

# Introduction to Probability II

STAT/MATH 395 Spring 2020

Instructor: Vincent Roulet

Teaching Assistant: Zhenman Yuan



# Logistics

## Lectures

- ▶ Online lectures via *Zoom*
- ▶ Participation through (bonuses points)
  - ▶ Discussions via *Canvas* about course/homeworks
  - ▶ Quizzes at the end of each lecture, polled and discussed the next lecture via *PollEverywhere*
- ▶ Lectures recorded via *Panepo* but quizzes not available once answered
- ▶ Reading material:  
*Introduction to probability* by Anderson D., Seppäläinen T., Valkò B.

## Office hours

- ▶ Monday 10:15 to 11:15 TA office hour by *Zoom*
- ▶ Friday 11:00 to 12:00 Instructor office hour by *Zoom*
- ▶ Register in advance, questions answered in order
- ▶ Use discussions on *Canvas*

## Grading

- ▶ 8 homeworks, one per week, starting week 2
- ▶ 3 exams (weeks 4, 7 and 11) (done remotely for a duration of one day)
- ▶ All answers require **clear and detailed** mathematical explanation

**Read in detail the home page on Canvas**

**Ask any questions on the logistics on the dedicated discussion on Canvas**

# Poll Everywhere

## Log in

1. Go to PollEv.com on the sidebar, insert *my* user name, i.e. vincentroulet (true for all polls)
2. Log in with your UW credentials

## Questions/Answers

1. The questions can be
  - ▶ open questions
  - ▶ multiple choice

⋮
2. In all cases, the goal is for you to participate, *wrong answers won't matter*

## Results

1. Results will be displayed either live or at the beginning of the next lecture

# Course Content

## Content

1. Review of basics: probability distributions, random variables
  2. Jointly distributed random variables
  3. Conditional probabilities, independence, . . .
  4. Moment generating function
  5. Convergence of probability distributions, limit theorems
- ⋮

# Review on Probability Distributions, Continuous and Discrete Random Variables

STAT/MATH 395 Spring 2020

Vincent Roulet  
(Courtesy of Néhémy Lim)

# Probability Measures

## Definition (Probability space)

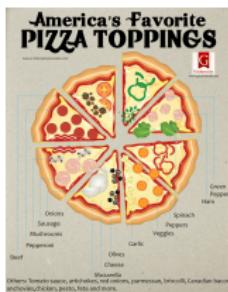
A probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  consists of three parts:

- ▶ A *sample space*  $\Omega$ , the set of all possible *outcomes* of a random action,
- ▶ A *set of events*  $\mathcal{F}$ , where each *event*  $E \in \mathcal{F}$  is a subset of  $\Omega$ ,
- ▶ A *probability measure*  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$  that assigns probabilities to events.

## Example (Cooking experiment)

Pick 5 toppings of a pizza from a total of 6.

Give the sample space  $\Omega$  of the experiment and examples of events on  $\Omega$ .



- ▶ Pepperoni
- ▶ Mozzarella
- ▶ Mushrooms
- ▶ Bacon
- ▶ Pineapple
- ▶ Green peppers

# Probability Measures

## Definition (Probability space)

A probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  consists of three parts:

- ▶ A *sample space*  $\Omega$ , the set of all possible *outcomes* of a random action,
- ▶ A *set of events*  $\mathcal{F}$ , where each event  $E \in \mathcal{F}$  is a subset of  $\Omega$ ,  
( $\mathcal{F} \subset 2^\Omega$  must be a  $\sigma$ -algebra)
- ▶ A *probability measure*  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$  that assigns probabilities to events.

## Axioms of probability

1. For all  $A \in \mathcal{F}$ ,  $0 \leq \mathbb{P}(A) \leq 1$ ,
2.  $\mathbb{P}(\emptyset) = 0$ ,  $\mathbb{P}(\Omega) = 1$
3. For any sequence  $A_1, A_2, \dots \in \mathcal{F}$  of *disjoint* sets,

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

# Random Variables

## Definition (Random variables)

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , a (real-valued) random variable (r.v.)  $X$  is defined as a mapping  $X : \Omega \rightarrow \mathbb{R}$  such that for any<sup>1</sup> subset  $B \subset \mathbb{R}$ ,

$$\{X \in B\} \triangleq \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F} \quad (1)$$

## Taxonomy of Random Variables

Random variables can be

1. discrete: it takes values in
  - ▶ a finite set, e.g.  $X(\Omega) = \{1, 2, 3\}$ ,
  - ▶ an infinite but countable set , e.g.  $X(\Omega) = \mathbb{N}$ ,
2. continuous: it takes values in a continuous set, e.g.  $X(\Omega) = [0, 1]$ .

---

<sup>1</sup>A formal definition requires  $B$  to be a Borel set of  $\mathbb{R}$ , this technicality is related to the definition of  $\sigma$ -algebra that ensures the definition of probability in continuous sets

## Discrete Random Variables

### Example

An instructor recklessly assigns a random grade (integer between 1 and 4) to his students uniformly at random

Let  $X$  be the grade of a student of his class.  
What distribution follows  $X$ ?



### Definition (Uniform random variable)

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space.  $X$  is a discrete r.v. with **uniform** distribution if it takes values on a finite set  $X(\Omega) = \mathcal{X}$ , such that

$$\mathbb{P}(X = x) = \frac{1}{\#\mathcal{X}} \quad \text{for any } x \in \mathcal{X}.$$

# Discrete Random Variables

## Example

Location of Federer's landing ball is uniform on the black box.

Let  $X$  be the score of Federer (1 or 0).

What distribution follows  $X$ ?



## Definition (Bernoulli random variable)

$X$  is a **Bernoulli** r.v. with parameter  $p \in [0, 1]$  if  $X(\Omega) = \{0, 1\}$  and

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p$$

## Discrete Random Variables

### Example

Ten people attending a match between Roger Federer and Novak Djokovic are randomly selected. A person is either a Federer fan with proba. 0.8 or a Djokovic fan with proba 0.2.

Give the sample space of this experiment.

Let  $X$  be the number of fans of Federer among the ten people. What distribution follows  $X$ ?



### Definition (Bernoulli/Binomial process)

A **Bernoulli or binomial process** with parameter  $(p, n)$  is defined by repeating  $n \in \mathbb{N}_*$  identical and independent trials of a Bernoulli r.v. with parameter  $p$ .

The associated r.v.  $X$  is the number of successes among the  $n$  trials, such that  $X(\Omega) = \{0, \dots, n\}$  and

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

# Review on Probability Distributions, Continuous and Discrete Random Variables

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 2, April 1st, 2020

# Overview

## Interactions

- ▶ Ask questions through the Zoom chat
- ▶ Answer questions via PollEverywhere (Pollev.com username: vincentroulet)

## Previous lecture

1. Model the random experiment with mathematical expressions
2. Define the random variable associated and its probability distribution

| R.V.      | Parameters                         | Value set                            | Probability Distributions                          |
|-----------|------------------------------------|--------------------------------------|--|
| Uniform   |                                    | $X(\Omega) = \mathcal{X}$ finite set | $\mathbb{P}(X=x) = 1/\#\mathcal{X}$                |
| Bernoulli | $p \in [0, 1]$                     | $X(\Omega) = \{0, 1\}$               | $\mathbb{P}(X=1) = p$                              |
| Binomial  | $n \in \mathbb{N}_*, p \in [0, 1]$ | $X(\Omega) = \{0, \dots, n\}$        | $\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)}$ |

**Not covered** (review by your own from MATH/STAT 394)

1. Rules of probability, e.g. probability of a union of non-disjoint sets
2. Conditional probability
3. Independence
4. Bayes formula and its extensions
5. Review e.g. birthday problem, its formulation and resolution

## Probability distribution

### Definition (Probability distribution of a random variable)

Let  $X$  be a r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ ,  
that is,  $X : \Omega \rightarrow \mathbb{R}$  and for any<sup>1</sup> subset  $B \subset \mathbb{R}$ ,

$$\{X \in B\} \triangleq X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{F}.$$

Denoting  $2^{\mathbb{R}} = \{B \subset \mathbb{R}\}$ , the **probability distribution** of  $X$  is the mapping

$$\mathbb{P}_X : \begin{cases} 2^{\mathbb{R}} & \rightarrow [0, 1] \\ B & \mapsto \mathbb{P}_X(B) \triangleq \mathbb{P}(\{X \in B\}) \end{cases}$$

We denote by  $X \sim \mathbb{P}_X$  the fact that "X follows a distribution  $\mathbb{P}_X$ "

### Remark

- ▶ We reduce our focus to probability spaces of the form  $(\mathbb{R}, 2^{\mathbb{R}}, \mathbb{P}_X)$
- ▶ Our goal is to
  - ▶ Analyze changes of prob. dist.  $\mathbb{P}_X$  under usual operations on r.v.
  - ▶ Generalize to multivariate r.v., i.e.,  $X = (X_1, \dots, X_d)$
  - ▶ Analyze convergence of the prob. dist.  $\mathbb{P}_X$  of r.v.

---

<sup>1</sup>A formal definition requires to restrict the subsets considered in the definition to belong to a  $\sigma$ -algebra. The canonical  $\sigma$ -algebra of  $\mathbb{R}$  is called the Borel algebra.

## Example

### Problem

In Settlers of Catan, at each round, a player rolls two dice, if your house is adjacent to an hexagon that has the same value as the dice, then you get resources.

How to choose best settling location in Catan?



### Model

The dice are 2 independent r.v.  $X_1, X_2$  with uniform prob. dist. on  $\{1, \dots, 6\}$ .

How do we describe the prob. dist. of the r.v.  $S = X_1 + X_2$  ?

Note that  $S(\Omega) = \{2, \dots, 12\}$

## Probability Mass Function

### Definition (Discrete Random variable)

A r.v.  $X$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be **discrete** if it takes values in a finite or countably infinite set  $X(\Omega)$  s.t.  $\sum_{k \in X(\Omega)} \mathbb{P}(X = k) = 1$

### Definition (Probability mass function)

Let  $X$  be a **discrete** r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The **probability mass function** (p.m.f.)  $p$  of  $X$  is defined by :

$$p : \begin{cases} X(\Omega) & \rightarrow [0, 1] \\ k & \rightarrow p(k) \triangleq \mathbb{P}(X = k) \end{cases}$$

### Example (Catan)

1. Probability mass function of  $S = X_1 + X_2$  ?

# Probability Mass Function

## Definition (Discrete Random variable)

A r.v.  $X$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is said to be **discrete** if it takes values in a finite or countably infinite set  $X(\Omega)$  s.t.  $\sum_{k \in X(\Omega)} \mathbb{P}(X = k) = 1$

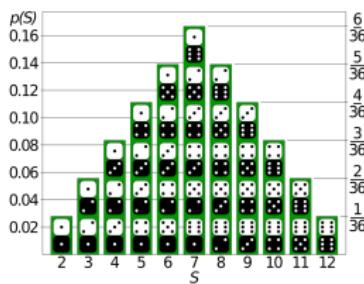
## Definition (Probability mass function)

Let  $X$  be a **discrete** r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The **probability mass function** (p.m.f.)  $p$  of  $X$  is defined by :

$$p : \begin{cases} X(\Omega) & \rightarrow [0, 1] \\ k & \rightarrow p(k) \triangleq \mathbb{P}(X = k) \end{cases}$$

## Example (Catan)

1. Probability mass function of  $S = X_1 + X_2$  ?  
→ just count frequencies



# Probability Mass Function

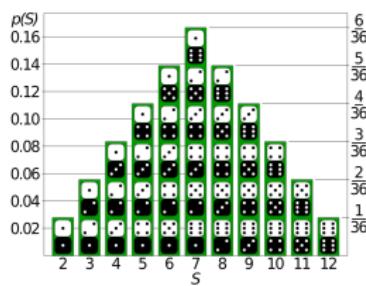
## Property

Let  $X$  be a discrete r.v. with p.m.f.  $p$ , then for any  $B \subset \mathbb{R}$ ,

$$\mathbb{P}(X \in B) = \sum_{k \in B \cap X(\Omega)} p(k)$$

## Example (Catan)

1. Compute the probability of getting resources from the location at the intersection of the hexagons 10, 8, 11



# Probability Mass Function

## Property

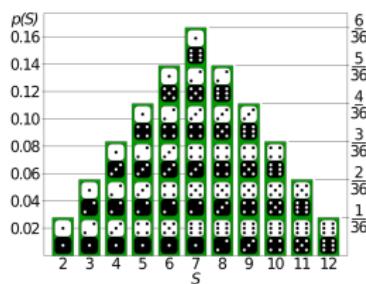
Let  $X$  be a discrete r.v. with p.m.f.  $p$ , then for any  $B \subset \mathbb{R}$ ,

$$\mathbb{P}(X \in B) = \sum_{k \in B \cap X(\Omega)} p(k)$$

## Example (Catan)

1. Compute the probability of getting resources from the location at the intersection of the hexagons 10, 8, 11

$$\begin{aligned}\mathbb{P}_S(\{8, 10, 11\}) &= \mathbb{P}(\{8\}) + \mathbb{P}(\{10\}) + \mathbb{P}(\{11\}) \\ &= \frac{5 + 3 + 2}{36} = \frac{5}{18} \approx 0.27\end{aligned}$$



## Example

You throw a bowling ball, it reaches the pin line with a uniform probability distribution on  $[-1, 1]$  where 0 is the position of the central pin.

What is the probability that the ball touches the central pin that is between  $[-0.1, 0.1]$ ?



## Probability Density Function

### Definition (Probability density function)

Let  $X$  be a r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . If a function  $f$  satisfies

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \quad \text{for any } a, b \in \mathbb{R} \cup \{-\infty, +\infty\},$$

then  $f$  is called the **probability density function** (p.d.f.) of  $X$ .

### Consequences 1

1.  $\int_{-\infty}^{+\infty} f(x)dx = 1$ ,  $f(x) \geq 0$  for all  $x \in \mathbb{R}$
2. More generally for any subset  $B \subset \mathbb{R}$ ,  $\mathbb{P}_X(B) = \int_B f(x)dx$

### Example (Bowling)

1. What is the probability that the ball touches the central pin?

# Probability Density Function

## Definition (Probability density function)

Let  $X$  be a r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . If a function  $f$  satisfies

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \quad \text{for any } a, b \in \mathbb{R} \cup \{-\infty, +\infty\},$$

then  $f$  is called the **probability density function** (p.d.f.) of  $X$ .

## Consequences 1

1.  $\int_{-\infty}^{+\infty} f(x)dx = 1, f(x) \geq 0$  for all  $x \in \mathbb{R}$
2. More generally for any subset  $B \subset \mathbb{R}$ ,  $\mathbb{P}_X(B) = \int_B f(x)dx$

## Example (Bowling)

1. What is the probability that the ball touches the central pin?

p.d.f is  $f(x) = \begin{cases} 1/2 & \text{if } x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$ , proba is given by  $\int_{-0.1}^{0.1} 1/2 dx = 0.1$

# Probability Density Function

## Consequences 2

1. Computing the probability of a r.v. on an interval amounts to a weighted **measure** of the interval

$$\mathbb{P}_X([a, b]) = \int_a^b f(x)dx$$

# Probability Density Function

## Consequences 2

1. Computing the probability of a r.v. on an interval amounts to a weighted **measure** of the interval

$$\mathbb{P}_X([a, b]) = \int_a^b f(x)dx$$

2. Manipulating r.v. with p.d.f. amounts to integral manipulations  
→ Analysis course

# Probability Density Function

## Consequences 2

1. Computing the probability of a r.v. on an interval amounts to a weighted **measure** of the interval

$$\mathbb{P}_X([a, b]) = \int_a^b f(x)dx$$

2. Manipulating r.v. with p.d.f. amounts to integral manipulations  
→ Analysis course
3. Yet, if  $X$  has a p.d.f., then  $\mathbb{P}(X = k) = \int_k^k f(x)dx = 0$   
→ Analysis course does not encompass manipulation of discrete r.v.!

