denote the number of balls being both **black** and marked with \*'s.
- ▶  $X \sim \text{NHGeom}(w, b, r)$ .
- ▶ The PMF of  $X$  is

$$P(X = k) = \frac{\binom{w}{r-1} \binom{b}{k}}{\binom{w+b}{r+k-1}} \frac{w-r+1}{w+b-r-k+1}; k = 0, 1, \dots$$

## Negative Hypergeometric: $\text{NHGeom}(w, b, r)$

If  $Z \sim \text{NHGeom}(w, b, r)$ , what is  $E(Z)$ ?

- The underlying sample space contain outcomes such as

$$bbbwwbw \cdots bb$$

which orders all balls.

For each black ball, it can choose  $w + 1$  positions with equal likeliness: before the first white ball, between 1st and 2nd white balls, between 2nd and 3rd white balls,  $\dots$ , between  $(w - 1)$ th and  $w$ th white balls, and after  $w$ th white balls.

- Let's label all black balls from 1 to  $b$ . Denote  $I_j$  the indicator random variable which equals one if the black ball  $j$  chooses any positions before the  $r$ th white ball otherwise zero.  $EI_j = r/(w + 1)$ .
- Let  $X = \sum_{i=1}^b I_i$ , and thus  $X \sim \text{NHGeom}(w, b, r)$ .

$$E(Z) = E(X) = \sum_{i=1}^b EI_i = br/(w + 1).$$

# Probability Theory for EOR

Some special discrete random variables

III (Poisson)

Some **discrete random variables** are associated very special/ubiquitous **distributions (PMFs)**, they get their own names!

## Definition

The **probability mass function (PMF)** of a **discrete r.v.**,  $X : S \mapsto \{x_1, x_2, \dots\}$ , is the function:

$$p_X(x) = P(X = x).$$

# Poisson!

# Poisson Process

Consider the following scenario:

**A web server would receive requests from computer A randomly.**

**How to properly model the random variable  $X$ ,  
the number of requests from computer A within a time interval of  
length  $t$ .**

**How to properly model the random variable  $X$ , the number of requests from computer A within a time interval of length  $t$ .**

- ▶ The probability of more than one hit in a very short interval is negligible.

The probability of a single arrival in a very short interval is proportional to the length of the interval.

- ▶ The numbers of hits in non-overlapping time intervals are independent.

E.g, what happens in interval  $[a, b]$  would be independent from what happens in  $[c, d]$  for all  $a, b, c, d$  from the real line as long as  $[a, b] \cap [c, d] = \emptyset$ .

**The probability of more than one hit in a very short interval is negligible.**

**The probability of a single arrival in a very short interval is proportional to the length of the interval.**

- ▶ The arrival of one request within a very small interval  $(s, s + \Delta t]$  follows  $\text{Bern}(\lambda\Delta t)$ , denoted as  $I_{(s,s+\Delta t]}$ ;
- ▶ The number of total requests within interval  $(s, s + t]$

$$X_{\Delta t} = \sum_{i=1}^{t/(\Delta t)} I_{(s+(i-1)\Delta t, s+i\Delta t]},$$

where for simplicity we assume that  $t/(\Delta t)$  is a positive integer.

**The number of hits in non-overlapping time intervals are independent.**

- ▶  $X_{\Delta t}$  follows  $\text{Bin}(t/\Delta t, \lambda\Delta t)$ .
- ▶  $P(X_{\Delta t} = k) = \binom{t/(\Delta t)}{k} (\lambda\Delta t)^k (1 - \lambda\Delta t)^{t/(\Delta t) - k}$ .

*Let's consider a special case such that  $\Delta t = t/n$ , what happens to the PMF of the random variable  $X_{\Delta t}$  if we let  $n \rightarrow \infty$  ( $\Delta t \rightarrow 0$ , we are splitting into smaller and smaller intervals). Denote the limit random variable as  $X$ .*

- ▶  $P(X = k) = \lim_{n \rightarrow \infty} \binom{n}{k} (\lambda t/n)^k (1 - \lambda t/n)^{n-k} = \lim_{n \rightarrow \infty} (\lambda t)^k / k! \left( \prod_{i=1}^k (n-i+1)/n \right) (1 - (\lambda t)/n)^n (1 - (\lambda t)/n)^{-k} = e^{-\lambda t} (\lambda t)^k / k!$ .

We use the fact that  $(1 + x/s)^s \rightarrow_{s \rightarrow \infty} e^x$ .

## Poisson distribution: $\text{Pois}(\lambda)$

Denote  $X$  the number of random requests within a time interval of **unit length** from the aforementioned scenario ( $t=1$ ).

- The PMF of  $X$  is

$$P(X = k) = e^{-\lambda} \lambda^k / k!; k = 0, 1, \dots$$

where  $\lambda > 0$ .

- $X \sim \text{Pois}(\lambda)$ .

# Expectation

$X \sim \text{Pois}(\lambda)$ .

- The expectation  $EX$ .

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda, \end{aligned}$$

where we use the fact that  $e^x = \sum_{k=0}^{\infty} x^k / k!$ .

**$\lambda$  describes the arrival rate, larger the arrival rate, larger the expected number of arrivals.**

# Variance

$$X \sim \text{Pois}(\lambda).$$

- The variance  $VX$ .

$$\begin{aligned}
 E[X^2] &= \sum_{x=0}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=1}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
 &= \sum_{x=1}^{\infty} (x-1) \frac{\lambda^x e^{-\lambda}}{(x-1)!} + \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
 &= \lambda \sum_{y=0}^{\infty} y \frac{e^{-\lambda} \lambda^y}{y!} + \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\
 &= \lambda E(X) + \lambda = \lambda^2 + \lambda,
 \end{aligned}$$

$$VX = E[X^2] - (E[X])^2 = \lambda.$$

Poisson Process

○○○○○○○●○○○○○

Special properties of Poisson

## **From Binomial to Poisson revisits: Poisson paradigm.**

**Previously,** The arrival of one request within a very small interval  $(s, s + \Delta t]$  follows  $\text{Bern}(\lambda \Delta t)$ , denoted as  $I_{(s,s+\Delta t]}$ ; and the number of total requests within interval  $(s, s + t]$

$$X_{\Delta t} = \sum_{i=1}^{t/(\Delta t)} I_{(s+(i-1)\Delta t, s+i\Delta t]}.$$

The limit once we let  $\Delta t$  go to zero follows

$$\text{Pois}(\lambda), \lambda = \sum_{i=1}^{t/(\Delta t)} \lambda \Delta t.$$

**Poisson paradigm** Let  $A_i, i = 1, \dots, n$  be **independent (or weakly dependent)** events with probability  $p_i$ , **n is large and  $p_i$  are very small**. Let

$$X = \sum_{j=1}^n I_{A_j}$$

count how many of the  $A_i$  occur (how many j visit one specific island within one day for example). Then **X is approximated distributed as**

$$\text{Pois}(\lambda); \lambda = \sum_j p_j.$$

- $\lambda$  as the *rate* (expected number within a certain time period) of occurrence of *rare events*.

Poisson Process

○○○○○○○○●○○○

Special properties of Poisson

**Sum of independent poisson is still Poisson.**

A web server would receive requests from computer A randomly, but also requests from computer B. A and B are independent.

**How to properly model the random variable  $Y$ ,  
the number of requests from either computer A or B within a time interval of unit length.**

- ▶ Simply sum their arrival rate, and all the rest follow the same logic.
- ▶  $A \sim \text{Pois}(\lambda_A)$ ,  $B \sim \text{Pois}(\lambda_B)$  and A,B are independent. Then

$$Y = A + B \sim \text{Pois}(\lambda_A + \lambda_B)$$

- ▶ Verify the result by looking at the distribution (PMF).

$$\begin{aligned} P(A + B = k) &= \sum_{i=0}^k P(A = k - i | B = i) P(B = i) = \sum_{i=0}^k P(A = k - i) P(B = i) \\ &= \sum_{i=0}^k \frac{1}{(k-i)!} \lambda_A^{k-i} e^{-\lambda_A} \cdot \frac{1}{i!} \lambda_B^i e^{-\lambda_B} = e^{-(\lambda_A + \lambda_B)} \frac{1}{k!} \sum_{i=0}^k \frac{k!}{i!(k-i)!} \lambda_A^{k-i} \lambda_B^i \\ &= e^{-(\lambda_A + \lambda_B)} \frac{1}{k!} \sum_{i=0}^k \binom{k}{i} \lambda_A^{k-i} \lambda_B^i = e^{-(\lambda_A + \lambda_B)} \frac{1}{k!} (\lambda_A + \lambda_B)^k \end{aligned}$$

- ▶  $(a + b)^k = (a + b)(a + b) \cdots (a + b) = \sum_{i=0}^k \binom{k}{i} a^{k-i} b^i$ .

Poisson Process

○○○○○○○○○○●○

Special properties of Poisson

**From Poisson to Binomial: Poisson given a sum of Poissons.**

- **Poisson conditional on the sum is Binomials .**
- If a web server receive total  $n$  requests from either computer A or B.
- Given these two computers are independent,  
**for each signal it could either be from A or B:**

$$I_{iA} \sim \text{Bern}\left(\frac{\lambda_A}{\lambda_A + \lambda_B}\right).$$

- If  $A \sim \text{Pois}(\lambda_A)$ ,  $B \sim \text{Pois}(\lambda_B)$ , A is independent from B, then the number of requests from computer A given  $A + B = n$ :

$$A \text{ given } \{A + B = n\} \text{ follows } \text{Bin}\left(n, \frac{\lambda_A}{\lambda_A + \lambda_B}\right)$$

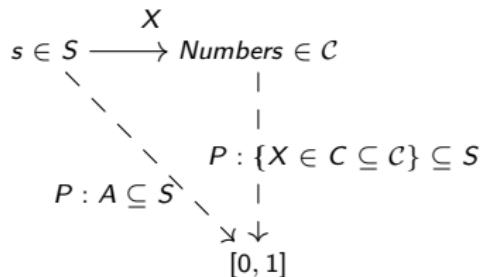
- Verify the result by looking at the distribution (PMF).

$$\begin{aligned} & P(A = k | A + B = n) \\ &= \frac{P(A = k, B = n - k)}{P(A + B = n)} = \frac{P(A = k) P(B = n - k)}{P(A + B = n)} \\ &= \frac{\left(e^{-\lambda_A} \frac{\lambda_A^k}{k!}\right) \left(e^{-\lambda_B} \frac{\lambda_B^{n-k}}{(n-k)!}\right)}{\left(e^{-(\lambda_A + \lambda_B)} \frac{(\lambda_A + \lambda_B)^n}{n!}\right)} = \frac{n!}{k!(n-k)!} \left(\frac{\lambda_A}{\lambda_A + \lambda_B}\right)^k \left(\frac{\lambda_B}{\lambda_A + \lambda_B}\right)^{n-k} \\ &= \binom{n}{k} \left(\frac{\lambda_A}{\lambda_A + \lambda_B}\right)^k \left(1 - \frac{\lambda_A}{\lambda_A + \lambda_B}\right)^{n-k}. \end{aligned}$$

## Layman's talk : what is continuous random variable

# Random variables are functions

$X$  maps elements from  $S$  to numbers! Now we can work with functions/numbers!



$X$  essentially is a function to numbers!

E.g.,  $X$  may assign random outcomes with real numbers, and thus it may take values from the real line.

To describe how the probability values are assigned to different events generated by a **real-valued** random variable  $X$ , we can look at the cumulative distribution function  $F_X$  such that

$$F_X(x) = P(X \leq x), x \in \mathbb{R}.$$

Random variables are functions

OO

Continuous r.v. example

●○○○○○○○

## Continuous r.v. example

Let's keep throwing a ten-sided fair die ((with numbers 0,1,...,9) ) without stopping!

- We have outcomes such as

$$a = 010310819\ldots$$

- Let's map these outcomes to numbers with [0,1] via the decimal expansion:

$$X(a) = 0.a.$$

Here we follow the convention such that  $0.99999\ldots = 1$ .

- What is the probability of the event  $\{X = 1\}$ ?  
**ZERO!**
- $P(X = 1) = P(\{99999999\ldots\}) = \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \cdots = 0$ .
- $P(X = a) = 0$ .

Let's keep throwing a ten-sided fair die (with numbers 0,1,...,9) without stopping!

- We have outcomes such as

$$a = 010310819\ldots$$

- Let's map these outcomes to numbers with  $[0,1]$ :  $X(a) = 0.a$ .
- How do we characterize the distribution of  $X$ : e.g., how likely is the event  $\{X \leq 0.5\}$ .

$\{X \leq 0.5\}$  contains all the outcomes such that the first throw is strictly smaller than five (0,1,2,3,4) and the specific outcome (only the first throw is a five the rests are all 0's)  
5000000000 ... :

$$P(X \leq 0.5) = 0.5 + P(\{5000000000 \cdots\}) = 0.5$$

$$P(X \leq a) = a, a \in [0, 1].$$

$$P(X \leq a) = a, a \in [0, 1].$$

- ▶ Let's cut  $[0, 1]$  into  $10^1$  grids such that each grid can be expressed as

$$I_i = [x_{i-1,1}, x_{i,1}], i \in \{1, \dots, 10^1\}$$

with  $x_j = j10^{-1}, j \in \{0, 1, \dots, 10^1\}$ . E.g.,  $i = 2, [0.1, 0.2]$ .

It is easy to check that

$$P(X \in \cup_{j=1}^i I_j) = P(X \leq x_i) = x_i = i * 10^{-1}, i \in \{1, \dots, 10\}$$

and thus for each grids  $P(X \in I_j) = 10^{-1}$ .

Let's approximate the probability value of  $\{X \leq a\}$  with

$$P(X \in \cup_{j=1}^{j_a} I_j) = \sum_{i=1}^{j_a} \Delta x_i, \Delta x_i = x_i - x_{i-1}, j_a = \min\{i \in \{1, \dots, 10^1\} : a \in I_i\}.$$

- ▶

$$\left| P(X \leq x) - \sum_{i=1}^{j_a} \Delta x_i \right| \leq 10^{-1}.$$

$$P(X \leq a) = a, a \in [0, 1].$$

- ▶ Let's cut  $[0, 1]$  into  $10^n$  grids such that each grid can be expressed as

$$I_i = [x_{i-1,n}, x_{i,n}], i \in \{1, \dots, 10^n\}$$

with  $x_j = j10^{-n}, j \in \{0, 1, \dots, 10^n\}$ . E.g.,  $i = 2, [10^{-n}, 2 * 10^{-n}]$ .  
It is easy to check that

$$P(X \in \bigcup_{j=1}^i I_j) = P(X \leq x_i) = x_i = i * 10^{-n}, i \in \{1, \dots, 10^n\}$$

and thus for each grids  $P(X \in I_j) = 10^{-n}$ .

Let's approximate the probability value of  $\{X \leq a\}$  with

$$\sum_{i=1}^{j_a} \Delta x_i, \Delta x_i = x_i - x_{i-1}, j_a = \min\{i \in \{1, \dots, 10^n\} : a \in I_i\}.$$

- ▶

$$\left| P(X \leq a) - \sum_{i=1}^{j_a} \Delta x_i \right| \leq 10^{-n}.$$

- ▶

$$\left| P(X \leq a) - \sum_{i=1}^{j_a} \Delta x_i \right| \leq 10^{-n}.$$

- ▶

$$F_X(a) = P(X \leq a) = a, a \in [0, 1].$$

- ▶ Let's cut  $[0, 1]$  into  $10^n$  grids such that each grid can be expressed as

$$I_i = [x_{i-1,n}, x_{i,n}], i \in \{1, \dots, 10^n\}$$

with  $x_j = j10^{-n}, j \in \{0, 1, \dots, 10^n\}$ . E.g.,  $i = 2, [10^{-n}, 2 * 10^{-n}]$ .

It is easy to check that

$$P(X \in \cup_{j=1}^i I_j) = P(X \leq x_i) = x_i = i * 10^{-n}, i \in \{1, \dots, 10^n\}$$

and thus for each grids  $P(X \in I_j) = 10^{-n}$ .

Let's approximate the probability value of  $\{X \leq a\}$  with

$$\sum_{i=1}^{j_a} \Delta x_i, \Delta x_i = x_{i-1}, j_a = \min\{i \in \{1, \dots, 10^n\} : a \in I_i\}.$$

- ▶

$$\sum_{i=1}^{j_a} \Delta x_i \rightarrow \int_0^a f_X(x) dx = a; f_X(x) = F'_X(x).$$

$$F_X(a) = P(X \leq a) = a, a \in [0, 1].$$

$$\int_{-\infty}^a f_X(x) dx = P(X \leq a) = a; f_X(x) = F'_X(x) = 1, x \in [0, 1] \text{ otherwise zero.}$$

- ▶ Two different ways describing the distribution of  $X$ , and particularly, there exists a probability density function  $f_X$  (integrate the (probability) density, we have the (probability) mass).

We have a probability density function  $f_X$  (integrate the (probability) density via **Riemann integral**, we have the (probability) value). If an interval has positive density function values, then there are **uncountably many** possible values in the interval taken by  $X$ .

$$\int_{-\infty}^a f_X(x) dx = P(X \leq a); a \in \mathbb{R}.$$

Any real-valued random variable that has a non-negative function  $f_X$  satisfying the above Riemann integral condition is a **continuous** real-valued random variable.

Let's compare.

- ▶ In the example, we see that we can choose  $f_X(x) = F'_X(x)$  and  $F'_X(x) = 1, x \in [0, 1]$ . There are **uncountably infinite** possible values in  $[0, 1]$  taken by  $X$ .
- ▶ For any discrete random variable  $Y$ , almost everywhere  $F'_Y$  is simply **zero**, there is no  $f_Y$  satisfying the above identity (via Riemann integral). There are **at most countably many** possible values taken by  $Y$ .

# Probability Theory for EOR

## Random variables (continuous)

Random variables are functions

●○○○

Continuous random variable

○○○○○○○

# Random variables are functions

## Definition (Random variables (r.v.'s) (preliminary))

A random variable (r.v.) is a function maps elements in sample space  $S$  to numbers in  $\mathcal{C}$ , e.g.,  $\mathcal{C} \subseteq \mathbb{R}$ .

There are functions from a sample space to a set of numbers that can be listed with natural numbers:  $x_1, x_2, x_3, \dots$  (which contains no non-empty open subsets/intervals).

## Definition (Discrete random variable)

A random variable  $X$  is a **discrete** random variable if there exists a set  $C$  with at most countably many numbers (so you can index all the elements within  $C$  using a set of natural numbers):

$$P(\{X \in C\}) = 1, \text{ or equivalently, } P(X \in C) = 1.$$

The **support** of a **discrete** random variable is the set of all numbers,  $x$ 's, such that  $P(X = x) > 0$  (**PMF**):

$$\text{supp}(X) = \{x \in C : P(X = x) > 0\}.$$

The support of a discrete random variable has empty interior (it contains no open intervals at all).

Random variables are functions

○○○○

Continuous random variable

●○○○○○○

## Continuous random variable

There are functions from a sample space to a set of numbers that contains non-empty open sets: e.g., random variables could take any positive values from the interval  $(0, \infty)$ .

## Example: waiting time

Customers come in according to a Poisson process such that:

*Poisson arrivals.*

The number of arrivals that occur in an interval of length  $t$  is a  $\text{Pois}(\lambda t)$  r.v.;

*Independence condition.*

The number of arrivals that occur in disjoint intervals are independent from each other.

**The distribution of the waiting time until the first customer?**

## Example: waiting time

**The distribution of the waiting time until the first customer?**

- ▶ Denote  $N_t$  the number of arrivals within  $[0, t]$ , then  $N_t$  follows  $\text{Pois}(\lambda t)$ .
- ▶ Denote  $T_1$  the time until the first arrival.
- ▶ Note that  $\{T_1 > t\} = \{N_t = 0\}$ .
- ▶  $P(T_1 > t) = P(N_t = 0) = e^{-\lambda t}$ .
- ▶ The CDF of  $T_1$  then

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - e^{-\lambda t}.$$

- ▶ Note that  $T_1$  can take any non-negative values, and its CDF is differentiable which can be expressed as a Riemann integral.

## Example: waiting time

We have two ways to specify the distribution of  $T_1$ .

$$F_{T_1}(t) = 1 - e^{-\lambda t} = \int_0^t f_{T_1}(t)dt.$$

$$f_{T_1}(t) = F'_{T_1}(t) = \lambda e^{-\lambda t}, t \in (0, \infty).$$

- ▶  $f_{T_1}$  is called the **probability density function (PDF)** of  $T_1$ .
- ▶  $T_1$  may take any value  $t$  where  $f_{T_1}(t) > 0: (0, \infty)$ .
- ▶ Integrate the (probability) **density** via the Riemann integral we get the (probability) **value**.
- ▶ If we work with Riemann integral, we can not find such  $f$  for discrete random variables.

E.g., For a  $\text{Bern}(p)$  distributed  $X$ :  $F_X(t) = \begin{cases} 0; & t < 0 \\ 1/2; & 0 \leq t < 1 \\ 1; & t \geq 1 \end{cases}$

## Definition (Continuous random variable)

A **real-valued** random variable  $X$  is a **continuous** random variable if there exists a non-negative function  $f_X$  such that:

$$P(X \leq x) = F_X(x) = \int_{-\infty}^x f_X(s)ds$$

If  $F_X$  is differentiable, we can choose  $f_X(x) = F'_X(x)$ .

The **support** of a **continuous** random variable with differentiable CDF  $F_X$ , conventionally, is the set of all numbers,  $x$ 's, such that  $f_X(x) = F'_X(x) > 0$  (**PDF**):

$$supp(X) = \{x \in C : F'_X(x) > 0\}.$$

The support of a continuous random variable has **non-empty interior** (it contains open intervals): e.g.,  $(0, \infty)$ .

- A **real-valued** random variable  $X$  is a **discrete** random variable if there exists a set  $C$  with at most countably many numbers (so you can index all the elements within  $C$  using a set of natural numbers):

$$P(\{X \in C\}) = 1, \text{ or equivalently, } P(X \in C) = 1.$$

The **support** of a **discrete real-valued** random variable is the set of all numbers,  $x$ 's, such that  $P(X = x) > 0$  (**PMF**):

$$\text{supp}(X) = \{x \in C : P(X = x) > 0\}.$$

The support of a discrete random variable has **empty interior** (it contains no open intervals at all): e.g.,  $\{0, 1, 2, 3, \dots\}$ .

- A **real-valued** random variable  $X$  is a **continuous** random variable if there exists a non-negative function  $f_X$  such that:

$$P(X \leq x) = F_X(x) = \int_{-\infty}^x f_X(s)ds$$

If  $F_X$  is differentiable, we can choose  $f_X(x) = F'_X(x)$ .

The **support** of a **continuous** random variable with differentialable CDF  $F_X$ , conventionally, is the set of all numbers,  $x$ 's, such that  $f_X(x) = F'_X(x) > 0$  (**PDF**):

$$\text{supp}(X) = \{x \in C : F'_X(x) > 0\}.$$

The support of a continuous random variable has **non-empty interior** (it contains open intervals): e.g.,  $(0, 1)$ .

# Probability Theory for EOR

## Random variables and their distributions

### II

Distributions: probability values of events associated with random variables



Distributions: probability values of events associated  
with random variables

**Real-valued r.v.  $X$  maps elements from  $S$  to real numbers!**

Now we can work with functions/numbers:

$$X : S \mapsto \mathbb{R}$$

- ▶ We can now use  $\{X \in C\}$  with some subsets  $C \subseteq \mathbb{R}$  to represent different random events (not all subsets  $C$  generate well-defined events, and usually we focus on different intervals: e.g.,  $(-\infty, a]$ ,  $[a, b] \cap (a, e) \dots$ ).

**The distribution of a given r.v. needs to specify all probability values of all those events generated by the r.v.:**

$$P : \{\{X \in C\} : \text{some subsets } C \subseteq \mathbb{R}\} \mapsto [0, 1].$$

For example, a distribution needs to be able to answer  $P(X \in [a, b] \cup [c, d])$ ,  $P(X = e)$ , ...

- ▶ Though there are so many associated events, **we only need to know probability values of some selected events** and the rest probability values can be inferred from these values.

## Definition (CDF of real-valued r.v.)

The **cumulative distribution function (CDF)** of an **real-valued r.v.**,  $X : S \rightarrow \mathbb{R}$ , is the function, usually denoted by  $F_X$ , which maps numbers to numbers within  $[0,1]$ .

$$F_X(x) = P(X \leq x).$$

- We only need to know probability values of the following types of events:  $\{X \leq x\}, x \in \mathbb{R}$ .