# Probability Density Function

## Consequences 2

1. Computing the probability of a r.v. on an interval amounts to a weighted **measure** of the interval

$$\mathbb{P}_X([a, b]) = \int_a^b f(x)dx$$

2. Manipulating r.v. with p.d.f. amounts to integral manipulations  
→ Analysis course
3. Yet, if  $X$  has a p.d.f., then  $\mathbb{P}(X = k) = \int_k^k f(x)dx = 0$   
→ Analysis course does not encompass manipulation of discrete r.v.!

That's why we are interested in **probability distributions/measures**, i.e.,

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) \quad \text{for } B \subset \mathbb{R}$$

## Distributions/Measures\*

### From probability density functions to measures

- ▶ The integrand  $dx$  in the def. of the p.d.f. is called the Lebesgue measure.  
It constraints r.v. with p.d.f. to satisfy  $\mathbb{P}(X = k) = \int_k^k f(x)dx = 0$
- ▶ Measure theory extends the classical definition of integrals by "considering other integrands  $dx$ "
- ▶ For example, the Dirac measure  $\delta_y$  on a point  $y \in \mathbb{R}$  is defined such that

$$\int_a^b d\delta_y(x) = \begin{cases} 1 & \text{if } y \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

→ It can encompass discrete r.v.!

## Cumulative Distribution Function

### Definition (Cumulative distribution function)

The **cumulative distribution function** (c.d.f.) of a r.v.  $X$  on  $(\Omega, \mathcal{F}, \mathbb{P})$  is

$$F(t) = \mathbb{P}(X \leq t) = \mathbb{P}_X([-\infty, t])$$

### Property

1.  $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a)$
2.  $\lim_{t \rightarrow -\infty} F(t) = 0, \quad \lim_{t \rightarrow +\infty} F(t) = 1$
3. (Monotonicity) If  $s \leq t$ ,  $F(s) \leq F(t)$

## Cumulative distribution function

### Cumulative distribution function for continuous r.v.

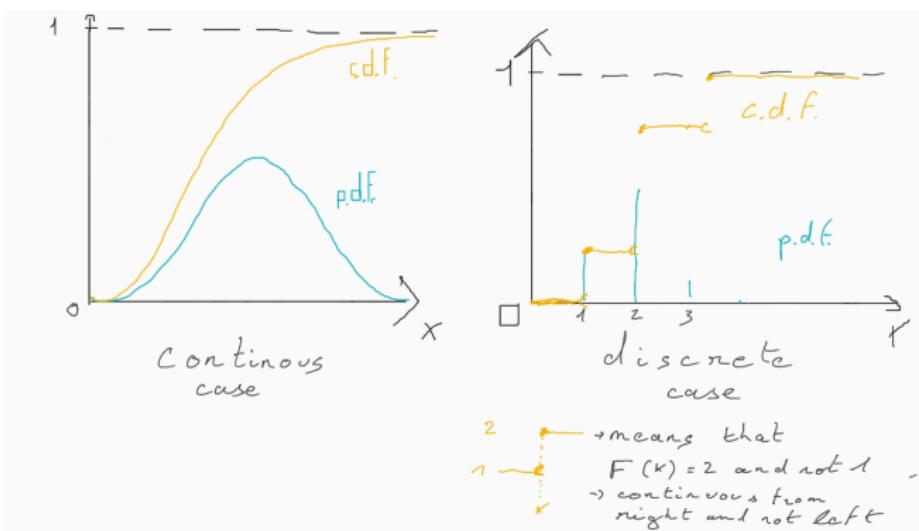
If a r.v.  $X$  has a p.d.f.  $f$  then its c.d.f. is continuous and reads

$$F(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^t f(x)dx$$

Reciprocally we have  $f(x) = F'(x)$  (Fundamental theorem of calculus)

### Cumulative distribution function for discrete r.v.

If a r.v.  $X$  is discrete, then its c.d.f. is piecewise constant.



# Cumulative Distribution function

## Continuity

1. (Right continuity)  $F(t) = \lim_{s \rightarrow t^+} F(s)$   
( $s \rightarrow t^+$  means  $s$  approaches  $t$  from the right)
2.  $\mathbb{P}(X < a) = \lim_{s \rightarrow a^-} F(s)$

## Why introducing c.d.f. ?

1. (Theoretical)\* Intervals of the form  $[-\infty, b]$  span all subsets we want to measure, i.e., the canonical  $\sigma$ -algebra of  $\mathbb{R}$ .  
→ characterizes a probability distribution

Note: one random variable can have multiple p.d.f.

Example:  $\tilde{f}(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, c) \cup (c, b] \\ 0 & \text{if } x \notin [a, b] \text{ or } x = c \end{cases}, f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

They are **equal almost everywhere**, i.e., their integral coincide

2. (Practical) Provides a way to sample from a distribution

## Generate Randomness

### Random Number Generator

"Generates a sequence of numbers  
that cannot be reasonably predicted better than by a random chance."

→ have access to uniform distribution  $U$  on  $[0, 1]$

#### Sampling from c.d.f

Assume that we have access to a c.d.f.  $F_X$  of a r.v.  $X$  with  $F$  strictly increasing

**Goal:** Generate  $Y$  such that c.d.f. of  $Y$  and c.d.f. of  $X$  coincide

**Method:** Define  $T = F_X^{-1}$  (possible if  $F$  strictly increasing)

Generate  $Y = T(U)$  s.t.

$$\mathbb{P}(Y \leq t) = \mathbb{P}(U \leq F_X(t)) = F_X(t)$$

where we used that  $\mathbb{P}(U \leq a) = \int_0^a dx = a$  for any  $a \in [0, 1]$ .

# Overview

## Until now

1. Probability mass/density functions
  - ▶ "unit measures" to compute probabilities (that are measures of sets)
2. Cumulative distribution function
  - ▶ Characterizes distributions
  - ▶ Offers way to sample

## Next

1. Expectation
  - ▶ key operator on probability distributions,
  - ▶ can get back probability distributions as expectations of a function
2. Moments of a r.v., in particular variance
3. Median, quantiles
4. Gaussian distribution

## Another Example of Discrete Random Variable

In France, the “galette des rois” (King cake) contains a figurine, the “fève”, hidden in the cake and the person who finds the fève in his or her slice becomes king/queen for the day. Assume that *galettes* are cut into 6 identical slices and that the fève is put at random in the galette.



Let  $X$  be the number of galettes you eat until you find a *fève*, i.e. until the slice you are given in the galette contains the *fève*.

1. What values can take  $X$ ?
2. What is the probability mass function of  $X$ ?
3. What is the name of this probability distribution?

# Review on Probability Distributions, Continuous and Discrete Random Variables

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 3, April 3rd, 2020

Ask questions via [chat on Zoom](#)  
Answer questions via [PollEv.com](#), (username *vincentroulet*)  
Whiteboard via [JamBoard](#) (link in the syllabus page)

## Overview

### Update on syllabus

- ▶ Full tentative schedule available (follows closely the book)
- ▶ Only 6 homeworks and not 8 (give you time to prepare for the exams)

### Previous lecture

- ▶ Probability distribution
- ▶ Probability mass function, probability density function
- ▶ Cumulative distribution function

### This lecture

- ▶ Expectation of a r.v.
- ▶ Function of a r.v., expectation fo a r.v.
- ▶ Moments of a r.v., in particular variance
- ▶ Gaussian distribution

# Expectation

## Definition (Expectation)

Let  $X$  be a r.v. on probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

1. **(Discrete case)** If  $X$  has a p.m.f  $p$  s.t.  $\sum_{k \in X(\Omega)} |k|p(k) < \infty$ ,  
the **expectation** (or **expected value**) of  $X$  exists and reads

$$\mathbb{E}[X] = \sum_{k \in X(\Omega)} kp(k)$$

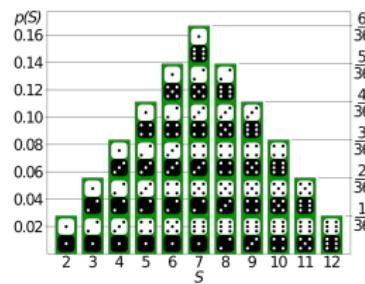
2. **(Continuous case)** If  $X$  has a p.d.f.  $f$  s.t.  $\int_{-\infty}^{+\infty} |x|f(x)dx < +\infty$   
the **expectation** (or **expected value**) of  $X$  exists and reads

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

# Expectation

## Example

1. Give the expected value of the roll of two dice
2. Give the expected reaching point of the bowling ball (Uniform on  $[-1, 1]$ )



## Linearity of Expectation

### Property (Linearity)

The expectation is **linear**, i.e.,

1. For any r.v. (discrete or continuous)  $X$  and  $Y$ ,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

2. For any r.v. (discrete or continuous)  $X$  and  $a, b \in \mathbb{R}$ ,

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

We treat  $b$  as a degenerate r.v., i.e. we identify  $b$  to  $\tilde{b}$ : 
$$\begin{cases} \Omega & \rightarrow \mathbb{R} \\ \omega & \rightarrow b \end{cases}$$

## Linearity of expectation

Linearity of expectation is very useful to decompose computations

### Example (Binomial)

Compute the expected value of a binomial r.v.  $X \sim \text{Bin}(n, p)$  that represents the number of success among  $n$  independent Bernoulli random variables (flip of a coin) each having a probability  $p$  of success

### Example (Birthday at office)

15 workers in an office. One birthday party last day of the month if at least one worker had his birthday this month. Average number of party during the year?

### Example (Party)

$n$  guest at a party. Each pair knows each other with proba 1/2. Average number of groups of 3 people where everybody knows each other?

## Expectation of a Function of a Random Variable

Property (Expectation of a function of a random variable)

Let  $X$  be a r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $g : X(\Omega) \rightarrow \mathbb{R}$ .

Then  $g(X)$  is a r.v. and

1. (Discrete case) if  $X$  has a p.m.f.  $p$ , and  $\sum_{k \in X(\Omega)} |g(k)|p(k) < +\infty$ , then

$$\mathbb{E}[g(X)] \text{ exists and } \mathbb{E}[g(X)] = \sum_{k \in X(\Omega)} g(k)p(k)$$

2. (Continuous case) if  $X$  has a p.d.f.  $f$ , and  $\int_{-\infty}^{+\infty} |g(x)|f(x)dx < +\infty$ , then

$$\mathbb{E}[g(X)] \text{ exists and } \mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

Proof on whiteboard

In the following we denote  $\mathbb{E}[|g(X)|] < +\infty$  the conditions in 1. and 2.

# From Expectation to Probability Distributions

## Probability distributions

For any subset  $B \subset \mathbb{R}$ , denote  $\mathbf{1}_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise} \end{cases}$

## Property

Let  $X$  be a r.v. with probability distribution  $\mathbb{P}_X$  and c.d.f.  $F_X$ , then

$$\mathbb{E}[\mathbf{1}_B(X)] = \mathbb{P}_X(B), \quad \mathbb{E}[\mathbf{1}_{(-\infty, t]}(X)] = \mathbb{P}(X \leq t) = F_X(t)$$

Proof on whiteboard

# Moments of a Random Variable

## Definition (Moment)

For a r.v.  $X$  and  $m \in \mathbb{N}$ , if  $\mathbb{E}[|X|^m] < +\infty$ , then

1. the  $m^{\text{th}}$  moment of  $X$  exists and is defined as  $\mathbb{E}(X^m)$
2. the  $m^{\text{th}}$  centered moment is defined as  $\mathbb{E}((X - \mathbb{E}(X))^m)$

## Comments

1. Provide meaningful information about the random variable, e.g.
  - ▶ For  $m = 1$ , the moment is the average (expected value)
  - ▶ For  $m = 2$ , the moment is the spread of the random variable around the average (variance)
  - ▶ For  $m = 3$ , the centered moment describes the asymmetry of the random variable around the mean

⋮

2. Question:

How many moments are necessary  
to fully characterize the probability distribution?

Answer will be provided in the course!

## Variance

### Definition (Variance–Standard Deviation)

Let  $X$  be a discrete r.v. on probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

If  $\mathbb{E}[|X|^2] < +\infty$ , the **variance** of  $X$  is defined by

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

The **standard deviation** of  $X$  is defined by  $\sigma_X = \sqrt{\text{Var}(X)}$

### Example (Bernoulli)

Give the variance of  $X \sim \text{Ber}(p)$  (Bernoulli with probability  $p$  of success)

### Example (Uniform)

Give the variance of  $X \sim \mathcal{U}([a, b])$  (Uniform on an interval  $[a, b]$ )

# Review on Probability Distributions, Continuous and Discrete Random Variables

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 4, April 6th, 2020

Ask questions via [chat on Zoom](#)  
Answer questions via [PollEv.com](#), (username *vincentroulet*)  
Whiteboard via [JamBoard](#) (link in the syllabus page)

## Overview

### Previous lecture

- ▶ Expectation, **linearity of expectation**, some computations
- ▶ Expectation of a function of a random variable
- ▶ Moments, Variance

### This lecture

- ▶ Variance properties, Quantiles
- ▶ Gaussian distribution
- ▶ Exponential distribution
- ▶ (Poisson distribution, Gamma distribution)

## Variance

### Definition (Variance–Standard Deviation)

Let  $X$  be a discrete r.v. on probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

If  $\mathbb{E}[|X|^2] < +\infty$ , the **variance** of  $X$  is defined by

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

The **standard deviation** of  $X$  is defined by  $\sigma_X = \sqrt{\text{Var}(X)}$

## Property

1. Let  $X$  be a r.v. such that  $\text{Var}(X)$  exists and  $a, b \in \mathbb{R}$ , then

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

2. For a r.v.  $X$ , we have  $\text{Var}(X) = 0$  if and only if  $\exists a \in \mathbb{R}$  s.t.  $\mathbb{P}(X = a) = 1$

## Median and quantiles

### Definition

1. A number  $m \in \mathbb{R}$  is called a **median** of a r.v.  $X$  if

$$\mathbb{P}(X \geq m) \geq 1/2 \quad \text{and} \quad \mathbb{P}(X \leq m) \geq 1/2$$

2. For  $0 < p < 1$ , a number  $q$  is called a  $p^{\text{th}}$  **quantile** of a r.v.  $X$  if

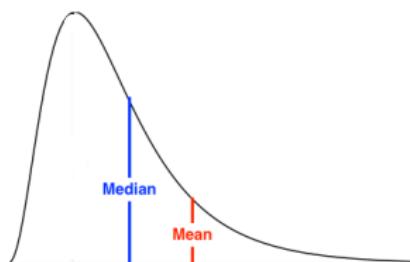
$$\mathbb{P}(X \geq q) \geq 1 - p \quad \text{and} \quad \mathbb{P}(X \leq q) \geq p$$

**Note:** Medians and quantiles are a priori not uniquely defined

### Example

$X$  uniformly distributed on  $\{1, 2, 3, 4, 5, 100\}$ .

What is the median? What is the mean?



## Gaussian Distribution

### Definition (Standard normal distribution)

A r.v.  $Z$  follows a **standard normal distribution** (denoted  $Z \sim \mathcal{N}(0, 1)$ ) if it has a p.d.f

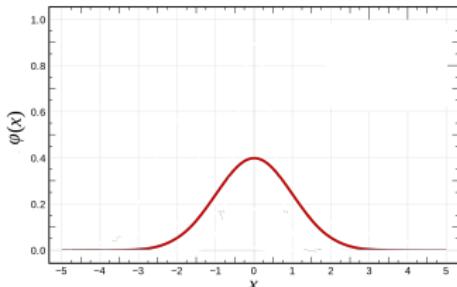
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

### Normalization

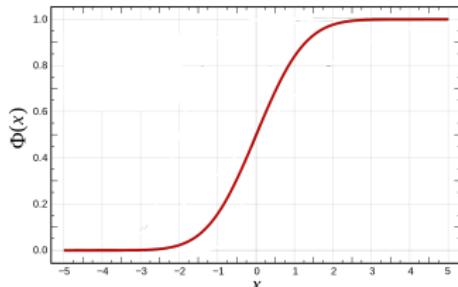
We have (see e.g. book page 120)  $\int_{-\infty}^{+\infty} e^{-x^2/2} dx = \sqrt{2\pi}$ ,

so  $\phi$  is well normalized,  $\int_{-\infty}^{+\infty} \phi(x) dx = 1$

**Cumulative distribution function:**  $\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$



p.d.f.



c.d.f.