### Properties:

- **Increasing function from zero to one:**

$$0 \leq F_X(x_1) \leq F_X(x_2) \leq 1, \forall -\infty < x_1 \leq x_2 < \infty.$$

$$(a): \lim_{x \rightarrow -\infty} F_X(x) = 0; \text{ think about } \emptyset!$$

$$(b): \lim_{x \rightarrow \infty} F_X(x) = 1; \text{ think about } S!$$

- **Right-continuous:**  $F_X(a) = \lim_{x \downarrow a} F(x)$ .

## Simple example revisits: waiting time

**Customers come in according to a Poisson process such that:**

*Poisson arrivals.*

The number of arrivals that occur in an interval of length  $t$ ,  $N_t$ , is a  $\text{Pois}(\lambda t)$  r.v.;

*Independence condition.*

The number of arrivals that occur in disjoint intervals are independent from each other.

**How long do you need to wait until the first customer?**

- ▶ Denote  $T_1$  the time until the first arrival.
- ▶ The CDF of  $T_1$  then

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - P(N_t = 0) = 1 - e^{-\lambda t}$$

$$= \int_0^t \lambda e^{-\lambda s} ds.$$

# From CDF to PDF

## Definition (CDF of real-valued r.v.)

The **cumulative distribution function (CDF)** of an **real-valued** r.v.,  $X : S \rightarrow \mathbb{R}$ , is the function, usually denoted by  $F_X$ :

$$F_X(x) = P(X \leq x).$$

## Definition (PDF of continuous real-valued r.v.)

The **probability density function (PDF)** of a **continuous** real-valued r.v. is a non-negative function  $f_X$  on the real line such that via the Riemann integral:

$$\int_{-\infty}^x f_X(s) ds = P(X \leq x).$$

For a **continuous** random variable with differentiable CDF  $F_X$ , conventionally,  $f_X(x) = F'_X(x)$ .

## Definition (PDF of continuous real-valued r.v.)

The **probability density function (PDF)** of a **continuous** real-valued r.v. is a non-negative function  $f_X$  on the real line such that via the Riemann integral:

$$\int_{-\infty}^x f_X(s)ds = P(X \leq x).$$

For a **continuous** random variable with differentiable CDF  $F_X$ , conventionally,  $f_X(x) = F'_X(x)$ .

- $f_X$  can be used to calculate the CDF function values, and thus  $f_X$  is also a way describing the distribution of **continuous real-valued** random variables.

### Properties:

- **Non-negativity:**  $f_X(x) > 0$  if  $x \in \text{Supp}(X)$ ;  $f_X(x) = 0$  otherwise.
- **Integrate to 1.**  $\int_{-\infty}^{\infty} f_X(x)dx = 1$ .

E.g.,  $\lambda = 3, 3e^{-3t} > 0, t > 0, 3 * e^{-0.03} > 2.911$ .

Distributions: probability values of events associated with random variables

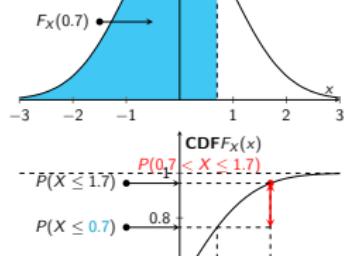
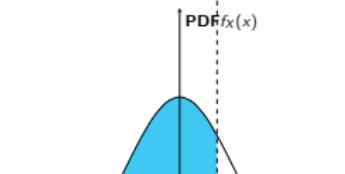
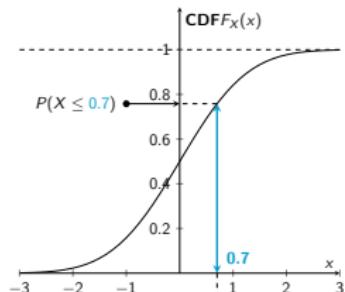
○○○○○●○○

**Continuous r.v.'s and their distributions**

The distribution tell us almost all we need to know about a random variable.

$\circ\circ\circ\circ\circ\circ\circ\bullet\circ$ 

Continuous r.v.'s and their distributions

Example I: standard normal distribution,  $N(0, 1)$ , CDF PDF

The standard normal distribution has  
PDF

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

and CDF

$$F(x) = \int_{-\infty}^x f(s)ds.$$

Suppose  $X$  follows the standard normal distribution.

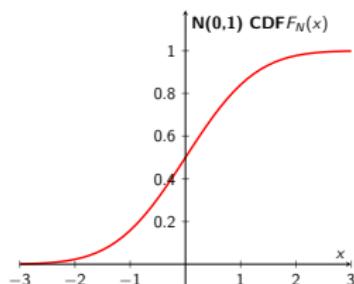
►  $P(X \leq 0.7)$ ?

$$P(X \leq 0.7) = F_X(0.7);$$

$$P(X \leq 0.7) = \int_{-\infty}^{0.7} f_X(s)ds.$$

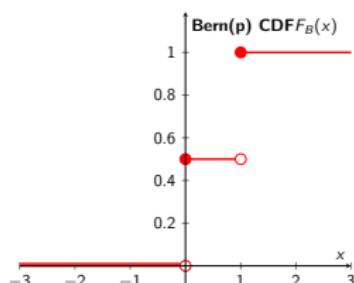
►  $P(0.7 < X \leq 1.7)$ ?

○○○○○○○●

**Continuous r.v.'s and their distributions****Example II: compare standard normal distribution  $N(0, 1)$  with Bernoulli distribution****Bern( $p$ ) distribution****The  $N(0, 1)$  distribution has CDF**

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds.$$

- ▶ **No jumps**, continuous function.
- ▶  $F'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} > 0$  at some open intervals.
- ▶ There exists  $f$  (e.g.,  $F'(x)$ ) such that  $F(x) = \int_{-\infty}^x f(s)ds$ .

**The Bern( $p$ ) distribution has CDF**

$$F(x) = 0I_{(-\infty, 0)}(x) + 1/2I_{[0, 1)} + 1I_{[1, \infty)}.$$

- ▶ **Jumps** at 0, and 1.
- ▶  $F'(x) = 0$  almost everywhere, and not well-defined at 0 and 1.
- ▶ **No such  $f$**  such that  $F$  can be rewritten as a Riemann integral of  $f$ .

# Probability Theory for EOR

Expectation of continuous random variables

Expected values of r.v.'s (if exist) are numbers.



## Expectation of continuous r.v.'s

## Definition (Expectation of discrete r.v.'s)

The expectation/expected value/first moment/mean/average of a **discrete** random variable  $X$  for  $x_1, x_2, x_3, \dots$  (if exists) is defined by

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i) \quad (= \sum_{i=1}^{\infty} x_i \Delta F_X(x_i))$$

(we order all values from the support of  $X$  by the increasing order:  $x_1 \leq x_2 \leq x_3, \dots$ .

We may choose  $\Delta F_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ , let  $x_0 < x_1, x_0 \notin \text{supp}(X)$  and thus  $F(x_0) = 0$ .

We may choose  $\Delta F_X(x) = F(x) - \lim_{y \uparrow x} F(y)$ , as another possible expression.

Essentially both expressions give:  $\Delta F_X(x_i) = P(X = x_i).$ )

If  $|\text{supp}(X)|$  is finite with elements  $x_1, x_2, x_3, \dots, x_n$ , then

$$\mathbb{E}[X] = \sum_{i=1}^n x_i \mathbb{P}(X = x_i) \quad (= \sum_{i=1}^n x_i \Delta F_X(x_i))$$

## Definition (Expectation of discrete r.v.'s)

The expectation/expected value/first moment/mean/average of a **discrete** random variable  $X$  for  $x_1, x_2, x_3, \dots$  (if exists) is defined by

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i) \quad (= \sum_{i=1}^{\infty} x_i \Delta F_X(x_i))$$

(we order all values from the support of  $X$  by the increasing order:  $x_1 \leq x_2 \leq x_3, \dots$ .

We may choose  $\Delta F_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ , let  $x_0 < x_1, x_0 \notin \text{supp}(X)$  and thus  $F(x_0) = 0$ .

We may choose  $\Delta F_X(x) = F(x) - \lim_{y \uparrow x} F(y)$ , as another possible expression.

Essentially both expressions give:  $\Delta F_X(x_i) = P(X = x_i).$ )

If  $|\text{supp}(X)|$  is finite with elements  $x_1, x_2, x_3, \dots, x_n$ , then

$$\mathbb{E}[X] = \sum_{i=1}^n x_i \mathbb{P}(X = x_i) \quad (= \sum_{i=1}^n x_i \Delta F_X(x_i))$$

## Example: average waiting time

**On average: how long do you need to wait until the first customer?**

- ▶ Denote  $N_t$  the number of arrivals within  $[0, t]$ ,  $N_t$  follows  $\text{Pois}(\lambda t)$ . Let  $T_1$  be the time until the first arrival:  $\{T_1 > t\} = \{N_t = 0\}$  and thus  $P(T_1 > t) = P(N_t = 0) = e^{-\lambda t}$ .
- ▶ The CDF of  $T_1$  then

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - e^{-\lambda t}.$$

- ▶ To calculate the average waiting time, we may consider this approximation by cutting time into pieces of unit length  $t_0 = 0, t_1 = 1, \dots$

$$\begin{aligned} & \frac{0+1}{2} P(0 < T_1 \leq 1) + \frac{1+2}{2} P(1 < T_1 \leq 2) + \dots \\ &= \sum_{i=0}^{\infty} \frac{t_i + t_{i+1}}{2} \times P(t_i < T_1 \leq t_{i+1}) = \sum_{i=0}^{\infty} \frac{t_i + t_{i+1}}{2} \times (F_{T_1}(t_{i+1}) - F_{T_1}(t_i)). \end{aligned}$$

- ▶ We can choose finer and finer grids such that  $t_{i+1} - t_i \rightarrow 0$ :

$$\sum_{i=0}^{\infty} \frac{t_i + t_{i+1}}{2} \times (F_{T_1}(t_{i+1}) - F_{T_1}(t_i)) \approx \sum_{i=0}^{\infty} t_i F'_{T_1}(t_i) (t_{i+1} - t_i)$$

$$\rightarrow \int_0^{\infty} t F'_{T_1}(t) dt.$$

- ▶ From the fact that  $F_{T_1}(t) = \int_{-\infty}^t f_{T_1}(s) ds$ , we know  $\int_0^{\infty} t F'_{T_1}(t) dt (= \int_{-\infty}^{\infty} t f_{T_1}(t) dt)$ .

## Definition (Expectation of continuous real-valued r.v.'s)

The expectation/expected value/first moment/mean/average of a **continuous** real-valued random variable  $X$  (if exists) with PDF  $f$  is defined via Riemann integral:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf_X(x)dx \\ &= \int_{-\infty}^{\infty} xdF_X(x) \end{aligned}$$

*the second line (color red) makes use of the Riemann–Stieltjes integral, a generalization of the Riemann integral*

## Law of the unconscious statistician (LOTUS)

For random variable  $g(X)$ , the expectation (if exists):

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

**Properties of Riemann integral imply the linearity and monotonicity.**

# Probability Theory for EOR

Discrete v.s. Continuous random variables

## Discrete v.s. Continuous random variables: differences and connections

**Real-valued random variables**

	CDF $\mathbb{P}(Z \leq z)$	PMF/PDF	Expectation $\mathbb{E}Z$
<b>Discrete</b> $R(X) = \text{supp}(X) = \{x_1 < \dots < x_n < \dots\}$ $R(\overset{\circ}{X}) = \emptyset.$	<b>Step function <math>F(x)</math></b> increase from 0 to 1 <b>Jumps</b>	$p_X(x_i) = F(x_i) - F(x_{i-1})$ $= F(x_i) - \lim_{s \uparrow x_i} F(s)$ $F(x) = \sum_{\substack{x_i \in R(x); \\ x_i \leq x}} p_X(x_i)$ $p_X(x) \geq 0; \sum_{x_i \in R(X)} p_X(x_i) = 1$	$\sum_{\substack{x_i \in R(x); \\ x_i < x_{i+1}}} x_i \Delta F(x_i)$ $\sum_{\substack{x_i \in R(x); \\ x_i < x_{i+1}}} x_i p_X(x_i)$ $(\sum_{n=0}^{\infty} G_X(n))$
<b>Continuous</b> $R(Y) = \text{supp}(Y)$ $R(\overset{\circ}{Y}) \neq \emptyset, \text{ e.g. } (a, b)$	<b>Continuous <math>F(y)</math></b> increase from 0 to 1 <b>No jumps</b> differentiable	$f_Y(y) = F'(y)$ $= \lim_{s \rightarrow y} \frac{F(y) - F(s)}{y - s}$ $F(y) = \int_{s < y} f_Y(s) ds$ $f_Y(y) \geq 0; \int_{R(Y)} f_Y(y) dy = 1$	$\int_{R(Y)} y dF(y)$ $\int_{R(Y)} y f(y) dy$ $\left( \int_0^\infty G_Y(y) dy \right)$

Let  $X$  be a **non-negative integer-valued** r.v. then if its expectation exists

$$\mathbb{E}X = \sum_{i=0}^{\infty} G_X(i)$$

where  $G_X$  is the survival function of  $X$  such that  $G_X(x) = 1 - F_X(x)$ .