## Gaussian Distribution

### Property

If  $Z \sim \mathcal{N}(0, 1)$ , then  $\mathbb{E}(Z) = 0$ ,  $\text{Var}(Z) = 1$ .

### Corollary

Consider  $X = \sigma Z + \mu$  for  $Z \sim \mathcal{N}(0, 1)$ . Then  $\mathbb{E}[X] = \mu$ ,  $\text{Var}(X) = \sigma^2$

### Proposition

Consider  $X = \sigma Z + \mu$  for  $Z \sim \mathcal{N}(0, 1)$ ,  $X$  has a p.d.f. (proof next slide)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

### Definition

A r.v.  $X$  follows a **normal/Gaussian distribution** with **mean**  $\mu$  and **variance**  $\sigma^2$ , denoted  $X \sim \mathcal{N}(\mu, \sigma^2)$ , if  $X$  has a p.d.f.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

## How to derive density functions from transformations of r.v.

- Derive expression of the c.d.f. for which transformations are easily expressed

Here  $X = \sigma Z + \mu$ , so (recall that we denote by  $\Phi$  the c.d.f. of  $Z$ )

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\sigma Z + \mu \leq t) = \mathbb{P}\left(Z \leq \frac{t - \mu}{\sigma}\right) = \Phi\left(\frac{t - \mu}{\sigma}\right)$$

- Get the expression of the p.d.f. as the derivative of the c.d.f.

We have recall that the p.d.f. of  $Z$  is  $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

$$f_X(x) = F'_X(x) = \frac{d}{dx} \left[ \Phi\left(\frac{x - \mu}{\sigma}\right) \right] = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

## Gaussian Distribution

### Exercise

If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , how can we choose  $a, b \in \mathbb{R}$  s.t.  $Z = aX + b \sim \mathcal{N}(0, 1)$ ?

1.  $a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma}$
2.  $a = \frac{1}{\sigma^2}, b = -\frac{\mu}{\sigma^2}$
3.  $a = \frac{1}{\sigma}, b = -\mu$
4.  $a = \frac{1}{\sigma^2}, b = -\mu$

### Note

- ▶ Mean and variance entirely characterize normal distributions.
- ▶ Gaussian distributions play a central role in probability through the central limit theorem:

The empirical mean of  $n$  observations of a random variable behaves asymptotically as a Gaussian

# Exponential Distribution

## Motivation

- ▶ The geometric distribution models the number of trials before the success of a Bernoulli r.v.
- ▶ Good for a *sequence* of trials but what about continuous time ?

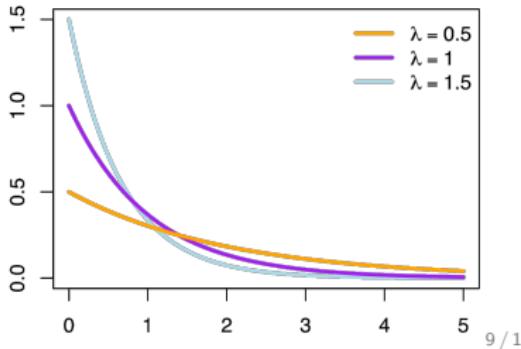
## Definition (Exponential distribution)

A r.v.  $X$  has an **exponential distribution** with parameter  $\lambda > 0$ , denoted  $X \sim \text{Exp}(\lambda)$ , if it has a probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

## Cumulative distribution function

$$F(t) = \int_{-\infty}^t f(x)dx = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}$$



## From Geometric to Exponential

**Objective:** Model the waiting time for a phone call.

**Idea:** Discretize time in intervals of size  $\tau$ .

### From Bernoulli to continuous

Assume that the probability that a phone call is answered during a time interval  $\tau$  is proportional to the size of this interval, i.e.  $\mathbb{P}(B_\tau = 1) = \lambda\tau$ , where  $B_\tau$  models the event that a phone call is answered during a time interval  $\tau < \lambda$ .

### From geometric to exponential

Let  $X_\tau$  be the number of intervals of time  $\tau$  we need to wait for the phone call

$$\mathbb{P}(X = k) = \mathbb{P}(B_\tau = 0)^{k-1} \mathbb{P}(B_\tau = 1) = (1 - \lambda\tau)^{k-1} \lambda\tau$$

Let  $T_\tau$  the approximate discretized time that we waited, i.e.,  $T_\tau \in \{0, \tau, 2\tau, \dots\}$

$$\mathbb{P}(T_\tau = k\tau) = (1 - \lambda\tau)^{k-1} \lambda\tau$$

"The distribution of  $T_\tau$  tends to an exponential distribution  $\text{Exp}(\lambda)$  as  $\tau \rightarrow 0$ ."

## From Geometric to Exponential

### Definition

A sequence of r.v.,  $X_1, X_2, \dots$  is said to converge **in distribution**, or **converge weakly**, or **converge in law** to a r.v.  $X$  if for any  $t \in \mathbb{R}$ ,

$$\lim_{n \rightarrow +\infty} F_{X_n}(t) = F_X(t)$$

where  $F_{X_n}, F_X$  are the c.d.f. of  $X_n$  and  $X$  respectively.

### Theorem

Let  $\lambda > 0$  and define for  $p \in \mathbb{N}$  ( $1/p = \tau$  previously) s.t.  $\lambda/p < 1$ , the discrete non-negative r.v.  $W_p$  such that

$$\mathbb{P}(W_p = k/p) = (1 - \lambda/p)^{k-1} \lambda/p$$

Then the r.v.  $(W_p)_{p_0}^{+\infty}$  defined for  $p$  sufficiently large, converge in law towards an exponential distribution.

## Exponential Distribution

### Exercise

Let  $X \sim \text{Exp}(\lambda)$ , then  $\mathbb{E}(X) = \frac{1}{\lambda}$ ,  $\text{Var}(X) = \frac{1}{\lambda^2}$ .

### Application

Waiting time of a phone call modeled by an exponential r.v.

The average waiting time is 5min.

What is the probability that the waiting time is more than 8 min?

### Property (Memoryless)

Let  $X \sim \text{Exp}(\lambda)$ , then for any  $s, t > 0$ ,

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s)$$

## Poisson Distribution

### Motivation

Models number of occurrences of Bernoulli r.v. for an infinite number of trials with an average of  $\lambda$  occurrences.

Can be seen as the limit of binomial r.v.  $B_n \sim \text{Bin}(n, \lambda/n)$ , whose average number of successes remain the same as  $n$  increases.

### Definition (Poisson distribution)

A discrete r.v.  $X$  has a **Poisson distribution** with parameter  $\lambda > 0$ , denoted  $X \sim \text{Poiss}(\lambda)$  if  $X$  has a p.m.f.

$$\mathbb{P}(X = k) = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} & \text{if } k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

### Normalization

We have  $e^\lambda = \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!}$  so  $\sum_{k=0}^{+\infty} \mathbb{P}(X = k) = 1$

### Exercise

If  $X \sim \text{Poiss}(\lambda)$ , then  $\mathbb{E}(X) = \lambda$ ,  $\text{Var}(X) = \lambda$ .

## Gamma Distribution\*

### Motivation

Poisson distribution models occurrences of a Bernoulli r.v. for an infinite sequence of trials.

What about continuous time and continuous number of success ?

### Definition (Gamma distribution)

A r.v.  $X$  follows a **Gamma distribution** with parameters  $r, \lambda > 0$ , denoted  $X \sim \text{Gamma}(r, \lambda)$ , if it has a p.d.f.

$$f(x) = \begin{cases} \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

where  $\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx$  s.t.  $\Gamma(n) = (n-1)!$  for  $n \in \mathbb{N}$ .

### Application

The time (in minutes) to wait for the  $n^{\text{th}}$  call can be modeled as  $X \sim \text{Gamma}(n, \lambda)$ .

What is the probability that I receive 2 calls after 5min for  $\lambda = 60$ ?

# Overview

## This lecture

- ▶ Variance properties, Quantiles
- ▶ Gaussian distribution
- ▶ Exponential distribution
- ▶ (Poisson distribution, Gamma distribution)

## Next lecture

- ▶ Joint distributions

# Joint Probability Distributions Discrete case

Section 6.1

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 5, April 8th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Lecture note

- ▶ A lecture note reviewing MATH/STAT 394 is available

### Homework

- ▶ 1st homework available tonight, due next Wednesday 11:59 pm
- ▶ No late homework accepted
- ▶ 1st homework is long, begin soon
- ▶ Provide **clear and detailed** answers
- ▶ One exercise chosen at random is graded to correct bad mathematical formulations
- ▶ The rest will be given points for completion

### Office hours

- ▶ Answer poll to maximize availability
- ▶ For the moment, might be updated
  - ▶ Mondays 10:00 to 11:00 with T.A. Z. Yuan by Zoom
  - ▶ Fridays 11:00 to 12:00 with V. Roulet by Zoom

# Overview

## Previous lectures

- ▶ Probability space, probability distributions
- ▶ Probability mass function, probability density function, cumulative distribution function
- ▶ Expectation, Variance
- ▶ Various discrete and continuous random variables

## This lecture

- ▶ Joint distributions discrete
- ▶ Marginal distributions
- ▶ Multinomial distribution

## Answer Previous Quizzes

### Exercise

If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , how can we choose  $a, b \in \mathbb{R}$  s.t.  $Z = aX + b \sim \mathcal{N}(0, 1)$ ?

### Answer

1.  $\text{Var}(Z) = a^2 \text{Var}(X) = a^2 \sigma^2$ .
2. So to have  $\text{Var}(Z) = 1$ , we need  $a = 1/\sigma$
3.  $\mathbb{E}(Z) = a\mathbb{E}(X) + b = a\mu + b$
4. So to have  $\mathbb{E}(Z) = 0$ , we need  $b = -\mu/\sigma$
5. Answer was 4. i.e.  $Z = \frac{X-\mu}{\sigma}$

## Answer Previous Quizzes

### Exercise

*Waiting time of a phone call modeled by an exponential r.v.*

*The average waiting time is 5min.*

*What is the probability that the waiting time is more than 8 min?*

### Answer

1.  $X \sim \text{Poisson}(\lambda)$
2.  $\mathbb{E}(X) = \frac{1}{\lambda} = 5$  so  $\lambda = 1/5$
3.  $\mathbb{P}(X \geq 8) = \mathbb{P}(X > 8) = 1 - F(8) = e^{-\lambda 8} = e^{-8/5} \approx 0.20$
4. **⚠ Typo in quiz**, none of the answers were correct

## Multivariate Random Variable/Random Vector

Definition (Multivariate random variable/Random vector)

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , a **multivariate random variable** or **random vector** is a vector  $X = (X_1, \dots, X_n)$ , whose components are real-valued random variables on  $(\Omega, \mathcal{F}, \mathbb{P})$ .

**Note:** Rather than speaking about the distribution of a random vector, we often speak about the joint distribution of the r.v. it is composed of

## Multivariate Random Variable/Random Vector

### Definition (Multivariate random variable/Random vector)

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , a **multivariate random variable** or **random vector** is a vector  $X = (X_1, \dots, X_n)$ , whose components are real-valued random variables on  $(\Omega, \mathcal{F}, \mathbb{P})$ .

**Note:** Rather than speaking about the distribution of a random vector, we often speak about the joint distribution of the r.v. it is composed of

### Example (Classic examples)

1. (Discrete case) Roll a die 100 times, denote  $X_1, \dots, X_6$  the number of 1, ..., 6 you got respectively, then  $X = (X_1, \dots, X_6)$  is a random vector
2. (Continuous case) Throw a dart uniformly at random on a disc, the coordinates  $(X, Y)$  of that throw form a random vector

## Joint Probability Mass Function

Definition (Joint probability mass function)

Let  $X_1, \dots, X_n$  be discrete r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , their **joint probability mass function** is defined as

$$\begin{aligned} p(k_1, \dots, k_n) &= \mathbb{P}(\{X_1 = k_1\} \cap \dots \cap \{X_n = k_n\}) \\ &\triangleq \mathbb{P}(X_1 = k_1, \dots, X_n = k_n) \end{aligned}$$

for any  $k_1, \dots, k_n \in X_1(\Omega) \times \dots \times X_n(\Omega)$  (any values taken by the random vector)

**Note:**

- ▶ Describe all joint values of the r.v.
- ▶ We then naturally have  $p(k_1, \dots, k_n) \geq 0$  and

$$\sum_{k_1, \dots, k_n \in X_1(\Omega) \times \dots \times X_n(\Omega)} p(k_1, \dots, k_n) = 1$$

## Joint Probability Mass Function

### Example

1. Roll two dice with **4 faces**, denote
  - (i)  $S$  the sum of the two dice
  - (ii)  $Y$  the indicator variable that you get a pair
2. Record which outcomes lead to different values of  $S, Y$

## Joint Probability Mass Function

### Example

1. Roll two dice with **4 faces**, denote
  - (i)  $S$  the sum of the two dice
  - (ii)  $Y$  the indicator variable that you get a pair
2. Record which outcomes lead to different values of  $S, Y$

|   | $Y$                         |        |
|---|-----------------------------|--------|
|   | 0                           | 1      |
| 2 |                             | (1,1)  |
| 3 | (1, 2) (2, 1)               |        |
| 4 | (1, 3) (3, 1)               | (2, 2) |
| 5 | (1, 4) (2, 3) (3, 2) (4, 1) |        |
| 6 | (2, 4) (4, 2)               | (3, 3) |
| 7 | (3, 4) (4, 3)               |        |
| 8 |                             | (4, 4) |

## Joint Probability Mass Function

### Example

1. Roll two dice with **4 faces**, denote
  - (i)  $S$  the sum of the two dice
  - (ii)  $Y$  the indicator variable that you get a pair
2. Record which outcomes lead to different values of  $S, Y$
3. Compute the corresponding joint probability mass function of  $S, Y$

|   |               | Y                           |       |
|---|---------------|-----------------------------|-------|
|   | 0             |                             | 1     |
| 2 |               |                             | (1,1) |
| 3 | (1, 2) (2, 1) |                             |       |
| 4 | (1, 3) (3, 1) | (2, 2)                      |       |
| S | 5             | (1, 4) (2, 3) (3, 2) (4, 1) |       |
| 6 | (2, 4) (4, 2) | (3, 3)                      |       |
| 7 | (3, 4) (4, 3) |                             |       |
| 8 |               | (4, 4)                      |       |

# Joint Probability Mass Function

## Example

1. Roll two dice with **4 faces**, denote
  - (i)  $S$  the sum of the two dice
  - (ii)  $Y$  the indicator variable that you get a pair
2. Record which outcomes lead to different values of  $S, Y$
3. Compute the corresponding joint probability mass function of  $S, Y$

|   |               | Y                           |       |
|---|---------------|-----------------------------|-------|
|   | 0             |                             | 1     |
| 2 |               |                             | (1,1) |
| 3 | (1, 2) (2, 1) |                             |       |
| 4 | (1, 3) (3, 1) | (2, 2)                      |       |
| S | 5             | (1, 4) (2, 3) (3, 2) (4, 1) |       |
| 6 | (2, 4) (4, 2) | (3, 3)                      |       |
| 7 | (3, 4) (4, 3) |                             |       |
| 8 |               | (4, 4)                      |       |

|   | Y   |      |   |
|---|-----|------|---|
| 0 | 0   | 1/16 |   |
| 1 | 1/8 | 0    |   |
| 2 | 0   | 1/16 |   |
| 3 | 1/8 | 0    |   |
| 4 | 1/8 | 1/16 |   |
| S | 5   | 1/4  | 0 |
| 6 | 1/8 | 1/16 |   |
| 7 | 1/8 | 0    |   |
| 8 | 0   | 1/16 |   |