► Proof:

$$X = I_{\{X \geq 1\}} + \cdots + I_{\{X \geq n\}} + \cdots = \sum_{i \in \mathbb{N}_+} I_{\{X \geq i\}}.$$

The above holds true since  $X$  and  $\sum_{i \in \mathbb{N}_+} I_{\{X \geq i\}}$  are the same function from  $S$  to the set of non-negative integers.

Linearity of expectation, fundamental bridge, and the fact that

$\{X \geq i\} = \{X > i - 1\}$ ,  $i \in \mathbb{N}_+$  give

$$\begin{aligned} EX &= E \sum_{i=1}^{\infty} I_{\{X \geq i\}} = \\ &\sum_{i=1}^{\infty} EI_{\{X \geq i\}} = \sum_{i=1}^{\infty} P(X \geq i) = \\ &\sum_{i=0}^{\infty} P(X > i) = \sum_{i=0}^{\infty} G_X(i). \end{aligned}$$

Let  $X$  be a **non-negative REAL-valued** (either discrete or continuous) r.v. then if its expectation exists

$$\mathbb{E}X = \int_0^{\infty} G_X(s)ds$$

where  $G_X$  is the survival function of  $X$  such that  $G_X(x) = 1 - F_X(x)$ .

► Proof:

$$X = \int_0^{\infty} I_{\{X > t\}} dt.$$

The above holds true since  $X$  and  $\int_0^{\infty} I_{\{X > t\}} dt$  are the same function from  $S$  to the set of non-negative real values:

when  $X(s) = x$ ,

$$\int_0^{\infty} I_{\{X(s) > t\}} dt = \int_0^{\infty} I_{\{x > t\}} dt = x$$

since  $I_{\{x > t\}} = 1$  only for  $t \in [0, x]$ .

Linearity of expectation, fundamental bridge give

$$\begin{aligned} EX &= E \int_0^{\infty} I_{\{X > t\}} dt = \\ &\int_0^{\infty} EI_{\{X > t\}} dt = \int_0^{\infty} P(X > t) dt \\ &= \int_0^{\infty} G_X(s)ds. \end{aligned}$$

## Discussions via problems

## I. what is the probability of $\{Z = z\}$ ?

$X$  follows  $\text{Bern}(0.5)$  with  
 $p_X(0) = p_X(1) = 1/2$ .

- ▶ The support of  $X$  is  $\{0, 1\}$  (**Finitely many**, at most countably infinite).
- ▶  $P(X = x) = 1/2$  for  $x \in \{0, 1\}$ , 0 otherwise.

$X$  has  $f_X(x) = 1, x \in [0, 1]$ .

- ▶ The support of  $X$  is  $[0, 1]$  (**Infinitely many**, uncountably infinite).
- ▶  $P(X = x) = 0$  for  $x \in \mathbb{R}$ . Since  

$$P(X = x) = \lim_{\delta \downarrow 0} \int_{x-\delta}^{x+\delta} f_X(s) ds = 0.$$

Uncountably many possible values, so it is of zero probability that you choose a specific number. True for all continuous r.v.'s

However, you could get outcomes in a neighborhood closer to the number  $B_\epsilon(x) = (x - \epsilon, x + \epsilon) \subseteq [0, 1]$  ( $\epsilon$  is an arbitrary and a very small positive number):

$$\begin{aligned} P(X \in (x - \epsilon, x + \epsilon)) \\ = \int_{x-\epsilon}^{x+\epsilon} f_X(s) ds = 2\epsilon. \end{aligned}$$

I. **uncountably many possible values, so it is of zero probability that you get a specific number. True for all continuous r.v.'s**

## II. what is the probability of $\{Z_1 < Z_2 < Z_3\}$ ?

$X_i, i = 1, 2, 3$  follows i.i.d.

Bern(1/2) with mass (PMF)

$$p(0) = p(1) = 1/2.$$

$X_i$  i.i.d. with density (PDF)

$$f(x) = 1, x \in [0, 1].$$

- ▶  $P(Z_1 < Z_2 < Z_3) = 0.$
- ▶ At least two of them would be equal to each other.
- ▶ First note that  $X_i - X_j, i \neq j$  is also a continuous r.v. and thus  $P(X_i = X_j) = P(X_i - X_j = 0) = 0.$
- ▶ Any two of them will be equal to each other with zero probability.
- ▶ Then one of the arbitrary order must happen  $X_{a_1} < X_{a_2} < X_{a_3}$  for any permutations  $a_1, a_2, a_3$  of 1, 2, 3, and by symmetry

$$P(X_{a_1} < X_{a_2} < X_{a_3}) = \frac{1}{3!}.$$

Can be extended to cases of  $n$  i.i.d. continuous r.v.'s.

$$P(X_{a_1} < X_{a_2} < X_{a_3} < \dots < X_{a_n}) = \frac{1}{n!}.$$

**II.  $X_i$  i.i.d. from a continuous distribution, then**

$$P(X_{a_1} < X_{a_2} < X_{a_3} < \cdots < X_{a_n}) = \frac{1}{n!}.$$

**for any permutations  $a_1, a_2, a_3, \dots, a_n$  of  $1, 2, 3, \dots, n$ .**

# Probability Theory for EOR

Some special continuous random variables

I (Exponential)

Some **continuous random variables** are associated very special/ubiquitous **distributions (PDFs)**, they get their own names!

Definition (PDF of continuous real-valued r.v.)

The **probability density function (PDF)** of a **continuous** real-valued r.v. is a non-negative function  $f_X$  on the real line such that via the Riemann integral:

$$\int_{-\infty}^x f_X(s)ds = P(X \leq x).$$

For a **continuous** random variable with differentiable CDF  $F_X$ , conventionally,  $f_X(x) = F'_X(x)$ .

## Exponential distribution

●○○○○○○○○○○○○○○○○

# Exponential distribution

# Waiting time

Customers come in according to a Poisson process such that:

*Poisson arrivals.*

The number of arrivals that occur in an interval of length  $t$  is a  $\text{Pois}(\lambda t)$  r.v.;

*Independence condition.*

The number of arrivals that occur in disjoint intervals are independent from each other.

**What is the distribution of the waiting time until the first customer?**

## Waiting time

**What is the distribution of the waiting time until the first customer?**

- ▶ Denote  $N_t$  the number of arrivals within  $[0, t]$ , then  $N_t$  follows  $\text{Pois}(\lambda t)$ .
- ▶ Denote  $T_1$  the time until the first arrival.
- ▶ Note that  $\{T_1 > t\} = \{N_t = 0\}$ .
- ▶  $P(T_1 > t) = P(N_t = 0) = e^{-\lambda t}$ .
- ▶ The CDF of  $T_1$  then

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - e^{-\lambda t}.$$

- ▶ **CDF has non-zero first derivative for  $t > 0$  such that CDF can be rewritten as a Riemann integral of the derivative**

$$f_{T_1}(t) = F'_{T_1}(t) = \lambda e^{-\lambda t}.$$

## Exponential distribution: $\text{Expo}(\lambda)$

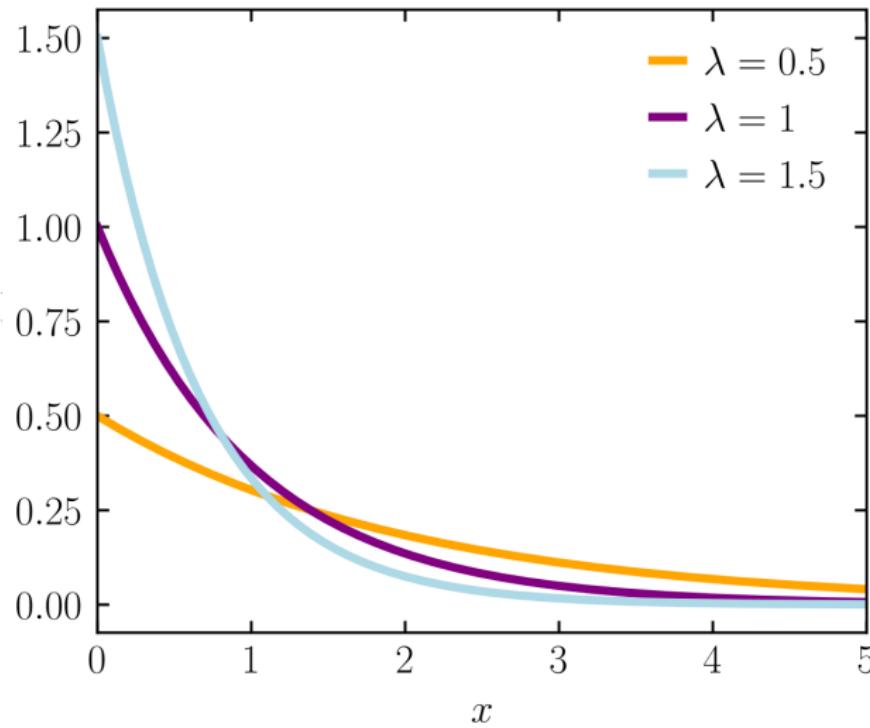
- A continuous real-valued r.v.  $X$  follows  $\text{Expo}(\lambda)$  if its PDF  $f_X(x)$  is

$$f_X(x) = \lambda e^{-\lambda t}, t > 0$$

and zero for  $t \leq 0$ .

- CDF:  $F_X(t) = 1 - e^{-\lambda t}$ .
- Note that if  $X \sim \text{Expo}(1)$ , then  $Y = \frac{X}{\lambda} \sim \text{Expo}(\lambda)$ .

$$P(Y \leq t) = P(X \leq \lambda t) = 1 - e^{-\lambda t}.$$

PDF  $\text{Expo}(\lambda)$ 

## Expectation and Variance

$$X \sim \text{Exp}(1), Y = \frac{X}{\lambda} \sim \text{Exp}(\lambda).$$

- The expectation  $EX$ .

$$E[X] = \int_0^{\infty} xe^{-x} dx = 1.$$

- The variance  $VX$ .

$$E[X^2] = \int_0^{\infty} x^2 e^{-x} dx = 2.$$

$$V[X] = E[X^2] - (E[X])^2 = 1.$$

- $EY = E\left(\frac{X}{\lambda}\right) = 1/\lambda, VY = V\left(\frac{X}{\lambda}\right) = 1/\lambda^2.$

**I. From Poisson process to exponential:  
waiting time between the first and the second arrivals is also Expo**

Customers come in according to a Poisson process such that:

*Poisson arrivals.*

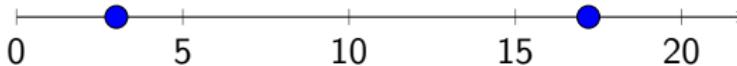
The number of arrivals that occur in an interval of length  $t$  is a  $\text{Pois}(\lambda t)$  r.v.;

*Independence condition.*

The number of arrivals that occur in disjoint intervals are independent from each other.

Denote  $T_i$  the waiting time until the  $i$ th arrivals.

## The distribution of $T_2 - T_1$ .



- ▶ Suppose at time  $t$ , there is one visit, what is the time of next visitor?
  - \* We re-start to wait and as if we are waiting for the *first* arrival.
  - \*  $(0, T_1)$  and  $(T_1, +\infty)$  are disjoint intervals and thus arrivals in these intervals are independent.
- ▶ Therefore,  $T_2 - T_1 \sim \text{Expo}(\lambda)$ .  
**Results can be generalised to the waiting time between the  $i$ th and the  $(i+1)$ th arrivals.**
- ▶ However,  $T_2$  does not follow exponential distribution: the sum of independent Expo's is not Expo.