# Joint Probability Mass Function

## Example

1. Roll two dice with **4 faces**, denote
  - (i)  $S$  the sum of the two dice
  - (ii)  $Y$  the indicator variable that you get a pair
2. Record which outcomes lead to different values of  $S, Y$
3. Compute the corresponding joint probability mass function of  $S, Y$
4. Read e.g.  $\mathbb{P}(S = 4, Y = 1) = 1/16$

|     | $Y$                                 |          |
|-----|-------------------------------------|----------|
|     | 0                                   | 1        |
| 2   |                                     | $(1, 1)$ |
| 3   | $(1, 2)$ $(2, 1)$                   |          |
| 4   | $(1, 3)$ $(3, 1)$                   | $(2, 2)$ |
| $S$ | $(1, 4)$ $(2, 3)$ $(3, 2)$ $(4, 1)$ |          |
| 6   | $(2, 4)$ $(4, 2)$                   | $(3, 3)$ |
| 7   | $(3, 4)$ $(4, 3)$                   |          |
| 8   |                                     | $(4, 4)$ |

|     | $Y$   |        |
|-----|-------|--------|
|     | 0     | 1      |
| 2   | 0     | $1/16$ |
| 3   | $1/8$ | 0      |
| 4   | $1/8$ | $1/16$ |
| $S$ | $1/4$ | 0      |
| 6   | $1/8$ | $1/16$ |
| 7   | $1/8$ | 0      |
| 8   | 0     | $1/16$ |

## Joint Probability Mass Function

### Lemma

Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $X_1, \dots, X_n$  be discrete r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with joint probability mass function  $p$ , then

$$\mathbb{E}[g(X_1, \dots, X_n)] = \sum_{k_1, \dots, k_n \in X_1(\Omega) \times \dots \times X_n(\Omega)} g(k_1, \dots, k_n) p(k_1, \dots, k_n)$$

**Note:** Extends naturally previous property for univariate r.v.

**Example:** Take  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  such that  $g(x_1, \dots, x_n) = \max_{i \in \{1, \dots, n\}} x_i$

## Joint Probability Mass Function

### Example

1. Roll two dices with **4 faces**, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Score is the sum of the dice, doubled if it is a pair

What is the average score?

|          | $Y$      |        |
|----------|----------|--------|
|          | 0        | 1      |
| 0        | 0        | $1/16$ |
| 2        | 0        | $1/16$ |
| 3        | $1/8$    | 0      |
| 4        | $1/8$    | $1/16$ |
| <b>S</b> | <b>5</b> | $1/4$  |
| 6        | $1/8$    | $1/16$ |
| 7        | $1/8$    | 0      |
| 8        | 0        | $1/16$ |

# Joint Probability Mass Function

## Example

1. Roll two dices with **4 faces**, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Score is the sum of the dice, doubled if it is a pair

What is the average score?

## Solution

1. The score is  $g(S, Y) = S(Y + 1)$

|   | Y   |      |
|---|-----|------|
|   | 0   | 1    |
| 0 | 0   | 1/16 |
| 2 | 0   | 1/16 |
| 3 | 1/8 | 0    |
| 4 | 1/8 | 1/16 |
| S | 1/4 | 0    |
| 5 | 1/4 | 0    |
| 6 | 1/8 | 1/16 |
| 7 | 1/8 | 0    |
| 8 | 0   | 1/16 |

# Joint Probability Mass Function

## Example

1. Roll two dices with 4 faces, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Score is the sum of the dice, doubled if it is a pair

What is the average score?

|   | Y   |      |
|---|-----|------|
| 0 | 0   | 1    |
| 2 | 0   | 1/16 |
| 3 | 1/8 | 0    |
| 4 | 1/8 | 1/16 |
| S | 1/4 | 0    |
| 5 | 1/4 | 0    |
| 6 | 1/8 | 1/16 |
| 7 | 1/8 | 0    |
| 8 | 0   | 1/16 |

## Solution

1. The score is  $g(S, Y) = S(Y + 1)$
2. The average score reads

$$\begin{aligned}\mathbb{E}[g(S, Y)] &= \sum_{s=2}^8 \sum_{y=0}^1 s(y+1)p(s, y) \\ &= \sum_{s=2}^8 sp(s, 0) + 2 \sum_{s=2}^8 sp(s, 1) \\ &= \frac{3+4+2\times 5+6+7}{8} + 2 \times \frac{2+4+6+8}{16} = 25/4 = 6.25\end{aligned}$$

## Marginal Probability Mass Function

### Definition

Let  $p_{X,Y}$  be the joint probability mass function of two r.v.  $(X, Y)$ . The probability mass function of  $X$  is given by,

$$p_X(k) \triangleq \mathbb{P}(X = k) = \sum_{\ell \in Y(\Omega)} p_{X,Y}(k, \ell)$$

The function  $p_X$  is called the **marginal probability distribution** of  $X$ .

**Proof** The events  $\{B_\ell = \{Y = \ell\}\}_{\ell \in Y(\Omega)}$  form a partition of  $\Omega$  by definition of a discrete random variable such that

$$\mathbb{P}(X = k) = \mathbb{P} \left( \{X = k\} \cap \bigcup_{\ell=-\infty}^{+\infty} B_\ell \right) = \sum_{\ell=-\infty}^{+\infty} \mathbb{P}(X = k, Y = \ell) = \sum_{\ell \in Y(\Omega)} p_{X,Y}(k, \ell)$$

## Marginal Probability Mass Function

### Definition

Let  $p$  be the joint probability mass function of  $n$  discrete r.v.  $X_1, \dots, X_n$ . The probability mass function of  $X_j$  for  $j \in \{1, \dots, n\}$  is given by for any  $k \in X_j(\Omega)$ ,

$$p_{X_j}(k) = \sum_{\substack{\ell_1, \dots, \ell_{j-1}, \ell_{j+1}, \dots, \ell_n \\ \in X_1(\Omega) \times \dots \times X_{j-1}(\Omega) \times X_{j+1}(\Omega) \times \dots \times X_n(\Omega)}} p(\ell_1, \dots, \ell_{j-1}, k, \ell_{j+1}, \dots, \ell_n)$$

The function  $p_{X_j}$  is called the **marginal probability distribution** of  $X_j$ .

**Proof** Denote  $p_{X_{i_1}, \dots, X_{i_j}}$  the joint p.m.f. of any subset  $X_{i_1}, \dots, X_{i_j}$  of r.v. with  $2 \leq j \leq n$  and  $1 \leq i_1 < \dots < i_j \leq i_n$ , then naturally

$$p_{X_{i_1}, \dots, X_{i_{j-1}}}(k_{i_1}, \dots, k_{i_{j-1}}) = \sum_{\ell_{i_j} \in X_{i_j}(\Omega)} p_{X_{i_1}, \dots, X_{i_j}}(k_{i_1}, \dots, \ell_{i_j})$$

By applying recursively this fact we get the result.

## Marginal Probability Mass Function

Previous result generalizes to the joint probability distribution of any subset.  
For example the joint probability of  $X_1, \dots, X_m$  given  $m < n$  is

$$p_{X_1, \dots, X_m}(k_1, \dots, k_m) = \sum_{\ell_{m+1}, \dots, \ell_n \in X_{m+1}(\Omega) \times \dots \times X_n(\Omega)} p(k_1, \dots, k_m, \ell_{m+1}, \dots, \ell_n)$$

## Marginal Probability Mass Function

### Example

1. Roll two dices with 4 faces, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Compute marginal distribution of  $Y$  from the joint p.m.f.

|     |       | $Y$    |   |
|-----|-------|--------|---|
|     | 0     | 1      |   |
| 2   | 0     | $1/16$ |   |
| 3   | $1/8$ | 0      |   |
| 4   | $1/8$ | $1/16$ |   |
| $S$ | 5     | $1/4$  | 0 |
| 6   | $1/8$ | $1/16$ |   |
| 7   | $1/8$ | 0      |   |
| 8   | 0     | $1/16$ |   |

## Marginal Probability Mass Function

### Example

1. Roll two dices with 4 faces, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Compute marginal distribution of  $Y$  from the joint p.m.f.

|     |   | $Y$   |        |
|-----|---|-------|--------|
|     |   | 0     | 1      |
|     | 2 | 0     | $1/16$ |
|     | 3 | $1/8$ | 0      |
|     | 4 | $1/8$ | $1/16$ |
| $S$ | 5 | $1/4$ | 0      |
|     | 6 | $1/8$ | $1/16$ |
|     | 7 | $1/8$ | 0      |
|     | 8 | 0     | $1/16$ |

### Solution:

Sum the columns of  $p(s, y)$

So you get  $\mathbb{P}(Y = 1) = 4/16$  and  $\mathbb{P}(Y = 0) = 12/16$

## Multinomial Distribution

### Motivation

Consider a trial with  $r$  possible outcomes, labeled  $1, \dots, r$ . Denote  $p_j$  the probability of the outcome  $j$  such that  $p_1 + \dots + p_r = 1$ . Perform  $n$  independent repetitions of this trial. Denote  $X_j$  the number of times the outcome  $j$  appeared among the  $n$  trials.

What is the joint probability mass function of  $(X_1, \dots, X_r)$ ?

## Multinomial Distribution

### Motivation

Consider a trial with  $r$  possible outcomes, labeled  $1, \dots, r$ . Denote  $p_j$  the probability of the outcome  $j$  such that  $p_1 + \dots + p_r = 1$ . Perform  $n$  independent repetitions of this trial. Denote  $X_j$  the number of times the outcome  $j$  appeared among the  $n$  trials.

What is the joint probability mass function of  $(X_1, \dots, X_r)$ ?

### Derivation

1. Let  $k_1, \dots, k_r \in \mathbb{N}$  such that  $k_1 + \dots + k_r = n$ .

# Multinomial Distribution

## Motivation

Consider a trial with  $r$  possible outcomes, labeled  $1, \dots, r$ . Denote  $p_j$  the probability of the outcome  $j$  such that  $p_1 + \dots + p_r = 1$ . Perform  $n$  independent repetitions of this trial. Denote  $X_j$  the number of times the outcome  $j$  appeared among the  $n$  trials.

What is the joint probability mass function of  $(X_1, \dots, X_r)$ ?

## Derivation

1. Let  $k_1, \dots, k_r \in \mathbb{N}$  such that  $k_1 + \dots + k_r = n$ .
2. Any outcome that leads to  $X_j = k_j$  for all  $j \in \{1, \dots, r\}$  has proba  $p_1^{k_1} \dots p_r^{k_r}$ .

# Multinomial Distribution

## Motivation

Consider a trial with  $r$  possible outcomes, labeled  $1, \dots, r$ . Denote  $p_j$  the probability of the outcome  $j$  such that  $p_1 + \dots + p_r = 1$ . Perform  $n$  independent repetitions of this trial. Denote  $X_j$  the number of times the outcome  $j$  appeared among the  $n$  trials.

What is the joint probability mass function of  $(X_1, \dots, X_r)$ ?

## Derivation

1. Let  $k_1, \dots, k_r \in \mathbb{N}$  such that  $k_1 + \dots + k_r = n$ .
2. Any outcome that leads to  $X_j = k_j$  for all  $j \in \{1, \dots, r\}$  has proba  $p_1^{k_1} \dots p_r^{k_r}$ .
3. The number of such outcomes is given by (in book page 392)

$$\binom{n}{k_1, \dots, k_r} = \frac{n!}{k_1! \dots k_r!}$$

4. Therefore we get  $\mathbb{P}(X_1 = k_1, \dots, X_r = k_r) = \binom{n}{k_1, \dots, k_r} p_1^{k_1} \dots p_r^{k_r}$

## Multinomial Distribution

### Definition (Multinomial distribution)

Let  $n, r \in \mathbb{N}_*$ , let  $p_1, \dots, p_r \in (0, 1)$  s.t.  $p_1 + \dots + p_r = 1$ , then a r.v.  $X$  has a **multinomial distribution** with parameters  $n, r, p_1, \dots, p_r$  if it is defined for any  $k_1, \dots, k_r \in \mathbb{N}$  s.t.  $k_1 + \dots + k_r = n$  with probability

$$\mathbb{P}(X_1 = k_1, \dots, X_r = k_r) = \binom{n}{k_1, \dots, k_r} p_1^{k_1} \dots p_r^{k_r}$$

We denote it  $(X_1, \dots, X_r) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$ .

**Note:** For  $r = 2$ , we necessarily have  $X_2 = n - X_1$  and we retrieve the binomial

### Example

Roll a fair die with 6 faces 100 times

The probability that the  $i^{\text{th}}$  face appear is  $p_i = 1/6$ , s.t.  $p_1 + \dots + p_6 = 1$

Denote  $X_1, \dots, X_6$  the number of times face 1, ..., 6 appeared

Then  $(X_1, \dots, X_6) \sim \text{Multinom}(100, \underbrace{6, 1/6, \dots, 1/6}_{6 \text{ times}})$

## Exercise for next lecture

### Exercise

*Roll a normal die 100 times. Find the probability that among the 100 rolls, we observe exactly 22 ones, 17 fives.*

### Solution next lecture

Try on your own without looking at the book :)

# Joint Probability Distributions Continuous Case

Section 6.2

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 6, April 10th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Anouncements

### **Office hours** (After poll)

- ▶ Mondays 14:30 to 15:30 with T.A. Z. Yuan by Zoom
- ▶ Fridays 11:30 to 12:30 with instructor V. Roulet
- ▶ Register in advance to access the zoom session

### **Lecture material**

Updated slides with solutions given at the end of the lecture

## Answer Previous Exercise

### Exercise

*Roll a normal die 100 times. Find the probability that among the 100 rolls, we observe exactly 22 ones, 17 fives.*

## Answer Previous Exercise

### Exercise

*Roll a normal die 100 times. Find the probability that among the 100 rolls, we observe exactly 22 ones, 17 fives.*

**Solution** Denote  $X_1, X_5$  the number of times you get a 1 or a 5 resp. among 100 rolls  
We have  $\mathbb{P}(\text{"face is 1"}) = \mathbb{P}(\text{"face is 5"}) = 1/6$

We could model  $X_1, X_2, X_3, X_4, X_5, X_6$  as a multinomial but that can be simplified

Denote  $Y = X_2 + X_3 + X_4 + X_6$  the number of times you get any other face

We have  $\mathbb{P}(\text{"face is not 1 or 5"}) = 4/6 = 2/3$

Then  $(X_1, X_5, Y) \sim \text{Multinom}(100, 3, 1/6, 1/6, 2/3)$

$$\text{So } \mathbb{P}(X_1 = 22, X_5 = 17, Y = 100 - (22 + 17)) = \frac{100!}{22!17!6!1!} \left(\frac{1}{6}\right)^{22} \left(\frac{1}{6}\right)^{17} \left(\frac{2}{3}\right)^{61} \approx 0.0037$$

## Joint Probability Density Functions

### Definition (Joint probability density function)

Random variables  $X_1, \dots, X_n$  are **jointly continuous** if there exists a **joint probability density function**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  such that for any<sup>1</sup>  $B \subset \mathbb{R}^n$ ,

$$\mathbb{P}(X_1, \dots, X_n \in B) = \int_B \dots \int f(x_1, \dots, x_n) dx_1 \dots dx_n$$

### Note:

- ▶  $f(x_1, \dots, x_n) \geq 0$  and  $\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(x_1, \dots, x_n) dx_1 \dots dx_n = 1$
- ▶  $X$  and  $Y$  have a p.d.f. does not imply that  $(X, Y)$  is jointly continuous!

---

<sup>1</sup>Think of  $B$  as for example  $[a, b]^n$ . Again a rigorous definition requires  $B$  to belong to the Borel algebra of  $\mathbb{R}^n$

## Joint Probability Density Functions

### Definition (Joint probability density function)

Random variables  $X_1, \dots, X_n$  are **jointly continuous** if there exists a **joint probability density function**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  such that for any<sup>1</sup>  $B \subset \mathbb{R}^n$ ,

$$\mathbb{P}(X_1, \dots, X_n \in B) = \int_B \dots \int f(x_1, \dots, x_n) dx_1 \dots dx_n$$

### Note:

- ▶  $f(x_1, \dots, x_n) \geq 0$  and  $\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(x_1, \dots, x_n) dx_1 \dots dx_n = 1$
- ▶  $X$  and  $Y$  have a p.d.f. does not imply that  $(X, Y)$  is jointly continuous!

*Example:* Take  $X$  any continuous r.v., define  $Y = X$ , s.t.  $\mathbb{P}(X = Y) = 1$ . If  $(X, Y)$  had a joint p.d.f.  $f$ , denoting  $D = \{(x, y) : x = y\}$ , we would have

$$\mathbb{P}(X = Y) = \int_D \int f(x, y) dx dy = \int_{-\infty}^{+\infty} \left( \int_x^\infty f(x, y) dy \right) dx = 0$$

---

<sup>1</sup>Think of  $B$  as for example  $[a, b]^n$ . Again a rigorous definition requires  $B$  to belong to the Borel algebra of  $\mathbb{R}^n$

## Joint Probability Density Functions

### Lemma

Let  $X_1, \dots, X_n$  be  $n$  jointly continuous r.v.. Then for any subset  $A \subset \mathbb{R}^n$  included in a linear subspace  $E \subset \mathbb{R}^n$  of dimension  $\dim(E) = m < n$ ,

$$\mathbb{P}((X_1, \dots, X_n) \in A) = 0$$

## Joint Probability Density Functions

### Lemma

Let  $X_1, \dots, X_n$  be  $n$  jointly continuous r.v.. Then for any subset  $A \subset \mathbb{R}^n$  included in a linear subspace  $E \subset \mathbb{R}^n$  of dimension  $\dim(E) = m < n$ ,

$$\mathbb{P}((X_1, \dots, X_n) \in A) = 0$$

**Proof** General case requires change of variables, let's consider  $A=[a, b]^m \subset \mathbb{R}^n$ . Denote  $f$  the joint p.d.f. of  $(X_1, \dots, X_n)$ ,

$$\mathbb{P}((X_1, \dots, X_n) \in A) = \underbrace{\int_a^b \dots \int_a^b}_{m \text{ times}} \underbrace{\int_0^0 \dots \int_0^0}_{n-m \text{ times}} f(x_1, \dots, x_n) dx_1 \dots dx_n = 0$$

## Joint Probability Density Functions

### Example (Synthetic)

Assume  $X, Y$  have a joint p.d.f.

$$f(x, y) = \begin{cases} \frac{3}{2}(xy^2 + y) & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Check that it is a valid joint p.d.f.

## Joint Probability Density Functions

### Example (Synthetic)

Assume  $X, Y$  have a joint p.d.f.

$$f(x, y) = \begin{cases} \frac{3}{2}(xy^2 + y) & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Check that it is a valid joint p.d.f.

**Solution** We have  $f(x, y) \geq 0$  and

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy &= \frac{3}{2} \int_0^1 \left( \int_0^1 xy^2 + y dx \right) dy \\ &= \frac{3}{2} \int_0^1 \left( \frac{1}{2}y^2 + y \right) dy = \frac{3}{2} \left( \frac{1}{6} + \frac{1}{2} \right) = 1 \end{aligned}$$

## Joint Probability Density Functions

### Example (Synthetic)

Assume  $X, Y$  have a joint p.d.f.

$$f(x, y) = \begin{cases} \frac{3}{2}(xy^2 + y) & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Compute  $\mathbb{P}(X < Y)$

## Joint Probability Density Functions

### Example (Synthetic)

Assume  $X, Y$  have a joint p.d.f.

$$f(x, y) = \begin{cases} \frac{3}{2}(xy^2 + y) & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Compute  $\mathbb{P}(X < Y)$

### Solution

$$\begin{aligned}\mathbb{P}(X < Y) &= \frac{3}{2} \int_0^1 \left( \int_0^y (xy^2 + y) dx \right) dy \\ &= \frac{3}{2} \int_0^1 \left( \frac{1}{2}y^4 + y^2 \right) dy \\ &= \frac{3}{2} \left( \frac{1}{10} + \frac{1}{3} \right) = 0.65\end{aligned}$$

## Uniform Continuous Random Variable in higher dimensions

Definition (Uniform continuous random variable in dimension 2 or 3)

Let  $D$  be a bounded subset of  $\mathbb{R}^2$  s.t.  $\text{Area}(D) < +\infty$ . The random point  $(X, Y)$  is **uniformly distributed on**  $D$  if its joint p.d.f. reads

$$f(x, y) = \frac{1}{\text{Area}(D)} \mathbf{1}_D(x, y) = \begin{cases} \frac{1}{\text{Area}(D)} & \text{if } (x, y) \in D \\ 0 & \text{otherwise} \end{cases}$$

Let  $D$  be a bounded subset of  $\mathbb{R}^3$  s.t.  $\text{Vol}(D) < +\infty$ . The random point  $(X, Y, Z)$  is **uniformly distributed on**  $D$  if its joint p.d.f. reads

$$f(x, y, z) = \frac{1}{\text{Vol}(D)} \mathbf{1}_D(x, y, z) \begin{cases} \frac{1}{\text{Vol}(D)} & \text{if } (x, y, z) \in D \\ 0 & \text{otherwise} \end{cases}$$

We denote  $(X, Y) \sim \text{Unif}(D)$  or  $(X, Y, Z) \sim \text{Unif}(D)$ .

## Uniform Continuous Random Variable in higher dimensions

### Lemma

Let  $(X, Y) \sim \text{Unif}(D)$  for  $D \subset \mathbb{R}^2$ , then for any  $G \subset D$ , (similar for  $\mathbb{R}^3$ )

$$\mathbb{P}((X, Y) \in G) = \frac{\text{Area}(G)}{\text{Area}(D)}$$

## Uniform Continuous Random Variable in higher dimensions

### Lemma

Let  $(X, Y) \sim \text{Unif}(D)$  for  $D \subset \mathbb{R}^2$ , then for any  $G \subset D$ , (similar for  $\mathbb{R}^3$ )

$$\mathbb{P}((X, Y) \in G) = \frac{\text{Area}(G)}{\text{Area}(D)}$$

### Proof

$$\Pr((X, Y) \in G) = \frac{1}{\text{Area}(D)} \int \int \mathbf{1}_G(x, y) \mathbf{1}_D(x, y) dx dy = \int \int \mathbf{1}_G(x, y) dx dy = \frac{\text{Area}(G)}{\text{Area}(D)}$$

## Uniform Continuous Random Variable in higher dimensions

### Lemma

Let  $(X, Y) \sim \text{Unif}(D)$  for  $D \subset \mathbb{R}^2$ , then for any  $G \subset D$ , (similar for  $\mathbb{R}^3$ )

$$\mathbb{P}((X, Y) \in G) = \frac{\text{Area}(G)}{\text{Area}(D)}$$

### Proof

$$\Pr((X, Y) \in G) = \frac{1}{\text{Area}(D)} \int \int \mathbf{1}_G(x, y) \mathbf{1}_D(x, y) dx dy = \int \int \mathbf{1}_G(x, y) dx dy = \frac{\text{Area}(G)}{\text{Area}(D)}$$

### Example

Denote  $D_r = \{(x, y) : x^2 + y^2 < r^2\}$  a disk of radius  $r$

Throw a dart uniformly at random on a disk of radius 2

What is the probability that the dart is in the central disk of radius one?

## Uniform Continuous Random Variable in higher dimensions

### Lemma

Let  $(X, Y) \sim \text{Unif}(D)$  for  $D \subset \mathbb{R}^2$ , then for any  $G \subset D$ , (similar for  $\mathbb{R}^3$ )

$$\mathbb{P}((X, Y) \in G) = \frac{\text{Area}(G)}{\text{Area}(D)}$$

### Proof

$$\Pr((X, Y) \in G) = \frac{1}{\text{Area}(D)} \int \int \mathbf{1}_G(x, y) \mathbf{1}_D(x, y) dx dy = \int \int \mathbf{1}_G(x, y) dx dy = \frac{\text{Area}(G)}{\text{Area}(D)}$$

### Example

Denote  $D_r = \{(x, y) : x^2 + y^2 < r^2\}$  a disk of radius  $r$

Throw a dart uniformly at random on a disk of radius 2

What is the probability that the dart is in the central disk of radius one?

**Solution**  $(X, Y) \sim \text{Unif}(D_2)$

$$\mathbb{P}((X, Y) \in D_1) = \frac{\pi 1^2}{\pi 2^2} = \frac{1}{4}$$

## Joint Probability Density Functions

### Lemma

Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $X_1, \dots, X_n$  be jointly continuous r.v. with joint p.d.f.  $f$ ,

$$\mathbb{E}[g(x_1, \dots, x_n)] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n$$

### Example

Throw a dart uniformly at random on a square of edge size 2 centered on 0

Assume your score is equal to the square distance to the center

What is your average score?

## Joint Probability Density Functions

### Lemma

Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $X_1, \dots, X_n$  be jointly continuous r.v. with joint p.d.f.  $f$ ,

$$\mathbb{E}[g(x_1, \dots, x_n)] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n$$

### Example

Throw a dart uniformly at random on a square of edge size 2 centered on 0

Assume your score is equal to the square distance to the center

What is your average score?

**Solution**  $(X, Y) \sim \text{Unif}(S)$  with  $S = \{(x, y) : -1 \leq x \leq 1, -1 \leq y \leq 1\}$

Score is  $g(x, y) = x^2 + y^2$

Average score

$$\mathbb{E}[g(X, Y)] = \frac{1}{4} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) \mathbf{1}_S(x, y) dx dy = \frac{1}{4} \int_{-1}^1 \int_{-1}^1 (x^2 + y^2) dx dy = 2/3$$

## Marginal Probability Density Function

**Definition (Marginal probability density function)**

Let  $X, Y$  be jointly continuous r.v. and denote  $f_{X,Y}$  their joint p.d.f. then the p.d.f. of  $X$  exists and is given by

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy$$

## Marginal Probability Density Function

**Definition (Marginal probability density function)**

Let  $X, Y$  be jointly continuous r.v. and denote  $f_{X,Y}$  their joint p.d.f. then the p.d.f. of  $X$  exists and is given by

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy$$

**Proof** We have by definition of the joint p.d.f. an expression of the c.d.f. of  $X$  as

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}(X \leq t, -\infty \leq Y \leq +\infty) = \int_{-\infty}^t \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy dx$$

Therefore  $f_X(x) = F'_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy$

## Marginal Probability Density Function

### Example

Consider a disk of radius  $r$ ,  $D_r = \{(x, y) : x^2 + y^2 \leq r\}$  and  $(X, Y) \sim \text{Unif}(D_r)$ .  
What is the marginal p.d.f. of  $X$ ?

## Marginal Probability Density Function

### Example

Consider a disk of radius  $r$ ,  $D_r = \{(x, y) : x^2 + y^2 \leq r^2\}$  and  $(X, Y) \sim \text{Unif}(D_r)$ . What is the marginal p.d.f. of  $X$ ?

**Solution** Joint p.d.f. is  $f_{X,Y}(x, y) = \frac{1}{\pi r^2} \mathbf{1}_{D_r}(x, y)$  where  $D_r = \{(x, y) : x^2 + y^2 \leq r^2\}$   
Marginal density is then  $f_X(x) = 0$  for  $|x| > r$ , and for  $|x| \leq r$ ,

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy = \frac{1}{\pi r^2} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} dy = \frac{2}{\pi r^2} \sqrt{r^2 - x^2}$$

## Marginal Probability Density Function

Definition (Marginal probability density function)

Let  $X_1, \dots, X_n$  be jointly continuous and denote  $f$  their joint p.d.f..

Then for any  $j \in \{1, \dots, n\}$ ,  $X_j$  is a continuous random variable with p.d.f.

$$f_{X_j}(x) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \dots, x_{j-1}, x, x_{j+1}, \dots, x_n) dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_n$$

( $n-1$  integrals)

## Joint Cumulative Distribution

Definition (Joint cumulative distribution)

The **joint cumulative distribution** of r.v.  $X_1, \dots, X_n$  is defined as

$$\begin{aligned} F(t_1, \dots, t_n) &= \mathbb{P}(\{X_1 \leq t_1\} \cap \dots \cap \{X_n \leq t_n\}) \\ &\triangleq \mathbb{P}(X_1 \leq t_1, \dots, X_n \leq t_n) \end{aligned}$$

### Lemma

1. If  $(X, Y)$  are jointly continuous with joint p.d.f.  $f$ ,

$$F(t, s) = \int_{-\infty}^t \int_{-\infty}^s f(x, y) dy dx$$

2. If  $(X, Y)$  are jointly continuous (i.e. there exists a joint p.d.f.) with joint c.d.f.  $F$

$$\frac{\partial^2}{\partial t \partial s} F(t, s) \Big|_{s=x, t=y} = f(x, y)$$

## Borel algebra in $\mathbb{R}^{n*}$

### Formal details

- ▶ Until now, we defined proba. distributions on any  $B \subset \mathbb{R}^n$  for  $n=1$  or  $n>1$ .
- ▶ Formal definitions require to restrict our focus to subsets  $B \subset \mathbb{R}^n$  that form a  $\sigma$ -algebra  $\mathcal{B}$

### Definition ( $\sigma$ -algebra)

Let  $\Omega$  be a set, a  $\sigma$ -algebra  $\mathcal{F}$  on  $\Omega$  is a subset of  $2^\Omega = \{B \subset \Omega\}$  such that

1.  $\Omega \in \mathcal{F}$
2. (Stable by complementarity) For any  $A \in \mathcal{F}$ ,  $A^c \triangleq \Omega \setminus A \in \mathcal{F}$
3. (Stable by countable union) For any  $A_1, A_2, \dots \in \mathcal{F}$ ,  $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$

### Why introducing $\sigma$ -algebra?

You want the probability measure to satisfy that

- ▶ the measure is non-negative
- ▶ the measure of the union of disjoint sets is the sum of the measure of union sets

Then you can build a union of sets  $V_k$  (see e.g. Vitali set on Wikipedia) s.t.

$$[0, 1] \subset \bigcup_{k=1}^{+\infty} V_k \subset [-1, 2] \quad \mathbb{P}(V_k) = \lambda \geq 0 \quad \text{for all } k$$

which leads to  $1 \leq \sum_{k=1}^{+\infty} \mathbb{P}(V_k) \leq 3$  which is impossible

## Borel algebra in $\mathbb{R}^{n*}$

Formally, we restrict our focus on the Borel algebra of  $\mathbb{R}^n$

### Definition (Borel algebra in $\mathbb{R}^n$ )

The Borel algebra in  $\mathbb{R}^n$ , denoted  $\mathcal{B}_n$ , is the smallest  $\sigma$ -algebra (in terms of inclusion) that contains

- ▶ all product of intervals  $[a_1, b_1] \times \dots \times [a_n, b_n]$  for  $a_i \leq b_i \in \mathbb{R}$

or equivalently defined as the smallest  $\sigma$ -algebra that contains

- ▶ all product of intervals of the form  $(-\infty, a_1] \times \dots \times (-\infty, a_n]$  for  $a_i \in \mathbb{R}$ .

### Consequence

1. If we can measure all intervals of the form  $(-\infty, a_1] \times \dots \times (-\infty, a_n]$  for  $a_i \in \mathbb{R}$ , then we can measure all subsets of interests, i.e. all  $B \in \mathcal{B}_n$ ,  
→ we know all the information necessary to describe the proba distribution
2. All the information necessary to describe any r.v. is contained in its c.d.f.