Exponential distribution

○○○○○○○○●○○○○○○

Special properties of Exponential

## II. Memoryless property

## Definition

### Memoryless property

$$\mathbb{P}(X - s \geq t | X \geq s) = \mathbb{P}(X \geq t)$$

If  $X$  is a positive continuous real-valued random variable with the memoryless property, then  $X$  has an exponential distribution.

- ▶ Let  $G(x) = 1 - F(x)$ , then memoryless implies

$$G(s+t) = G(s)G(t).$$

- ▶ (1)  $G(mt) = (G(t))^m$ .
- ▶ (2)  $G(t/n) = (G(t))^{1/n}$ .
- ▶ (3)  $G(m/nt) = (G(t))^{m/n}$ .
- ▶ (4)  $G(xt) = (G(t))^\times$  and thus  $G(x) = (G(1))^\times = e^{-\log(G(1))x}$
- ▶ The CDF of  $X$  would be the exponential distribution CDF ( $\text{Expo}(\log(G(1)))$ ):  $F(x) = 1 - e^{-\log(G(1))x}$ .
- ▶ One discrete r.v. also satisfies the memoryless property: Geometric!

Exponential distribution

○○○○○○○○○○●○○○

Special properties of Exponential

### III. Minimum of independent Expo's is still Expo

A web server receives requests from independent computers  $1, \dots, n$ ,  
**the number of requests from the computer  $i$  follows Poisson process with unit arrival rate  $\lambda_i$ .**

**the waiting time before the first request from computer  $i$ ,  $T_i$ , follows  $\text{Expo}(\lambda_i)$ .**

What is the distribution of  $T = \min_{1 \leq i \leq n} T_i$  ?

- ▶ Note that the number of total requests in a time interval of length  $t$ ,  $N_t$ , follows Poisson distribution with arrival rate  $\left(\sum_{1 \leq i \leq n} \lambda_i\right) t$ .
- ▶ Therefore, the waiting time before any requests would follow  $\text{Expo}\left(\left(\sum_{1 \leq i \leq n} \lambda_i\right) t\right)$ .
- ▶

$$F_T(t) = 1 - P(T > t) = 1 - P(N_t = 0) = 1 - e^{-\left(\sum_{1 \leq i \leq n} \lambda_i\right)t}.$$

Exponential distribution

○○○○○○○○○○○○●○

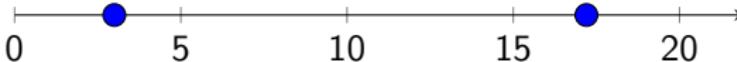
Special properties of Exponential

## IV. From Exponential to Poisson

The economy starting from time 0 follows a business cycle which switches between peak and trough periods, suppose each period lasts for a random time length following i.i.d.  $\text{Expo}(\lambda)$ .

**What is the distribution of the number of transitions from one status to another in the time interval  $(0,t)$ ?**

- It is like the economy would receive a signal, such that receiving the signal would change economic status.



Time duration between each successive signals follows i.i.d.  $\text{Expo}(\lambda)$ .

- Therefore, the number of signals follow  $\text{Pois}(\lambda t)$ .

The waiting time between two Poisson process arrivals follows independent exponential distributions, while if the waiting time between two successive arrivals follows i.i.d. exponential distributions then the number of arrivals follows a Poisson process.

# Probability Theory for EOR

Some special continuous random variables  
II (Uniform)

Some **continuous random variables** are associated very special/ubiquitous **distributions (PDFs)**, they get their own names!

Definition (PDF of continuous real-valued r.v.)

The **probability density function (PDF)** of a **continuous** real-valued r.v. is a non-negative function  $f_X$  on the real line such that via the Riemann integral:

$$\int_{-\infty}^x f_X(s)ds = P(X \leq x).$$

For a **continuous** random variable with differentiable CDF  $F_X$ , conventionally,  $f_X(x) = F'_X(x)$ .

## Uniform distribution: $\text{Unif}((a, b))$

- A continuous real-valued r.v.  $U$  is said to have the uniform distribution on  $(a, b)$  (or follows  $\text{Unif}((a, b))$ ), if its PDF is

$$f_U(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{otherwise} \end{cases}$$

- $U$  takes values in  $(a, b)$ .
- Probability is proportional to length (e.g.,  $(c, d) \subseteq (a, b)$ ):

$$P(c \leq U \leq d) = \int_c^d \frac{1}{b-a} dx = \frac{c-d}{b-a}.$$

If  $U$  follows  $\text{Unif}(0,1)$ , what is the distribution of  $X = a + (b - a)U$ ?

- ▶  $\text{Unif}(a,b)$ .
- ▶ Verify by CDF ( $x \in (a, b)$ ):

$$\begin{aligned}F_X(x) &= P(a + (b - a)U \leq x) \\&= P(U \leq (x - a)/(b - a)) = (x - a)/(b - a).\end{aligned}$$

## Expectation and Variance of $\text{Unif}(a,b)$

$$Y \sim \text{Unif}(a,b), X \sim \text{Unif}(0,1)$$

- The expectation  $EX$ .

$$\mathbb{E}[X] = \int_0^1 x dx = 1/2.$$

- The variance  $VX$ .

$$\mathbb{E}[X^2] = \int_0^1 x^2 dx = 1/3.$$

$$V[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1/12.$$

- $EY = E(a + (b-a)X) = a + \frac{b-a}{2} = \frac{a+b}{2}, VY = V(a + (b-a)X) = (b-a)^2/12.$

## I. Conditional on a unif is a unif

**Let  $U \sim \text{Unif}(a,b)$ ,  $(c, d) \subseteq (a, b)$ . Then the conditional distribution of  $U$  given  $U \in (c, d)$  is  $\text{Unif}(c,d)$ .**

- ▶ Proof: for  $x \in (c, d)$

$$\begin{aligned} & P(U \leq x | U \in (c, d)) \\ &= P(U \in (c, x]) / P(U \in (c, d)) = \frac{x - c}{d - c} \end{aligned}$$

which is exactly the CDF of  $\text{Unif}(c,d)$ .

## **II. Universality of the Unif**

**Let  $F$  be a strictly increasing (so  $F^{-1}$  the inverse function is well defined) CDF of a continuous r.v., we then have**

- Let  $U \sim \text{Unif}(0,1)$ ,  $F^{-1}(U)$  is an r.v. with CDF  $F$ .

*Proof:*

$$\begin{aligned} F_{F^{-1}(U)}(x) &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) = F(x) \end{aligned}$$

- Let  $X$  be an r.v. with CDF  $F$ , then  $F(X) \sim \text{Unif}(0, 1)$ .

*Proof:*

$$\begin{aligned} F_{F(X)}(x) &= P(F(X) \leq x) \\ &= P(X \leq F^{-1}(x)) = F(F^{-1}(x)) = x. \end{aligned}$$

If you have a uniformly distributed random variable you can generate many random variables: e.g.,  $\log\left(\frac{1}{1-U}\right) \sim \text{Expo}(1)$ .

# Probability Theory for EOR

Some special continuous random variables

III (Normal)

Part A

Some **continuous random variables** are associated very special/ubiquitous **distributions (PDFs)**, they get their own names!

Definition (PDF of continuous real-valued r.v.)

The **probability density function (PDF)** of a **continuous** real-valued r.v. is a non-negative function  $f_X$  on the real line such that via the Riemann integral:

$$\int_{-\infty}^x f_X(s)ds = P(X \leq x).$$

For a **continuous** random variable with differentiable CDF  $F_X$ , conventionally,  $f_X(x) = F'_X(x)$ .

**Normal/Gaussian!! Part A.**

Normal distribution

●○○○○○○

## Normal distribution

- ▶ Suppose there is a super server receiving many requests from independent computers  $1, \dots, n$ , the number of requests from each computer within a time interval of unit length,  $N_i, i = 1, \dots, n$ , would follow i.i.d. Pois(1).
- ▶ One is wondering about the distribution of the number of the total requests when  $n$  is getting larger and larger, and decided to look at the following random variable (a rescaled of the total requests shifted by the average):
 
$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i - EN_i),$$

which is  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i - 1)$ .

- ▶ Note that  $\sum_{i=1}^n N_i$  is a sum of i.i.d. Pois(1)-distributed r.v.'s, which is still a Poisson distribution (Pois(n)), and thus we can calculate the CDF function of the term  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i - 1)$ :

$$P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i - 1) \leq m\right) = P\left(\sum_{i=1}^n N_i \leq n + \sqrt{n}m\right) = \sum_{i=0}^{\lfloor n + \sqrt{n}m \rfloor} e^{-n} \frac{n^i}{i!},$$

which surprisingly has a limit as  $n \rightarrow \infty$  :

$$\int_{-\infty}^m \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds.$$

- ▶ The limit is true not only for  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i - EN_i)$  but also valid for more general choices of  $X_i$ 's (zero mean and unit variance):  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - EX_i)$ .

## Normal distribution: $N(\mu, \sigma^2)$

- A continuous r.v.  $Z$  is said to have the *standard normal distribution*  $N(0, 1)$  (mean zero and variance one) if its PDF  $\psi$  is given by

$$\psi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

with corresponding CDF  $\Psi(z) = \int_{-\infty}^z \psi(z) dz$ .

- As for  $X = \mu + \sigma Z$  ( $\sigma > 0$ ), it is said to have the *normal distribution*  $N(\mu, \sigma^2)$  with mean  $\mu$  and variance  $\sigma^2$ .

The PDF of  $X$ :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $F_X(x) = P(\mu + \sigma Z \leq x) = \Psi\left(\frac{x-\mu}{\sigma}\right)$ ,  $f_X(x) = F'_X(x) = \psi\left(\frac{x-\mu}{\sigma}\right) \frac{1}{\sigma}$ .

## $\psi(z)$ is a proper PDF

- ▶ Non-negative.
- ▶ Integrate to one:

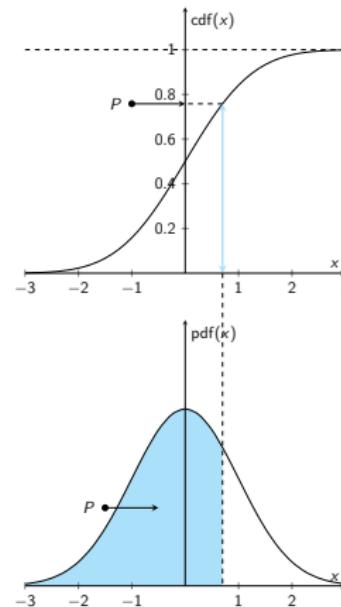
$$\begin{aligned} \left( \int_{-\infty}^{\infty} \psi(z) dz \right)^2 &= \left( \int_{-\infty}^{\infty} \psi(x) dx \right) \left( \int_{-\infty}^{\infty} \psi(y) dy \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi(x)\psi(y) dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} dxdy \\ &= \int_0^{2\pi} \int_0^{\infty} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r dr d\theta = \int_0^{2\pi} \frac{1}{2\pi} d\theta = 1 \end{aligned}$$

Normal distribution

○○○○●○○

Basic properties of Normal

## Standard Normal Distribution $N(0, 1)$ CDF, PDF



- Symmetry of PDF ( $\psi(z) = \psi(-z)$ ); of CDF ( $\Psi(z) = 1 - \Psi(-z)$ ); of Z and -Z.