## Quiz for next lecture

### Exercise

I am shooting an arrow on a target on a wall  $W = \{(x, y) : -1 \leq x \leq 1, 0 \leq y \leq 1\}$ . A wind affects my shoot from the left and the gravity also affects my shoot such that the position of the arrow has a p.d.f. proportional to  $\frac{e^x}{\sqrt{y+1}}$

What is the probability  
that I touch the target  $T = \{(x, y) : -0.1 \leq x \leq 0.1, 0.4 \leq y \leq 0.6\}$ ?

# Joint Probability Distributions & Independence

Section 6.3

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 7, April 13th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Optional exercises in homeworks

- ▶ Additional material for you to master the course
- ▶ Adds up to the grade of the homework up to the total score
- ▶ Taken into account for any recommendation letter

### Previous lectures

- ▶ Joint distributions, discrete and continuous cases

### This lecture

- ▶ Joint distributions and independence,
- ▶ Discrete independent random variables
- ▶ Continuous independent random variables
- ▶ Functions of independent random variables
- ▶ Minimum, maximum of independent random variables

## Answer Previous Quiz

### Exercise

I am shooting an arrow at a target on a wall  $W = \{(x, y) : -1 \leq x \leq 1, 0 \leq y \leq 1\}$ . Wind from the left and gravity affect my shot such that the position of the arrow has a p.d.f. proportional to  $\frac{e^x}{\sqrt{y+1}}$

What is the probability  
that I touch the target  $T = \{(x, y) : -0.1 \leq x \leq 0.1, 0.4 \leq y \leq 0.6\}$ ?

---

<sup>1</sup>When limits of an integral are not specified this means that the integral goes from  $-\infty$  to  $+\infty$

## Answer Previous Quiz

### Exercise

I am shooting an arrow at a target on a wall  $W = \{(x, y) : -1 \leq x \leq 1, 0 \leq y \leq 1\}$ .

Wind from the left and gravity affect my shot such that the position of the arrow has a p.d.f. proportional to  $\frac{e^x}{\sqrt{y+1}}$

What is the probability  
that I touch the target  $T = \{(x, y) : -0.1 \leq x \leq 0.1, 0.4 \leq y \leq 0.6\}$ ?

### Solution

1.  $f(x, y) = \frac{1}{\lambda} \frac{e^x}{\sqrt{y+1}} \mathbf{1}_W(x, y)$  with  $\lambda \geq 0$ , we have  $\int_{-\infty}^{+\infty} f(x, y) dx dy = 1$  and so<sup>1</sup>

$$\lambda = \int \int \frac{e^x}{\sqrt{y+1}} \mathbf{1}_W(x, y) dx dy = \int_0^1 \left( \int_{-1}^1 \frac{e^x}{\sqrt{y+1}} dx \right) dy = 2(\sqrt{2} - 1)(e - e^{-1})$$

2.

$$\mathbb{P}((X, Y) \in T) = \frac{1}{\lambda} \int_{0.4}^{0.6} \int_{-0.1}^{0.1} \frac{e^x}{\sqrt{y+1}} dx dy = \frac{2(\sqrt{1.6} - \sqrt{1.4})(e^{0.1} - e^{-0.1})}{\lambda} \approx 0.017$$

---

<sup>1</sup>When limits of an integral are not specified this means that the integral goes from  $-\infty$  to  $+\infty$

## Independent Random Variables

### Definition (Independent random variables)

Random variables  $X_1, \dots, X_n$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  are **independent** if for any<sup>2</sup> subsets  $B_1, \dots, B_n \subset \mathbb{R}$ ,

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \dots \mathbb{P}(X_n \in B_n)$$

or equivalently if their joint c.d.f.  $F$  factorizes into the marginal c.d.f. as

$$F(t_1, \dots, t_n) = F_{X_1}(t_1) \dots F_{X_n}(t_n)$$

---

<sup>2</sup>Again a formal definition requires these subsets to be Borel subsets of  $\mathbb{R}^n$

# Independent Random Variables

## Definition (Independent random variables)

Random variables  $X_1, \dots, X_n$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  are **independent** if for any<sup>2</sup> subsets  $B_1, \dots, B_n \subset \mathbb{R}$ ,

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \dots \mathbb{P}(X_n \in B_n)$$

or equivalently if their joint c.d.f.  $F$  factorizes into the marginal c.d.f. as

$$F(t_1, \dots, t_n) = F_{X_1}(t_1) \dots F_{X_n}(t_n)$$

**Proof** If they are independent then the joint c.d.f. factorizes by definition

If the c.d.f. factorizes into the marginals, the idea is that all the Borel subsets we want to measure can be generated by intervals of the form  $(-\infty, t]$  for  $t \in \mathbb{R}$  by taking intersections or unions of these intervals.

---

<sup>2</sup>Again a formal definition requires these subsets to be Borel subsets of  $\mathbb{R}^n$

# Independent Random Variables

## Definition (Independent random variables)

Random variables  $X_1, \dots, X_n$  on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  are **independent** if for any<sup>2</sup> subsets  $B_1, \dots, B_n \subset \mathbb{R}$ ,

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \dots \mathbb{P}(X_n \in B_n)$$

or equivalently if their joint c.d.f.  $F$  factorizes into the marginal c.d.f. as

$$F(t_1, \dots, t_n) = F_{X_1}(t_1) \dots F_{X_n}(t_n)$$

How can we understand independence  
by simply looking at the joint distribution?

What are the consequences in terms of p.m.f., p.d.f., c.d.f.?

---

<sup>2</sup>Again a formal definition requires these subsets to be Borel subsets of  $\mathbb{R}^n$

## Independent Discrete Random Variables

### Lemma

Let  $X_1, \dots, X_n$  be  $n$  discrete random variables on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Then  $X_1, \dots, X_n$  are independent if and only if their joint p.m.f.  $p$  factorizes into the marginals  $p_{X_i}$ ,

$$p(k_1, \dots, k_n) = p_{X_1}(k_1) \dots p_{X_n}(k_n)$$

# Independent Discrete Random Variables

## Lemma

Let  $X_1, \dots, X_n$  be  $n$  discrete random variables on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Then  $X_1, \dots, X_n$  are independent if and only if their joint p.m.f.  $p$  factorizes into the marginals  $p_{X_i}$ ,

$$p(k_1, \dots, k_n) = p_{X_1}(k_1) \dots p_{X_n}(k_n)$$

**Proof** If  $X_1, \dots, X_n$  are independent the result comes from the definition.  
If the joint p.m.f. factorizes into the marginal distributions, then

$$\begin{aligned}\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) &= \sum_{k_1 \in B_1, \dots, k_n \in B_n} p(k_1, \dots, k_n) \\ &= \sum_{k_1 \in B_1, \dots, k_n \in B_n} p_{X_1}(k_1) \dots p_{X_n}(k_n) \\ &= \left( \sum_{k_1 \in B_1} p_{X_1}(k_1) \right) \dots \left( \sum_{k_n \in B_n} p_{X_n}(k_n) \right) = \prod_{i=1}^n \mathbb{P}(X_i \in B_i)\end{aligned}$$

# Independent Discrete Random Variables

## Example

1. Roll two dices with **4 faces**, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Are  $S, Y$  independent?

|     | $Y$   |        |
|-----|-------|--------|
|     | 0     | 1      |
| 2   | 0     | $1/16$ |
| 3   | $1/8$ | 0      |
| 4   | $1/8$ | $1/16$ |
| $S$ | $1/4$ | 0      |
| 6   | $1/8$ | $1/16$ |
| 7   | $1/8$ | 0      |
| 8   | 0     | $1/16$ |

# Independent Discrete Random Variables

## Example

1. Roll two dices with **4 faces**, denote
  - (i)  $S$  the sum of the two dices
  - (ii)  $Y$  the indicator variable that you get a pair
2. Are  $S, Y$  independent?

|     | $Y$   |        |
|-----|-------|--------|
|     | 0     | 1      |
| 2   | 0     | $1/16$ |
| 3   | $1/8$ | 0      |
| 4   | $1/8$ | $1/16$ |
| $S$ | $1/4$ | 0      |
| 6   | $1/8$ | $1/16$ |
| 7   | $1/8$ | 0      |
| 8   | 0     | $1/16$ |

**Solution** Check for example  $\mathbb{P}(S = 2, Y = 0) = 0 \neq \mathbb{P}(S = 2)\mathbb{P}(Y = 0) > 0$

Note: one counterexample suffices to show that  $S, Y$  are dependent,

but to prove independence one would need to show the equality for all values of  $S, Y$

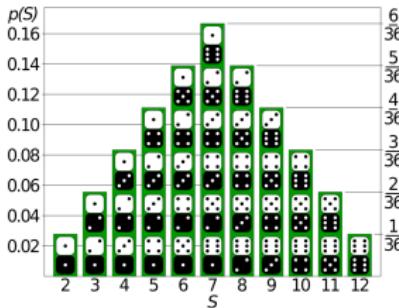
# Independent Discrete Random Variables

## Example

Roll repeatedly a pair of dice.

Denote  $N$  the number of rolls until the sum of the dice is 2 or a 6

1. What is the distribution of  $N$ ?
2. Denote  $X$  the sum you finally get (2 or 6), are  $X$  and  $N$  independent?



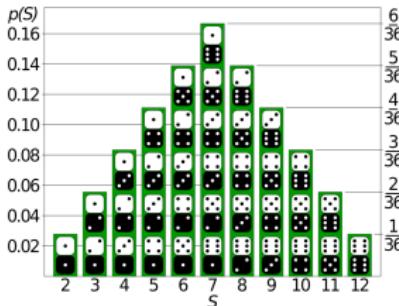
# Independent Discrete Random Variables

## Example

Roll repeatedly a pair of dice.

Denote  $N$  the number of rolls until the sum of the dice is 2 or a 6

1. What is the distribution of  $N$ ?
2. Denote  $X$  the sum you finally get (2 or 6), are  $X$  and  $N$  independent?



## Solution

1. Let  $Y_i$  be the sum of the two dice at the  $i^{\text{th}}$  roll.

We have  $\mathbb{P}(Y_i \in \{2, 6\}) = 1/36 + 5/36 = 1/6$  and so  $N \sim \text{Geom}(1/6)$

2.  $\mathbb{P}(N = n, X = 6) = \mathbb{P}(Y_1 \notin \{2, 6\}, \dots, Y_{n-1} \notin \{2, 6\}, Y_n = 6) = \left(\frac{5}{6}\right)^{n-1} \frac{5}{36}$

Therefore  $\mathbb{P}(X = 6) = \sum_{n=1}^{+\infty} \left(\frac{5}{6}\right)^{n-1} \frac{5}{36} = \frac{5/36}{1 - 5/6} = 5/6$

So  $\mathbb{P}(N = n, X = 6) = \left(\frac{5}{6}\right)^{n-1} \frac{1}{6} \frac{5}{6} = \mathbb{P}(N = n) \mathbb{P}(X = 6)$

Same argument shows  $\mathbb{P}(N = n, X = 2) = \mathbb{P}(N = n) \mathbb{P}(X = 2)$

$\rightarrow N$  and  $X$  are independent

# Independent Continuous Random Variables

## Lemma

Let  $X_1, \dots, X_n$  be  $n$  r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Assume that for  $j \in \{1, \dots, n\}$ , the rv.  $X_j$  has p.d.f.  $f_{X_j}$ .

1. If  $X_1, \dots, X_n$  have a joint p.d.f. that factorizes in the marginal p.d.f. as

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

then  $X_1, \dots, X_n$  are independent.

2. Conversely if  $X_1, \dots, X_n$  are independent then they are jointly continuous with joint p.d.f.

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

## Note:

1. Checking if  $X_1, \dots, X_n$  are independent can be done by looking at the joint p.d.f.
2. Conversely if they are independent, we know that they have a joint p.d.f. (remember that it was not always the case a priori)

# Independent Continuous Random Variables

## Lemma

Let  $X_1, \dots, X_n$  be  $n$  r.v. on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Assume that for  $j \in \{1, \dots, n\}$ , the rv.  $X_j$  has p.d.f.  $f_{X_j}$ .

1. If  $X_1, \dots, X_n$  have a joint p.d.f. that factorizes in the marginal p.d.f. as

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

then  $X_1, \dots, X_n$  are independent.

2. Conversely if  $X_1, \dots, X_n$  are independent then they are jointly continuous with joint p.d.f.

$$f(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

**Proof** For  $n = 2$  with two r.v.  $(X, Y)$ , denote  $A, B \subset \mathbb{R}$ ,

$$\begin{aligned}\mathbb{P}(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x, y) dy dx = \int_A \int_B f_X(x) f_Y(y) dy dx \\ &= \int_A f_X(x) dx \int_B f_Y(y) dy = \mathbb{P}(X \in A) \mathbb{P}(Y \in B)\end{aligned}$$

Conversely, if  $X, Y$  are independent

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B) = \int_A \int_B f_X(x) f_Y(y) dy dx$$

## Independent Continuous Random Variables

### Example (Shooting an arrow)

Consider  $X, Y$  with p.d.f.  $f(x, y) = \frac{1}{\lambda} \frac{e^x}{\sqrt{y+1}} \mathbf{1}_W(x, y)$  for  $\lambda=2(\sqrt{2}-1)(e - e^{-1})$   
where  $W=\{(x, y) : -1 \leq x \leq 1, 0 \leq y \leq 1\}$ .

1. Are  $X, Y$  independent?
2. What consequences it had when computing the probability to get the target  $T=\{(x, y) : -0.1 \leq x \leq 0.1, 0.4 \leq y \leq 0.6\}$ ?

# Independent Continuous Random Variables

## Example (Shooting an arrow)

Consider  $X, Y$  with p.d.f.  $f(x, y) = \frac{1}{\lambda} \frac{e^x}{\sqrt{y+1}} \mathbf{1}_W(x, y)$  for  $\lambda=2(\sqrt{2}-1)(e-e^{-1})$   
where  $W=\{(x, y) : -1 \leq x \leq 1, 0 \leq y \leq 1\}$ .

1. Are  $X, Y$  independent?
2. What consequences it had when computing the probability to get the target  $T=\{(x, y) : -0.1 \leq x \leq 0.1, 0.4 \leq y \leq 0.6\}$ ?

## Solution

1. Note that  $\mathbf{1}_W(x, y) = \mathbf{1}_{[-1,1]}(x) \mathbf{1}_{[0,1]}(y)$ ,

then one has  $f_X(x) = \frac{1}{e-e^{-1}} e^x \mathbf{1}_{[-1,1]}(x)$ ,  $f_Y(y) = \frac{1}{2(\sqrt{2}-1)\sqrt{y+1}} \mathbf{1}_{[0,1]}(y)$

So  $X, Y$  are independent

2.  $\mathbb{P}((X, Y) \in T) = \mathbb{P}(X \in [-0.1, 0.1]) \mathbb{P}(Y \in [0.4, 0.6])$  where  $\mathbb{P}(X \in [-0.1, 0.1])$ ,  $\mathbb{P}(Y \in [0.4, 0.6])$  can be computed from  $f_X, f_Y$  respectively.

## Quiz for next lecture

### Example

Let  $X, Y$  be two jointly continuous r.v., are  $X, Y$  independent if

1. their joint p.d.f. is  $f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$ ?
2. their joint p.d.f. is  $f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}}$  for  $-1 < \rho < 1$ ?

# Functions of Random Variables

Sections 5.2 6.4

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 8, April 15th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Previous lectures

- ▶ Joint distributions, discrete, continuous
- ▶ Joint distributions & independence,  
characterizations in terms of p.m.f./p.d.f.

In particular, if one has access to the joint p.m.f./p.d.f. of  $X, Y$ ,

marginal p.m.f./p.d.f.  $X$  is given by  
summing joint p.m.f./p.d.f. of  $(X, Y)$  over  $Y$

### This lecture

- ▶ Distributions of functions of random variables

## Quiz Previous Lecture

### Exercise

Let  $X, Y$  be two jointly continuous r.v., are  $X, Y$  independent if

1. their joint p.d.f. is  $f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$  ?
2. their joint p.d.f. is  $f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}}$  for  $-1 < \rho < 1$ ?

## Quiz Previous Lecture

### Exercise

Let  $X, Y$  be two jointly continuous r.v., are  $X, Y$  independent if

1. their joint p.d.f. is  $f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$ ?
2. their joint p.d.f. is  $f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}}$  for  $-1 < \rho < 1$ ?

### Solution

1. Yes, can easily check that  $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$  and  $f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$
2. We have

$$\begin{aligned} f_X(x) &= \frac{e^{-\frac{x^2}{2(1-\rho^2)}}}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} e^{-\frac{(y-\rho x)^2-\rho^2 x^2}{2(1-\rho^2)}} dy \\ &= \frac{e^{-\frac{x^2}{2}}}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}} dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \end{aligned}$$

which by the way proves that  $\int \int f_{X,Y}(x,y) dx dy = 1$

$$\text{Similarly } f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

Therefore  $X, Y$  are not independent

## Functions of Random Variables

### Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

## Functions of Random Variables

### Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

### Example

Let  $X \sim \text{Unif}(\{-1, 0, 1, 2\})$ , and  $Y = X^2$ . Distribution of  $Y$ ?

# Functions of Random Variables

## Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

## Example

Let  $X \sim \text{Unif}(\{-1, 0, 1, 2\})$ , and  $Y = X^2$ . Distribution of  $Y$ ?

## Solution

1.  $Y \in \{0, 1, 4\}$
- 2.

$$\mathbb{P}(Y = 0) = \mathbb{P}(X^2 = 0) = \mathbb{P}(X = 0) = \frac{1}{4}$$

$$\mathbb{P}(Y = 1) = \mathbb{P}(X = 1 \text{ or } X = -1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$\mathbb{P}(Y = 4) = \mathbb{P}(X = 2) = \frac{1}{4}$$

## Functions of Random Variables

### Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

# Functions of Random Variables

## Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

## Classical Method

1. Compute the c.d.f. of  $Y$ ,  $F_Y(t) = \mathbb{P}(Y \leq t)$

# Functions of Random Variables

## Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

## Classical Method

1. Compute the c.d.f. of  $Y$ ,  $F_Y(t) = \mathbb{P}(Y \leq t)$
2. Get
  - a. (*Discrete case*) if  $X$  is discrete, the p.m.f. of  $Y$  as

$$p_Y(k) = \mathbb{P}(k - 1 < Y \leq k) = \mathbb{P}(Y \leq k) - \mathbb{P}(Y \leq k - 1) = F_Y(k) - F_Y(k - 1)$$

- b. (*Continuous case*) if  $X$  is continuous, the p.d.f. of  $Y$  as

$$f_Y(y) = F'_Y(y)$$

# Functions of Random Variables

## Functions of random variables

Consider either

1. Let  $X$  be a r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  and denote  $Y = g(X)$
2. Let  $X_1, \dots, X_n$  be r.v.,  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and denote  $Y = g(X_1, \dots, X_n)$

What is the p.m.f./p.d.f. of  $Y$ ?

## Classical Method

1. Compute the c.d.f. of  $Y$ ,  $F_Y(t) = \mathbb{P}(Y \leq t)$
2. Get
  - a. (*Discrete case*) if  $X$  is discrete, the p.m.f. of  $Y$  as
$$p_Y(k) = \mathbb{P}(k - 1 < Y \leq k) = \mathbb{P}(Y \leq k) - \mathbb{P}(Y \leq k - 1) = F_Y(k) - F_Y(k - 1)$$
  - b. (*Continuous case*) if  $X$  is continuous, the p.d.f. of  $Y$  as

$$f_Y(y) = F'_Y(y)$$

## Why using the c.d.f. ?

Because operations on random variables can easily be expressed in the c.d.f.

## Functions of Random Variables

### Example

Let  $X$  be a continuous r.v. with joint p.d.f.  $f_X$

What is the p.d.f. of  $Y = aX + b$  with  $a \neq 0$ ?

## Functions of Random Variables

### Example

Let  $X$  be a continuous r.v. with joint p.d.f.  $f_X$

What is the p.d.f. of  $Y = aX + b$  with  $a \neq 0$ ?

### Solution

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(aX + b \leq t) = \begin{cases} \mathbb{P}(X \leq \frac{t-b}{a}) = F_X(\frac{t-b}{a}) & \text{if } a > 0 \\ \mathbb{P}(X \geq \frac{t-b}{a}) = 1 - F_X(\frac{t-b}{a}) & \text{if } a < 0 \end{cases}$$

$$\begin{aligned} f_Y(y) &= \begin{cases} \frac{1}{a}f_X(\frac{t-b}{a}) & \text{if } a > 0 \\ -\frac{1}{a}f_X(\frac{t-b}{a}) & \text{if } a < 0 \end{cases} \\ &= \frac{1}{|a|}f_X\left(\frac{t-b}{a}\right) \end{aligned}$$

## Function of one Random Variable

### Lemma

Let  $X$  be a continuous r.v. with p.d.f.  $f_X$

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be differentiable and strictly monotonic

with inverse denoted  $\gamma = g^{-1}$ , then the p.d.f of  $Y = g(X)$  exists<sup>1</sup> and it reads

$$f_Y(y) = \begin{cases} |\gamma'(y)| f_X(\gamma(y)) & \text{if } y \in g(\mathbb{R}) \\ 0 & \text{otherwise} \end{cases}$$

where  $\gamma'(y) = \frac{1}{g'(g^{-1}(y))}$

---

<sup>1</sup>We admit that fact

# Function of one Random Variable

## Lemma

Let  $X$  be a continuous r.v. with p.d.f.  $f_X$

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be differentiable and strictly monotonic

with inverse denoted  $\gamma = g^{-1}$ , then the p.d.f of  $Y = g(X)$  exists<sup>1</sup> and it reads

$$f_Y(y) = \begin{cases} |\gamma'(y)| f_X(\gamma(y)) & \text{if } y \in g(\mathbb{R}) \\ 0 & \text{otherwise} \end{cases}$$

where  $\gamma'(y) = \frac{1}{g'(g^{-1}(y))}$

**Proof** Denote  $a = \inf_x g(x)$ ,  $b = \sup_x g(x)$ , (potentially  $a = -\infty$ ,  $b = +\infty$ )

1. If  $t < a$ ,  $F_Y(t) = \mathbb{P}(g(X) \leq t) = 0$  so  $f_Y(t) = 0$
2. If  $t > b$ ,  $F_Y(t) = \mathbb{P}(g(X) \leq b) = 1$  so  $f_Y(t) = 0$
3. Since the probability on a point does not matter we can define  
 $f_Y(b) = f_Y(a) = 0$  if  $a, b$  are finite
4. if  $g$  is strictly increasing, for  $t \in (a, b)$  s.t.  $g^{-1}(t)$  is defined,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(g(X) \leq t) = \mathbb{P}(X \leq g^{-1}(t)) = F_X(\gamma(t))$$

$$\text{so } f_Y(t) = \gamma'(t) f_X(\gamma(t))$$

5. if  $g$  is strictly decreasing, for  $t \in (a, b)$  s.t.  $g^{-1}(t)$  is defined,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(g(X) \leq t) = \mathbb{P}(X \geq g^{-1}(t)) = 1 - F_X(\gamma(t))$$

$$\text{so } f_Y(t) = -\gamma'(t) f_X(\gamma(t))$$

For  $t \in (a, b)$ ,  $g \circ g^{-1}(t) = t$  so  $\gamma'(t) = \frac{1}{g'(g^{-1}(t))}$  so  $\gamma'(t) < 0$  for  $g$  decreasing.

<sup>1</sup>We admit that fact

## Function of one Random Variable

### Practical method

$X$  continuous r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  continuous<sup>2</sup>,  $Y = g(X)$ ,  $h_t = \mathbf{1}_{(-\infty, t]}$  for  $t \in \mathbb{R}$

---

<sup>2</sup>This ensures that  $Y$  is also a continuous r.v.

## Function of one Random Variable

### Practical method

$X$  continuous r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  continuous<sup>2</sup>,  $Y = g(X)$ ,  $h_t = \mathbf{1}_{(-\infty, t]}$  for  $t \in \mathbb{R}$

**Idea:** One one hand, using that  $\mathbf{1}_{(-\infty, t]}(g(x)) = \mathbf{1}_{\{x: g(x) \leq t\}}(x)$

$$F_Y(t) = \mathbb{P}(g(X) \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(g(X))] = \int_{-\infty}^{+\infty} h_t(g(x)) f_X(x) dx$$

On the other hand,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(Y)] = \mathbb{E}[h_t(Y)] = \int_{-\infty}^{+\infty} h_t(y) f_Y(y) dy$$

---

<sup>2</sup>This ensures that  $Y$  is also a continuous r.v.

## Function of one Random Variable

### Practical method

$X$  continuous r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  continuous<sup>2</sup>,  $Y = g(X)$ ,  $h_t = \mathbf{1}_{(-\infty, t]}$  for  $t \in \mathbb{R}$

**Idea:** One one hand, using that  $\mathbf{1}_{(-\infty, t]}(g(x)) = \mathbf{1}_{\{x: g(x) \leq t\}}(x)$

$$F_Y(t) = \mathbb{P}(g(X) \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(g(X))] = \int_{-\infty}^{+\infty} h_t(g(x)) f_X(x) dx$$

On the other hand,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(Y)] = \mathbb{E}[h_t(Y)] = \int_{-\infty}^{+\infty} h_t(y) f_Y(y) dy$$

**Principle:** To get  $f_Y(y)$ , it suffices to perform changes of variables in

$$\int_{-\infty}^{+\infty} h_t(g(x)) f_X(x) dx \quad \text{until getting something of the form} \quad \int_{-\infty}^{+\infty} h_t(y) \phi(y) dy$$

such that

$$f_Y(y) = F'_Y(t) = \phi(y)$$

---

<sup>2</sup>This ensures that  $Y$  is also a continuous r.v.

## Function of one Random Variable

### Example

Let  $X$  be a continuous r.v.,  $g : x \rightarrow x^2$ ,  $Y = g(x)$ . What is the p.d.f. of  $Y$ ?

### Note

1.  $g$  is not strictly monotonic on  $\mathbb{R}$  so we cannot apply previous lemma
2. but  $g$  is strictly decreasing on  $(-\infty, 0]$  and strictly increasing on  $[0, +\infty)$

## Function of one Random Variable

### Example

Let  $X$  be a continuous r.v.,  $g : x \rightarrow x^2$ ,  $Y = g(x)$ . What is the p.d.f. of  $Y$ ?

### Note

1.  $g$  is not strictly monotonic on  $\mathbb{R}$  so we cannot apply previous lemma
2. but  $g$  is strictly decreasing on  $(-\infty, 0]$  and strictly increasing on  $[0, +\infty)$

**Solution** For  $t \in \mathbb{R}$  and  $h_t = \mathbf{1}_{(-\infty, t]}$ ,

$$\begin{aligned}\mathbb{E}_X(h_t(g(X))) &= \int_{-\infty}^0 h_t(x^2) f_X(x) dx + \int_0^\infty h_t(x^2) f_X(x) dx \\ &= \int_0^{+\infty} h_t(x^2) f_X(-x) dx + \int_0^\infty h_t(x^2) f_X(x) dx\end{aligned}$$

On  $[0, +\infty)$   $g$  is invertible, so we can safely change variables  $y=x^2$ ,  $x=\sqrt{y}$ ,  $dx=\frac{1}{2\sqrt{y}}dy$

$$\mathbb{E}_X(h_t(g(X))) = \int_0^{+\infty} h_t(y) \frac{1}{2\sqrt{y}} (f_X(-\sqrt{y}) + f_X(\sqrt{y})) dy$$

Therefore  $f_Y(y) = (f_X(\sqrt{y}) + f_X(-\sqrt{y})) \frac{1}{2\sqrt{y}} \mathbf{1}_{[0, +\infty)}(y)$

## Function of Random Variables

### Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

## Function of Random Variables

### Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

1.  $g$  is invertible on  $S$  with inverse  $\gamma(u, v) = (\alpha(u, v), \beta(u, v))$   
 $(\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta : \mathbb{R}^2 \rightarrow \mathbb{R})$

## Function of Random Variables

### Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

1.  $g$  is invertible on  $S$  with inverse  $\gamma(u, v) = (\alpha(u, v), \beta(u, v))$   
 $(\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta : \mathbb{R}^2 \rightarrow \mathbb{R})$
2.  $\gamma$  is continuously differentiable on  $g(S)$  (partial derivatives are continuous)

## Function of Random Variables

### Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

1.  $g$  is invertible on  $S$  with inverse  $\gamma(u, v) = (\alpha(u, v), \beta(u, v))$   
 $(\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta : \mathbb{R}^2 \rightarrow \mathbb{R})$
2.  $\gamma$  is continuously differentiable on  $g(S)$  (partial derivatives are continuous)
3. The determinant of the Jacobian  $J_\gamma(u, v)$  of  $\gamma$  does not vanish on  $g(S)$ , where

$$J_\gamma(u, v) = \begin{pmatrix} \frac{\partial \alpha}{\partial u} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial u} & \frac{\partial \beta}{\partial v} \end{pmatrix}$$

Then  $(U, V) = g(X, Y)$  is jointly continuous with joint p.d.f.

$$f_{U,V}(u, v) = f_{X,Y}(\gamma(u, v)) |\det(J(u, v))| \mathbf{1}_{g(S)}(u, v)$$

# Function of Random Variables

## Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

1.  $g$  is invertible on  $S$  with inverse  $\gamma(u, v) = (\alpha(u, v), \beta(u, v))$   
 $(\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta : \mathbb{R}^2 \rightarrow \mathbb{R})$
2.  $\gamma$  is continuously differentiable on  $g(S)$  (partial derivatives are continuous)
3. The determinant of the Jacobian  $J_\gamma(u, v)$  of  $\gamma$  does not vanish on  $g(S)$ , where

$$J_\gamma(u, v) = \begin{pmatrix} \frac{\partial \alpha}{\partial u} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial u} & \frac{\partial \beta}{\partial v} \end{pmatrix}$$

Then  $(U, V) = g(X, Y)$  is jointly continuous with joint p.d.f.

$$f_{U,V}(u, v) = f_{X,Y}(\gamma(u, v)) |\det(J(u, v))| \mathbf{1}_{g(S)}(u, v)$$

**Proof** Denote  $h_{a,b} = \mathbf{1}_{(-\infty, a] \times (-\infty, b]}$  for  $a, b \in \mathbb{R}$ ,

Then the theorem comes from change of variables in 2 dimensions, such that

$$\int \int h_{a,b}(g(x, y)) f_{X,Y}(x, y) dx dy = \int \int h_{a,b}(u, v) f_{X,Y}(\gamma(u, v)) |\det(J(u, v))| du dv$$

# Function of Random Variables

## Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

1.  $g$  is invertible on  $S$  with inverse  $\gamma(u, v) = (\alpha(u, v), \beta(u, v))$   
 $(\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta : \mathbb{R}^2 \rightarrow \mathbb{R})$
2.  $\gamma$  is continuously differentiable on  $g(S)$  (partial derivatives are continuous)
3. The determinant of the Jacobian  $J_\gamma(u, v)$  of  $\gamma$  does not vanish on  $g(S)$ , where

$$J_\gamma(u, v) = \begin{pmatrix} \frac{\partial \alpha}{\partial u} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial u} & \frac{\partial \beta}{\partial v} \end{pmatrix}$$

Then  $(U, V) = g(X, Y)$  is jointly continuous with joint p.d.f.

$$f_{U,V}(u, v) = f_{X,Y}(\gamma(u, v)) |\det(J(u, v))| \mathbf{1}_{g(S)}(u, v)$$

## Note:

- To use such theorem (like previous one) it can be useful to first consider alternatives p.d.f. such that the mapping is one to one.
- Namely if there exists  $(x^*, y^*)$  s.t.  $f_{X,Y}(x^*, y^*) > 0$  and  $g$  is not invertible at that point, we can consider  $\tilde{f}_{X,Y}$  s.t.  $\tilde{f}_{X,Y} = f(x, y)$  if  $(x, y) \neq (x^*, y^*)$  and 0 o.w.
- This is still a valid p.d.f. for  $(X, Y)$  and the theorem can be applied with  $\tilde{f}_{X,Y}$

## Function of Random Variables

### Example

Let  $X, Y$  be two independent standard  $\text{Exp}(\lambda)$  r.v.