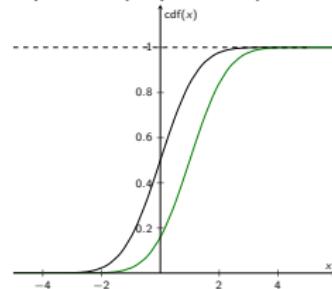
Normal distribution

○○○○●○

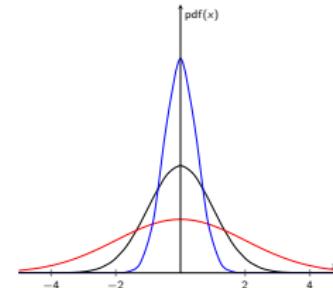
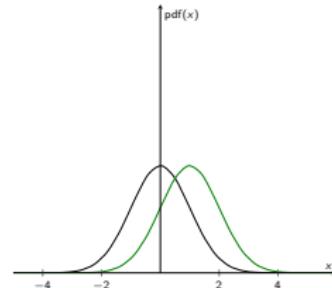
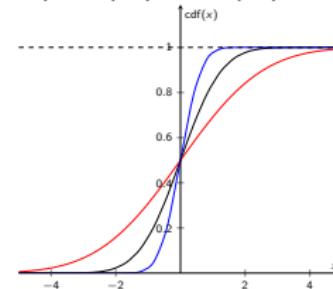
Basic properties of Normal

## Normal Distribution $N(\mu, \sigma^2)$ CDF, PDF

$\sigma = 1, \mu =$  (black 0), (green 1)



$\mu = 0, \sigma =$  (red 2), (black 1), (blue 1/2)



# Expectation and Variance

$$Z \sim N(0, 1), X \sim N(\mu, \sigma^2).$$

- The expectation  $EZ$ .

$$E[Z] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.$$

- The variance  $VZ$ .

$$\begin{aligned} E[Z^2] &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 2 \int_0^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \frac{2}{\sqrt{2\pi}} \left( -xe^{-x^2/2} \Big|_0^\infty + \int_0^\infty e^{-\frac{x^2}{2}} dx \right) \\ &= \frac{2}{\sqrt{2\pi}} \left( \cancel{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right) = \frac{2}{\sqrt{2\pi}} \left( \cancel{\sqrt{2\pi}} \frac{1}{2} \right) \\ &= 1. \end{aligned}$$

$$V[Z] = E[Z^2] - (E[Z])^2 = 1.$$

- $EX = E(\mu + \sigma Z) = \mu, VX = V(\mu + \sigma Z) = \sigma^2$ .

Layman's talk : what is moment generating function

A real-valued r.v.  $X$  and its nth moment  $\mathbb{E}X^n$ .

Can we find something to store the information of a sequence of moments?

	$\mathbb{E}X^0$	$\mathbb{E}X$	$\mathbb{E}X^2$	$\mathbb{E}X^3$	$\mathbb{E}X^4$	$\dots$
	$t^0/0!$	$t^1/1!$	$t^2/2!$	$t^3/3!$	$t^4/4!$	$\dots$

We use polynomials to store these moment values, label each moment with a unique polynomial!

$$M_X(t) = \sum_{i=0}^{\infty} E(X^i)t^i/i! = E \sum_{i=0}^{\infty} (X^i)t^i/i! = Ee^{tX}.$$

**Moment generating function (MGF)  $M_X(t)$  is a function.**

We say it is well-defined, if we can find one  $a > 0$  such that  
 $M_X(t) : (-a, a) \mapsto \mathbb{R}$ .

- Moments via the derivative of the MGF:

$$EX^n = M_X^{(n)}(0)$$

- MGF (if exists) determines the distribution type:

If two r.v.s have the same MGF, they have the same distribution!

# Probability Theory for EOR

## Moments

The  $n$ th moment of  $X$ :  $\mathbb{E}X^n$

●○○○

Sample moments

○○○○○○

The  $n$ th moment of  $X$ :  $\mathbb{E}X^n$

For a real-valued r.v.  $X$  (assume the following expectation values exist):

►  **$n$ th Moment**

$$\mathbb{E}X^n.$$

►  **$n$ th Central Moment**

$$\mathbb{E}(X - \mu)^n.$$

►  **$n$ th Standardized Moment**

$$\mathbb{E}((X - \mu)/\sigma)^n.$$

► **Distributions determine moments** (**moments are numbers**, or you may regard them as parameters of fixed numbers for given random variables) since they are defined based on the concept: expectation.

- Central/standarized moments are linear combinations of moments,  
e.g.,

	$\mathbb{E}X^0$	$\mathbb{E}X$	$\mathbb{E}X^2$	$\mathbb{E}X^3$	$\mathbb{E}X^4$	...
$\mu: \mathbb{E}X$	0	1	0	0	0	0
$\sigma^2: \mathbb{E}X^2 - (\mathbb{E}X)(\mathbb{E}X)$	0	$-\mu$	1	0	0	0
$\text{Skew}(X): \mathbb{E} \left( \frac{X-\mu}{\sigma} \right)^3$	$-\frac{\mu^3}{\sigma^3}$	$\frac{3\mu^2}{\sigma^2}$	$-\frac{3\mu}{\sigma}$	1	0	0
...	...					

- Moments describe the shapes of distributions: e.g., mean.

Moments are quite useful. Let's how to use random draws/obeserved data to get estimates (approximations) for our moments/parameters.

The  $n$ th moment of  $X$ :  $\mathbb{E}X^n$

○○○○

Sample moments

●○○○○○

## Sample moments

Toss a coin many many ( $n$ ) times, denote  $X_i$  which equals one if the  $i$ th toss is a head otherwise zero for a tail,

**the sample average (first sample moment) value**  $\frac{1}{n} \sum_{i=1}^n X_i$  is a random variable, but from our daily experience, we know it is more likely to take values in a smaller and smaller neighborhood of the successful rate of getting a head when  $n$  grows,  $p$ ,

which is the **1st moment**  $EX_i$ .

For each time, we throw a fair coin for  $n$  tosses, and calculate one number: the proportion of heads (one realisation of the sample average), we repeat this for 2000 times. Then we have 2000 numbers, we draw the histogram (density) plot of these 2000 numbers.

As  $n$  increases, we see the realised sample averages (random realisations) are centering towards  $p = 1/2$ .

We can generalise this results to all moments:  
from the sample values to theoretical values!

We focus on the i.i.d. sequences.

Let  $X_i, i = 1, \dots, n$  be i.i.d. r.v.'s;

**The  $k$ th sample moment:**

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

- E.g., the sample mean is  $M_1$  (the first sample moment)

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

- **Sample moments serve as approximations (observed from realized data) to the moments (theoretical values).**

**Sample moments serve as approximations (observed from realized data) to the moments (theoretical values).**

Let  $X_i, i = 1, \dots, n$  be i.i.d. r.v.'s with  $\mu, \sigma^2$ .

- **Sample mean**  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ .

By the linearity of expectation,  $\mathbb{E}\bar{X}_n = \mu$ ;

due to the independence,  $\text{Var}\bar{X}_n = \sigma^2/n \rightarrow 0$ . (The only random variable type with zero variance is the degenerated random variable: constants.)

- $X_i^k, i = 1, \dots, n$  are also i.i.d. r.v.'s for any non-negative integer  $k$ , so the above results also hold true if  $X_i^k$  has first and second moments.

- **Sample variance**  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

- From  $\sum_i^n (X_i - c)^2 = \sum_i^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - c)^2$ , we know  $\mathbb{E}S_n^2 = \sigma^2$ .

Median and mode

ooo

Symmetry and skewness

ooooo

Tail and Kurtosis

ooooo

# Probability Theory for EOR

## How to describe the distribution

Median and mode

○○○

Symmetry and skewness

○○○○○

Tail and Kurtosis

○○○○○

Distributions determine probability values of events generated by random variables. We distinguish random variables by their distributions.

How do we describe distributions: e.g., **the shape of PDF/PMF?**

Median and mode

●○○

Symmetry and skewness

○○○○○

Tail and Kurtosis

○○○○○

## Median and mode

## Median.

- ▶  $c$  (**may not be unique for a given r.v.**) is a median if  $\mathbb{P}(X \leq c) \geq 1/2$  and  $\mathbb{P}(X \geq c) \geq 1/2$ .
- ▶ Median  $c$  is different from mean/expectation  $\mu$ .
  - ▶ The value that minimizes the mean squared error,  $\mathbb{E}(X - x)^2$ , is the **mean**  $\mu$ .  
 Sketch of proof:  $\mathbb{E}(X - x)^2 = \mathbb{E}(X - \mu + (\mu - x))^2 = \mathbb{E}(X - \mu)^2 + 2\mathbb{E}((X - \mu)(\mu - x)) + \mathbb{E}(\mu - x)^2 = \text{Var}(X) + (\mu - x)^2$ .
  - ▶ A value that minimizes the mean absolute error,  $\mathbb{E}|X - x|$ , is a **median**  $c$ .  
 Sketch of proof: compare  $E|X - x|$  with  $E|X - c|$  for  $x < c$  and  $x > c$ .

## Example:

- ▶ Median for  $\text{Bern}(p)$ -distributed  $X$  with  $p = 1/2$ :  
 $\forall c \in [0, 1]$ .
- ▶ Median for a constant  $c$  (degenerated random variable  $X$ ,  $P(X = c) = 1$ ):  
 $c$ .

## Mode.

- ▶ **c (may not be unique for a given r.v.)** is a mode if the PMF/PDF takes its maximum value at  $c$ .

### Example:

- ▶ Mode for  $\text{Bern}(p)$ -distributed  $X$  with  $p > 1/2$ :  
 $c = 1$ . (Unimodal distribution, a probability distribution with a PDF/PMF which has a single peak).
- ▶ Mode for  $\text{Bern}(p)$ -distributed  $X$  with  $p = 1/2$ :  
 $\{0, 1\}$ . (Multimodal distribution, a probability distribution with a PDF/PMF which has multiple peaks, here we have a bimodal case).
- ▶ Mode for a constant  $c$  (degenerated random variable  $X$ ,  $P(X = c) = 1$ ):  
 $c$ .
- ▶ Mode for  $N(\mu, \sigma^2)$ -distributed  $X$ :  

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
 takes its maximum at  $c = \mu$ .

Median and mode

○○○

Symmetry and skewness

●○○○○

Tail and Kurtosis

○○○○○

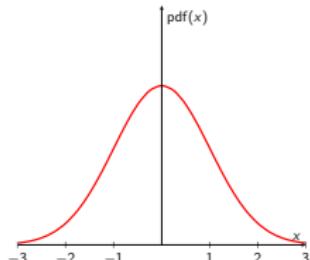
## Symmetry and skewness

## Symmetry.

- We say that an r.v.  $X$  has a symmetric distribution about  $\mu$  (or  $X$  is symmetric) if  $X - \mu$  has the same/identical distribution as  $\mu - X$ .

### Example:

- If  $Z \sim N(0, 1)$ , both  $Z$  and  $-Z$  have the identical **PDF**:



$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

- Flip one coin, betting on the head is the same as betting on the bottom. Here we may consider  $X = I_{\text{head}}$  ( $\text{Bern}(1/2)$ ).

Symmetric as **PMF** satisfies

$$P(X - \mu = k) = P(\mu - X = k).$$

*What else?*

$$\mathbb{E}(X - \mu) \equiv 0, \quad \mathbb{E}(X - \mu)^3 = 0, \quad \mathbb{E}(X - \mu)^5 = 0, \dots$$

## Skewness.

- The skewness of an r.v.  $X$  with  $\mu, \sigma^2$  is:

$$\text{Skew}(X) = \mathbb{E} \left( \frac{X - \mu}{\sigma} \right)^3$$

- We normalize by  $\sigma$  since a good measure of the symmetry should give the same number to measure the symmetry of  $X$  and  $Y = bX$ : **skewness is a third standardized moment.**

Symmetry implies that  $\text{Skew}(X) = 0$ ;

**$\text{Skew}(X) \neq 0$  implies asymmetry.**

### Example:

- Consider  $Y \sim \text{Bern}(1/2)$ ,  $\mathbb{E} \left( \frac{Y - 1/2}{1/2} \right)^3 = -1 \times 1/2 + 1 \times 1/2 = 0$ .
- **$\text{Skew}(X) = 0$  does not necessarily imply symmetry:**  
Consider a discrete r.v.  $X$  such that  $\text{supp}(X) = \{-1, -2, 3\}$  with **PMF**  $p_X(x) = 1/3, x \in \{-1, -2, 3\}$ ,  $X$  is not symmetric.

Median and mode

○○○

Symmetry and skewness

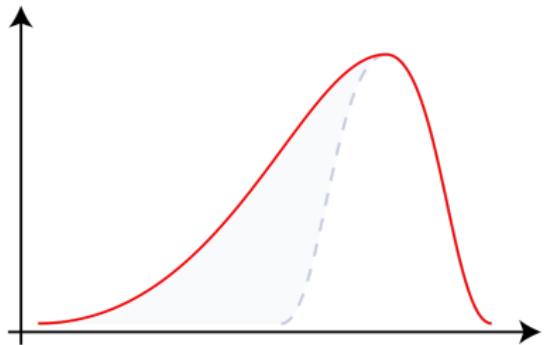
○○○●○

Tail and Kurtosis

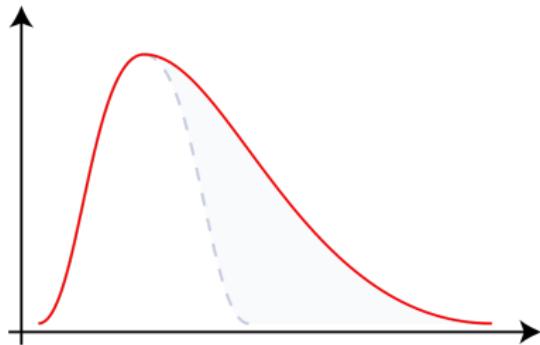
○○○○○

## Example figure I:

Figures from wikipedia (<https://en.wikipedia.org/wiki/Skewness>) on negative (left) skewed and positive (right) skewed distributions.



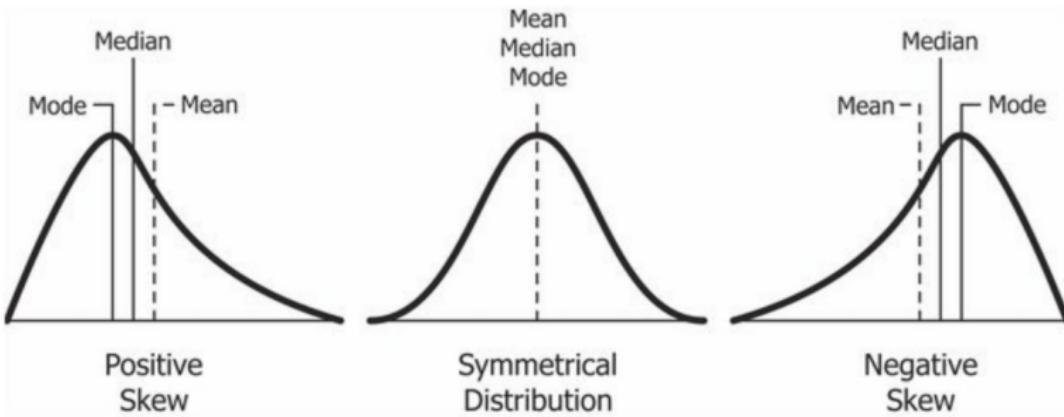
Negative Skew



Positive Skew

## Example figure II:

Figures from wikipedia (<https://en.wikipedia.org/wiki/Skewness>) on negative (left) skewed and positive (right) skewed **unimodal** distributions.



Median and mode

○○○

Symmetry and skewness

○○○○○

Tail and Kurtosis

●○○○○

## Tail and Kurtosis

## Kurtosis.

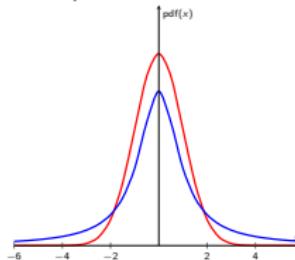
- The Kurtosis of an r.v.  $X$  with  $\mu, \sigma^2$  is (**fourth standardized moment shifted by 3**):

$$\text{Kurt}(X) = \mathbb{E} \left( \frac{X - \mu}{\sigma} \right)^4 - 3$$

- A measure of how heavy the tail of the probability distribution of a real-valued random variable is.

### Example:

- Student-t( $n$ ) PDF ( $n=5$ ) (Kurtosis is 5) and standard normal bell-shape PDF (Kurtosis is 0)



Median and mode

○○○

Symmetry and skewness

○○○○○

Tail and Kurtosis

○○●○○

## Student-t( $n$ ) PDF:

Median and mode

○○○

Symmetry and skewness

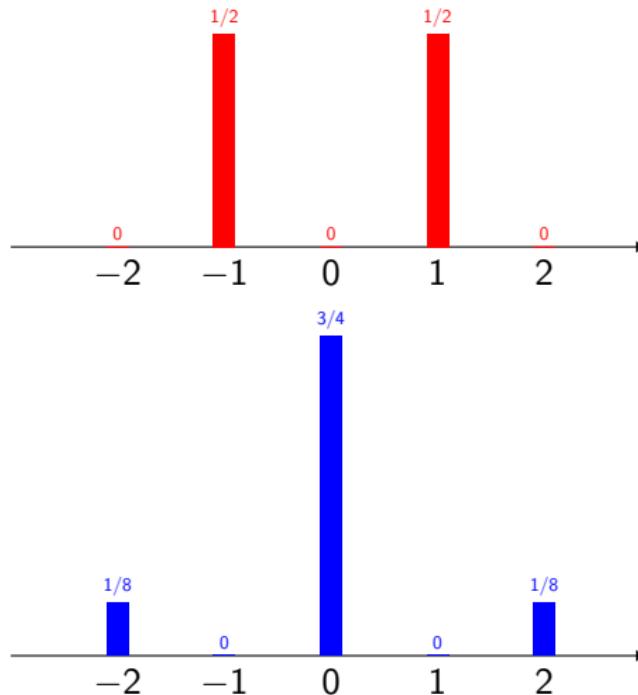
○○○○○

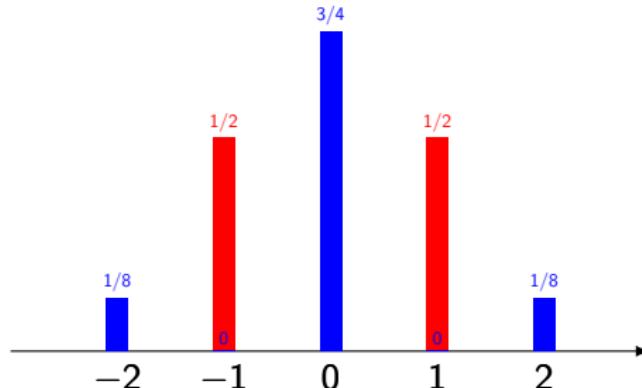
Tail and Kurtosis

○○○●○

## Example:

Fair coin v.s. unfair 3-sided die



**Example:****Fair coin v.s. unfair 3-sided die**

1. Mean the same
2. Variance the same
3. Symmetric (Skew zero)
4. \*\*  $\text{Kurt}(\text{red}) < \text{Kurt}(\text{blue})$

# Probability Theory for EOR

## Moment generating functions (MGFs)

Moments  $EX^k$  are pretty useful in characterizing random variables.  
But there are so many of them.

**Can we find something to store the information of a sequence of moments?**

$\mathbb{E}X^0$	$\mathbb{E}X$	$\mathbb{E}X^2$	$\mathbb{E}X^3$	$\mathbb{E}X^4$	$\dots$
$t^0/0!$	$t^1/1!$	$t^2/2!$	$t^3/3!$	$t^4/4!$	$\dots$

**All moments are uniquely encoded by one associated polynomial.**  
We use polynomials to store these information, label each moment with a unique polynomial! And we sum them up

$$\sum_{i=0}^{\infty} E(X^i) t^i / i!$$

Surprisingly, if the above summation is finite then it is equal to  $Ee^{tX}$ .

$$M_X(t) = Ee^{tX}.$$

## Moment generating function (MGF)

$$M_X(t) = Ee^{tX}.$$

We say it is well-defined if we can find  $a > 0$  such that

$$M_X(t) : (-a, a) \mapsto \mathbb{R}.$$

Namely,  $M_X(t)$  is only well defined if the summation below is finite for any  $t$  from some given open interval containing 0:

$$\sum_{i=0}^{\infty} E(X^i)t^i/i! = E \sum_{i=0}^{\infty} (X^i)t^i/i! = Ee^{tX}.$$

- Example (MGF for  $X \sim \text{Expo}(1)$ ):  $M_X(t) = Ee^{tX} = \int_0^{\infty} e^{tx} e^{-x} dx = \frac{1}{1-t}, \forall t \neq 1$ ,  
 $t = 0, M_X(t) = Ee^{tX} = 1$ , note that  $\frac{1}{1-t}$  is finite for, e.g.,  $t \in (-0.5, 0.5)$  so we have a well-defined MGF for  $X$ .
- Question: If  $M_Y(t) = M_X(t), \forall t$  in some open interval containing zero, what do we know about  $(\mathbb{E}(X^n))$  and  $(\mathbb{E}(Y^n))$ ?

**linearly independence of polynomials:**

$\sum_{i=0}^{\infty} a_i / i! t^i = 0, \forall t$  in some open interval containing zero if and only if  $a_i = 0$ .

So if

$$M_X(t) = \sum_{i=0}^{\infty} E(X^i) t^i / i! = \sum_{i=0}^{\infty} E(Y^i) t^i / i! = M_Y(t) = c < \infty$$

we must have

$$E(X^i) = E(Y^i), \forall i.$$

## **I. We can derive moments from the MGF (if exists).**

Not a surprise, as all moments are encoded in the MGF function by design.

We can derive moments from the MGF.

- I. taking  $n$ th derivative w.r.t.  $t$  and evaluate the  $n$ th derivative function at  $t = 0$ :

$$\mathbb{E}X^n = M_X^{(n)}(0)$$

*Proof sketch.* Notice  $M_X(t) = \sum_{i=0}^{\infty} \mathbb{E}(X^i)t^i/i!$ , then (under certain conditions)

$$M_X^{(n)}(t) = \left( \sum_{i=0}^{\infty} \mathbb{E}(X^i)t^i/i! \right)^{(n)} = \sum_{i=0}^{\infty} \left( \mathbb{E}(X^i)t^i/i! \right)^{(n)} = \mathbb{E}(X^n) + \sum_{i=n+1}^{\infty} (\mathbb{E}(X^i)t^{i-n}/(i-n)!)$$

- II. use Taylor expansion/or other expansions to write the MGF as:

$$M_X(t) = \sum_{n=0}^{\infty} a_n t^n / n!$$

then  $a_n$  is the  $n$ th moment.

## Example (moments for $X \sim \text{Exp}(1)$ ):

- $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}(X^n) t^n / n!$  (either expansion, e.g., Taylor expansion, geometric series summation..)

$$M_X(t) = \frac{1}{1-t} = \sum_{n=0}^{\infty} t^n = \sum_{n=0}^{\infty} n! \frac{t^n}{n!}$$

- $M_X(t) = \sum_{n=0}^{\infty} M_X^{(n)}(0) t^n / n!$  (or taking derivative and evaluate at 0)

$$M_X^{(n)}(t) = \left(\frac{1}{1-t}\right)^{(n)} = \frac{n!}{(1-t)^{n+1}}; M_X^{(n)}(0) = n!$$

- $\mathbb{E}X^n = n!$
- Easy to calculate  $VX = EX^2 - (EX)^2 = 2! - (1!)^2 = 1.$

## **II. MGFs (if exist) determine the distributions.**

If two r.v.s have the same MGF, they have the same distribution!  
CDF, PMF/PDF, MGF (if exists).

Very useful when dealing with **sum of independent r.v.'s** once you notice that for independent X and Y:  $E(XY) = E(X)E(Y)$  (a way of calculating the expectation of the product) and  $e^{X+Y} = e^X e^Y$  (a way of transforming a sum into a product).

**Independence means products:**

$$M_{X+Y}(t) = Ee^{X+Y} = Ee^X Ee^Y = M_X(t)M_Y(t)$$

In order to know the distribution of  $X + Y$ , besides CDF, PMF/PDF, we can also look at its MGF (if exists) which is a much easier way.

### Example I:

Show that a sum of two i.i.d.  $\text{Expo}(1)$ -distributed r.v.'s is not exponentially distributed.

- MGF of a  $\text{Expo}(\lambda)$ -distributed  $X$ :

$$M_X(t) = Ee^{tX} = \int_0^\infty e^{tx} e^{-\lambda x} dx = \frac{1}{\lambda - t},$$

which is finite for, e.g.,  $t \in (-\lambda/2, \lambda/2)$  (so the MGF is well-defined).

- MGF of  $Y_1 + Y_2$  with  $Y_i \sim \text{i.i.d. } \text{Expo}(1)$ :

$$M_{Y_1+Y_2}(t) = M_{Y_1}(t)M_{Y_2}(t) = \frac{1}{(1-t)^2}.$$

which is finite for, e.g.,  $t \in (-1/2, 1/2)$  (so the MGF is well-defined).

- $\frac{1}{(1-t)^2}$  can not be written in the form of  $\frac{1}{\lambda-t}$  for any  $\lambda$ 's. We prove by contradiction, suppose they are equal for some  $\lambda$  then  $\left. \frac{1}{(1-t)^2} \right|_{t=0} = \left. \frac{1}{\lambda-t} \right|_{t=0}$  and thus  $\lambda = 1$ , however,  $\left. \frac{1}{(1-t)^2} \right|_{t=a} \neq \left. \frac{1}{1-t} \right|_{t=a}$  for any  $a \in \{a \neq 0, |a| < \}$ .
- The MGF of  $Y_1 + Y_2$  is not the MGF of a exponential distribution, then we know a sum of two i.i.d.  $\text{Expo}(1)$ -distributed r.v.'s is not exponentially distributed as MGF determines distributions (different MGF forms, different distributions).

## Example II:

$\text{Bin}(n, \lambda/n)$  converges to  $\text{Pois}(\lambda)$  with  $n \rightarrow \infty$ .

- Derive the corresponding MGFs:

$X_i \sim \text{i.i.d. Bern}(p)$ ,

$$M_{X_i}(t) = p(e^t - 1) + 1, t \in \mathbb{R};$$

$Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ ,

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = (M_{X_1}(t))^n = (1 + p(e^t - 1))^n, t \in \mathbb{R}.$$

$Z \sim \text{Pois}(\lambda)$ ,

$$M_Z(t) = \sum_{n=0}^{\infty} e^{tn} e^{-\lambda} \lambda^n / n! = e^{-\lambda} \sum_{n=0}^{\infty} (\lambda e^t)^n / n! = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}, t \in \mathbb{R}.$$

- Note that  $\lim_{n \rightarrow \infty, p=\lambda/n} (1 + p(e^t - 1))^n = \lim_{n \rightarrow \infty} \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n = e^{\lambda(e^t - 1)}$ .
- The MGFs are the same in the limit, which implies in the limit these two distributions coincide.

**Example III:**

Show that a sum of  $n$  independent  $\text{Pois}(\lambda_i)$ ,  $i \leq n$  is still Possion.

- MGF of a  $\text{Pois}(\lambda)$ -dsitributed  $X$ :

$$M_X(t) = Ee^{tX} = e^{\lambda(e^t - 1)},$$

which is finite for, e.g.,  $t \in (-\lambda/2, \lambda/2)$  (so the MGF is well-defined).

- MGF of  $\sum_{i=1}^n Y_i$  with  $Y_i \sim \text{independent Pois}(\lambda_i)$ :

$$M_{\sum_{i=1}^n Y_i}(t) = \prod_{i=1}^n e^{\lambda_i(e^t - 1)} = e^{\sum_{i=1}^n \lambda_i(e^t - 1)}.$$

which is finite for, e.g.,  $t \in (-1/2, 1/2)$  (so the MGF is well-defined) and is the MGF of a  $\text{Pois}(\sum_{i=1}^n \lambda_i)$ -dsitributed  $X$ .

- The MGF of  $\sum_{i=1}^n Y_i$  is the MGF of  $\text{Pois}(\sum_{i=1}^n \lambda_i)$ , then we know a sum of  $n$  independent  $\text{Pois}(\lambda_i)$ ,  $i \leq n$  is still Possion.

# Probability Theory for EOR

Some special continuous random variables

III (Normal)

Part B

Some **continuous random variables** are associated very special/ubiquitous **distributions (PDFs)**, they get their own names!

Definition (PDF of continuous real-valued r.v.)

The **probability density function (PDF)** of a **continuous** real-valued r.v. is a non-negative function  $f_X$  on the real line such that via the Riemann integral:

$$\int_{-\infty}^x f_X(s)ds = P(X \leq x).$$

For a **continuous** random variable with differentiable CDF  $F_X$ , conventionally,  $f_X(x) = F'_X(x)$ .

**Normal/Gaussian!! Part B.**

# Normal distribution and MGF

**Derive the MGF of  $N(\mu, \sigma^2)$ .**

- Derive  $M_Z(t)$  for  $Z \sim N(0, 1)$ .

$$M_Z(t) = \mathbb{E}e^{tZ} = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-t)^2}{2}} dz = e^{t^2/2}$$

which is finite for  $t \in \mathbb{R}$ , so the MGF,  $M_Z$ , is well defined.

- Derive  $M_X(t)$  for  $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$ .

$$M_X(t) = \mathbb{E}e^{tX} = \mathbb{E}e^{t(\mu+\sigma Z)} = e^{t\mu} \mathbb{E}e^{t\sigma Z} = e^{t\mu} \mathbb{E}e^{(t\sigma)Z} = e^{t\mu} M_Z(t\sigma) = e^{\mu t + \frac{1}{2}\sigma^2 t^2},$$

which is finite for  $t \in \mathbb{R}$ , so the MGF,  $M_X$ , is well defined.

*Remark: The relation  $M_X(t) = e^{t\mu} M_Z(t\sigma)$  is true for general local-scale transformation relation  $X = \mu + \sigma Z$ .*

Show that  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - 1)$  with  $X_i \sim \text{i.i.d. Pois}(1)$  can be nicely approximated by  $N(0, 1)$  when  $n$  grows to infinity.

- Derive the corresponding MGFs:

MGF of a Pois(1)-distributed  $X$ :

$$M_X(t) = Ee^{tX} = e^{(e^t - 1)},$$

which is finite for, e.g.,  $t \in (-1/2, 1/2)$  (so the MGF is well-defined).

MGF of  $Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - 1)$ ,

$$M_Y(t) = \prod_{i=1}^n M_{(X_i - 1)}(t/\sqrt{n}) = (e^{-t/\sqrt{n}} M_{X_1}(t/\sqrt{n}))^n = e^{n \left( e^{\frac{t}{\sqrt{n}}} - 1 - \frac{t}{\sqrt{n}} \right)},$$

which is finite for  $t \in \mathbb{R}$ , so the MGF,  $M_Y$ , is well defined.

MGF  $M_\xi(t)$  for  $\xi \sim N(\mu, \sigma^2)$ .

$$M_\xi(t) = e^{\mu t + \frac{1}{2} \sigma^2 t^2},$$

which is finite for  $t \in \mathbb{R}$ , so the MGF,  $M_\xi$ , is well defined.

- Note that  $\lim_{n \rightarrow \infty} e^{n \left( e^{\frac{t}{\sqrt{n}}} - 1 - \frac{t}{\sqrt{n}} \right)} = \lim_{n \rightarrow \infty} e^{n \left( \sum_{i=0}^{\infty} \left( \frac{t}{\sqrt{n}} \right)^i / i! - 1 - \frac{t}{\sqrt{n}} \right)} = e^{\frac{1}{2} t^2}$ .
- The MGFs of  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - 1)$  and  $N(0, 1)$  are the same in the limit, which implies in the limit these two distributions coincide.

Show that a sum of n independent  $N(\mu_i, \sigma_i^2)$ ,  $i \leq n$  is still normal.

- MGF of  $X \sim N(\mu, \sigma^2)$ .

$$M_X(t) = e^{\mu t + \frac{1}{2} \sigma^2 t^2},$$

which is finite for  $t \in \mathbb{R}$ , so the MGF,  $M_X$ , is well defined.

- MGF of  $\sum_{i=1}^n X_i$  with  $X_i \sim$  independent  $N(\mu_i, \sigma_i^2)$ ,  $i \leq n$ :

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n e^{\mu_i t + \frac{1}{2} \sigma_i^2 t^2} = e^{(\sum_{i=1}^n \mu_i)t + \frac{1}{2} (\sum_{i=1}^n \sigma_i^2)t^2}.$$

which is finite for, e.g.,  $t \in (-1, 1)$  (so the MGF is well-defined) and is the MGF of a  $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ -distributed r.v.

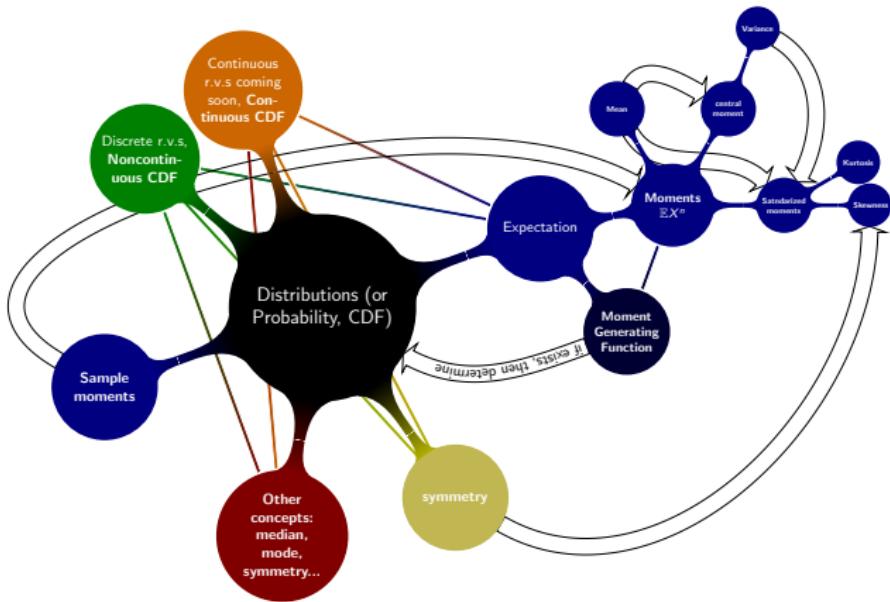
- The MGF of  $\sum_{i=1}^n X_i$  is the MGF of  $N(\sum_{i=1}^n \mu_i, \frac{1}{2} (\sum_{i=1}^n \sigma_i^2))$ , then we know a sum of n independent  $N(\mu_i, \sigma_i^2)$ ,  $i \leq n$  is still Normal/Gaussian.

# Probability Theory for EOR

Essential things you need to know from the course

From random outcomes, we have learned basic probability language:  
sample space, random events, (conditional) probability, ...

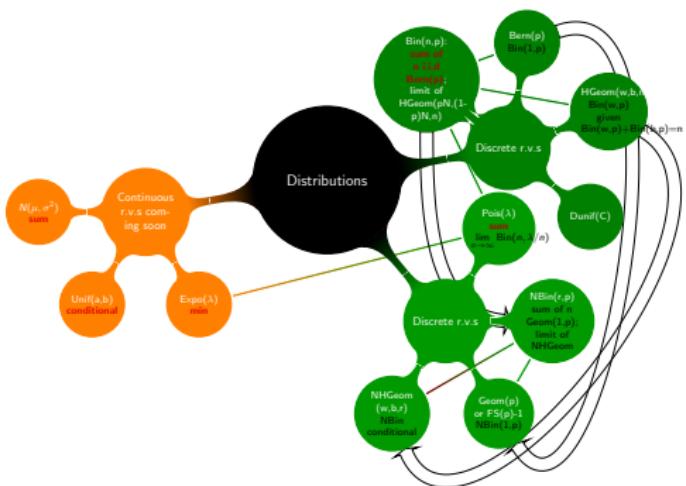
Once we have r.v.'s and their distributions, there are many derivative concepts:



We have learned many commonly used distributions with background stories for intuitions.

(1) We look at their CDFs, PDF/PMFs, MGFs (means, variances, modes, medians);

(2) Some of their transformations (sum, min, conditional...) and relations (e.g., from Pois to Expo).



It never harms to learn a bit more, but here are essential things that you need to know to pass this course.

Be able to make use of the following concepts:

1. Quantify uncertainty using a coherent framework for probability.  
*probability, sample space, outcomes, random events, random variables (indicator random variables)...*
2. Understand the concept of a conditional probability and use conditional probability as a problem solving tool.  
*conditional probability, LOTP, Bayes' rule, independence.*
3. Understand the concept of a random variable and derive properties of well-known discrete and continuous random variables.  
*random variables, and distributions (CDF, PMF/PDF), and some specific random variables and their properties (Bin, Pois, Expo, Normal, Unif).*
4. Calculate the expected value of discrete and continuous random variables.  
*From expectations (determined by distributions) we have mean, variance, moments and MGFs. The MGFs(if exist) determine distributions.*
5. Simulate and visualize the outcomes of chance experiments on your computer.  
*Simulations are experiments, which help us to uncover some interesting results.*