Find the joint p.d.f. of  $U = X + Y$  and  $V = \frac{X}{X+Y}$

## Function of Random Variables

### Example

Let  $X, Y$  be two independent standard  $\text{Exp}(\lambda)$  r.v.

Find the joint p.d.f. of  $U = X + Y$  and  $V = \frac{X}{X+Y}$

**Solution** Classical joint p.d.f. of  $(X, Y)$  is  $f_{X,Y}^0(x,y) = \lambda^2 e^{-\lambda(x+y)} \mathbf{1}_{[0,+\infty)^2}(x,y)$

We rather consider  $f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} \mathbf{1}_{(0,+\infty)^2}(x,y)$  which defines same distrib.

$g : (x,y) \rightarrow (x+y, \frac{x}{x+y})$  well defined on  $(0,+\infty)^2$  and  $g((0,+\infty)^2) = (0,+\infty) \times (0,1)$

Inverse mapping is given by

$$u = x + y, \quad v = \frac{x}{x+y} \iff x = \alpha(u,v) = uv, \quad y = \beta(u,v) = (1-v)u,$$

$$\det(J_\gamma(u,v)) = \det \begin{pmatrix} \frac{\partial \alpha}{\partial u} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial u} & \frac{\partial \beta}{\partial v} \end{pmatrix} = \det \begin{pmatrix} v & u \\ 1-v & -u \end{pmatrix} = (v-1)u - uv = -u$$

Applying the formula

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(\gamma(u,v)) |\det(J(u,v))| \mathbf{1}_{g(S)}(u,v) \\ &= \lambda^2 u e^{-\lambda u} \mathbf{1}_{(0,+\infty) \times (0,1)}(u,v) \end{aligned}$$

# Functions of Independent Random Variables

## Minimum, Maximum

Section 6.3

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 9, April 17th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Homeworks

- ▶ Homework 1 due tonight
- ▶ Homework 2 due Wednesday

### Course feedback

- ▶ A course feedback is available at <https://uw.iasystem.org/survey/222954>
- ▶ Please complete it before **Monday, 27th April**

### Previous lecture

- ▶ Function of random variables, change of distributions

### This lecture

- ▶ Functions of independent random variables
- ▶ Min, max of independent random variables
- ▶ Sum of independent random variables

## Function of one Random Variable

### Practical method

$X$  continuous r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  continuous<sup>1</sup>,  $Y = g(X)$ ,  $h_t = \mathbf{1}_{(-\infty, t]}$  for  $t \in \mathbb{R}$

---

<sup>1</sup>This ensures that  $Y$  is also a continuous r.v.

## Function of one Random Variable

### Practical method

$X$  continuous r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  continuous<sup>1</sup>,  $Y = g(X)$ ,  $h_t = \mathbf{1}_{(-\infty, t]}$  for  $t \in \mathbb{R}$

**Idea:** One one hand, using  $\mathbf{1}_{(-\infty, t]}(g(x)) = \mathbf{1}_{\{x: g(x) \leq t\}}(x)$ ,  $\mathbb{E}(\mathbf{1}_B(X)) = \mathbb{P}(X \in B)$ ,

$$F_Y(t) = \mathbb{P}(g(X) \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(g(X))] = \int_{-\infty}^{+\infty} h_t(g(x)) f_X(x) dx$$

On the other hand,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(Y)] = \mathbb{E}[h_t(Y)] = \int_{-\infty}^{+\infty} h_t(y) f_Y(y) dy$$

---

<sup>1</sup>This ensures that  $Y$  is also a continuous r.v.

## Function of one Random Variable

### Practical method

$X$  continuous r.v.,  $g : \mathbb{R} \rightarrow \mathbb{R}$  continuous<sup>1</sup>,  $Y = g(X)$ ,  $h_t = \mathbf{1}_{(-\infty, t]}$  for  $t \in \mathbb{R}$

**Idea:** One one hand, using  $\mathbf{1}_{(-\infty, t]}(g(x)) = \mathbf{1}_{\{x: g(x) \leq t\}}(x)$ ,  $\mathbb{E}(\mathbf{1}_B(X)) = \mathbb{P}(X \in B)$ ,

$$F_Y(t) = \mathbb{P}(g(X) \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(g(X))] = \int_{-\infty}^{+\infty} h_t(g(x)) f_X(x) dx$$

On the other hand,

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{E}[\mathbf{1}_{(-\infty, t]}(Y)] = \mathbb{E}[h_t(Y)] = \int_{-\infty}^{+\infty} h_t(y) f_Y(y) dy$$

**Principle:** To get  $f_Y(y)$ , it suffices to perform changes of variables in

$$\int_{-\infty}^{+\infty} h_t(g(x)) f_X(x) dx \quad \text{until getting something of the form} \quad \int_{-\infty}^{+\infty} h_t(y) \phi(y) dy$$

such that

$$f_Y(y) = F'_Y(t) = \phi(y)$$

---

<sup>1</sup>This ensures that  $Y$  is also a continuous r.v.

## Function of Random Variables

### Theorem

Let  $(X, Y)$  jointly continuous with p.d.f.  $f_{X,Y}$ , denote  $S = \{(x, y) : f_{X,Y}(x, y) > 0\}$

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

1.  $g$  is invertible on  $S$  with inverse  $\gamma(u, v) = (\alpha(u, v), \beta(u, v))$   
 $(\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta : \mathbb{R}^2 \rightarrow \mathbb{R})$
2.  $\gamma$  is continuously differentiable on  $g(S)$  (partial derivatives are continuous)
3. The determinant of the Jacobian  $J_\gamma(u, v)$  of  $\gamma$  does not vanish on  $g(S)$ , where

$$J_\gamma(u, v) = \begin{pmatrix} \frac{\partial \alpha}{\partial u} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial u} & \frac{\partial \beta}{\partial v} \end{pmatrix}$$

Then  $(U, V) = g(X, Y)$  is jointly continuous with joint p.d.f.

$$f_{U,V}(u, v) = f_{X,Y}(\gamma(u, v)) |\det(J(u, v))| \mathbf{1}_{g(S)}(u, v)$$

## Function of Random Variables

### Example

Let  $X, Y$  be two independent standard  $\text{Exp}(\lambda)$  r.v.

Find the joint p.d.f. of  $U = X + Y$  and  $V = \frac{X}{X+Y}$

1. Are  $U, V$  independent?
2. What distributions follow  $U, V$ ?

# Function of Random Variables

## Example

Let  $X, Y$  be two independent standard  $\text{Exp}(\lambda)$  r.v.

Find the joint p.d.f. of  $U = X + Y$  and  $V = \frac{X}{X+Y}$

1. Are  $U, V$  independent?
2. What distributions follow  $U, V$ ?

**Solution** Classical joint p.d.f. of  $(X, Y)$  is  $f_{X,Y}^0(x,y) = \lambda^2 e^{-\lambda(x+y)} \mathbf{1}_{[0,+\infty)^2}(x,y)$

We rather consider  $f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} \mathbf{1}_{(0,+\infty)^2}(x,y)$  which defines same distrib.

$g : (x,y) \rightarrow (x+y, \frac{x}{x+y})$  well defined on  $(0,+\infty)^2$  and  $g((0,+\infty)^2) = (0,+\infty) \times (0,1)$

Inverse mapping is given by

$$u = x + y, \quad v = \frac{x}{x+y} \iff x = \alpha(u,v) = uv, \quad y = \beta(u,v) = (1-v)u,$$

$$\det(J_\gamma(u,v)) = \det \begin{pmatrix} \frac{\partial \alpha}{\partial u} & \frac{\partial \alpha}{\partial v} \\ \frac{\partial \beta}{\partial u} & \frac{\partial \beta}{\partial v} \end{pmatrix} = \det \begin{pmatrix} v & u \\ 1-v & -u \end{pmatrix} = (v-1)u - uv = -u$$

Applying the formula

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(\gamma(u,v)) |\det(J(u,v))| \mathbf{1}_{g(S)}(u,v) \\ &= \lambda^2 u e^{-\lambda u} \mathbf{1}_{(0,+\infty) \times (0,1)}(u,v) \end{aligned}$$

$U, V$  are independent with  $U \sim \text{Gamma}(2, \lambda)$ ,  $V \sim \text{Unif}((0,1))$

## Functions of Independent Random Variables

### Lemma

Let  $X_1, \dots, X_{m+n}$  be  $m+n$  independent r.v. (discrete or continuous).

Let  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  and  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ .

Then  $Y = g(X_1, \dots, X_m)$  and  $Z = h(X_{m+1}, \dots, X_{m+n})$  are independent.

### Example

Let  $B_1, \dots, B_{m+n} \sim \text{Ber}(p)$  be  $m+n$  independent random variables.

Denote  $S_1 = \sum_{i=1}^m B_i$  and  $S_2 = \sum_{i=m+1}^n B_i$ , then  $S_1$  and  $S_2$  are independent.  
Are  $Z = S_1 + S_2$  and  $S_1$  independent?

## Functions of Independent Random Variables

### Lemma

Let  $X_1, \dots, X_{m+n}$  be  $m+n$  independent r.v. (discrete or continuous).

Let  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  and  $h : \mathbb{R}^n \rightarrow \mathbb{R}$ .

Then  $Y = g(X_1, \dots, X_m)$  and  $Z = h(X_{m+1}, \dots, X_{m+n})$  are independent.

### Example

Let  $B_1, \dots, B_{m+n} \sim \text{Ber}(p)$  be  $m+n$  independent random variables.

Denote  $S_1 = \sum_{i=1}^m B_i$  and  $S_2 = \sum_{i=m+1}^n B_i$ , then  $S_1$  and  $S_2$  are independent.  
Are  $Z = S_1 + S_2$  and  $S_1$  independent?

**Solution**  $\mathbb{P}(S_1 = 1, Z = 0) = 0 \neq \mathbb{P}(S_1 = 1) \mathbb{P}(Z = 0) > 0$

## Minimum, Maximum of Independent Random Variables

### Lemma

Let  $X_1, \dots, X_n$  be  $n$  independent random variables.

Denote  $Y = \max(X_1, \dots, X_n)$  and  $Z = \min(X_1, \dots, X_n)$ , then

$$F_Y(t) = \prod_{i=1}^n F_{X_i}(t) \quad 1 - F_Z(t) = \prod_{i=1}^n (1 - F_{X_i}(t))$$

# Minimum, Maximum of Independent Random Variables

## Lemma

Let  $X_1, \dots, X_n$  be  $n$  independent random variables.

Denote  $Y = \max(X_1, \dots, X_n)$  and  $Z = \min(X_1, \dots, X_n)$ , then

$$F_Y(t) = \prod_{i=1}^n F_{X_i}(t) \quad 1 - F_Z(t) = \prod_{i=1}^n (1 - F_{X_i}(t))$$

## Proof

$$F_Y(t) = \mathbb{P}(\max(X_1, \dots, X_n) \leq t) = \mathbb{P}(X_1 \leq t, \dots, X_n \leq t) = \prod_{i=1}^n \mathbb{P}(X_i \leq t) = \prod_{i=1}^n F_{X_i}(t)$$

Similarly  $\mathbb{P}(\min(X_1, \dots, X_n) > t) = \mathbb{P}(X_1 > t, \dots, X_n > t) = \prod_{i=1}^n \mathbb{P}(X_i > t)$ ,  
hence the second result

## Minimum, Maximum of Independent Random Variables

### Example

Let  $X_1, \dots, X_n$  be  $n$  independent r.v. following  $X_i \sim \text{Geom}(p_i)$ ,  $p_i \in (0, 1)$

What is the p.m.f. of  $Y = \min(X_1, \dots, X_n)$ ?

## Minimum, Maximum of Independent Random Variables

### Example

Let  $X_1, \dots, X_n$  be  $n$  independent r.v. following  $X_i \sim \text{Geom}(p_i)$ ,  $p_i \in (0, 1)$   
What is the p.m.f. of  $Y = \min(X_1, \dots, X_n)$ ?

**Solution** For  $k \in \mathbb{N}$ ,  $1 - F_{X_i}(k) = \mathbb{P}(X_i > k) = (1 - p_i)^k$ ,

So, by previous lemma,  $1 - F_Y(k) = \mathbb{P}(Y > k) = \prod_{i=1}^n (1 - p_i)^k$

Then denoting  $q = \prod_{i=1}^n (1 - p_i)$  and  $r = 1 - q$ ,

$$\mathbb{P}(Y = k) = \mathbb{P}(Y > k - 1) - \mathbb{P}(Y > k) = q^{k-1} - q^k = q^{k-1}(1 - q) = (1 - r)^{k-1}r$$

So we recognize  $Y \sim \text{Geom}(r)$ .

## Quiz for next lecture

### Exercise

Let  $X \sim \text{Exp}(\lambda)$ ,  $Y \sim \text{Exp}(\mu)$ , what is the distribution of  $\min(X, Y)$ ?

# Sums of Independent Random Variables

Section 7.1

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 10, April 20th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

Course feedback at <https://uw.iasystem.org/survey/222954> until **Monday, 27th April**

## Homework

- ▶ Solution to homework 1 has been published

## Mid-term

- ▶ Mid-term will be available on Friday for one day
- ▶ Once started you will have 60 minutes to complete the homework and 15 additional minutes to send your solutions
- ▶ Next session (on Wednesday) will review course until now

## Previous lecture

- ▶ Functions of independent random variables
- ▶ Min, max of independent random variables

## This lecture

- ▶ Sum of independent random variables
- ▶ (Exchangeable random variables)

## Quiz Previous Lecture

### Exercise

*Let  $X \sim \text{Exp}(\lambda)$ ,  $Y \sim \text{Exp}(\mu)$ , what is the distribution of  $\min(X, Y)$ ?*

## Quiz Previous Lecture

### Exercise

Let  $X \sim \text{Exp}(\lambda)$ ,  $Y \sim \text{Exp}(\mu)$ , what is the distribution of  $\min(X, Y)$ ?

**Solution** Let  $Z = \min(X, Y)$ ,

$$1 - F_Z(t) = \mathbb{P}(Z > t) = \mathbb{P}(X > t, Y > t) = \mathbb{P}(X > t) \mathbb{P}(Y > t) = e^{-\lambda t} e^{-\mu t}$$

So  $f_Z(t) = F'_Z(t) = (\lambda + \mu)e^{-(\lambda+\mu)t}$ , i.e.,  $Z \sim \text{Exp}(\lambda + \mu)$

## Sums of Independent Random Variables

### Lemma

Let  $X, Y$  be two independent random variables.

1. If  $X, Y$  are discrete r.v. with p.m.f.  $p_X, p_Y$  (defined w.l.o.g. on  $\mathbb{Z}$ ), then for  $n \in \mathbb{Z}$ ,

$$p_{X+Y}(n) = \sum_{k \in \mathbb{Z}} p_X(k)p_Y(n - k) = \sum_{k \in \mathbb{Z}} p_X(n - k)p_Y(k)$$

# Sums of Independent Random Variables

## Lemma

Let  $X, Y$  be two independent random variables.

1. If  $X, Y$  are discrete r.v. with p.m.f.  $p_X, p_Y$  (defined w.l.o.g. on  $\mathbb{Z}$ ), then for  $n \in \mathbb{Z}$ ,

$$p_{X+Y}(n) = \sum_{k \in \mathbb{Z}} p_X(k)p_Y(n-k) = \sum_{k \in \mathbb{Z}} p_X(n-k)p_Y(k)$$

## Proof

$$\mathbb{P}(X + Y = n) = \sum_{k \in \mathbb{Z}} \mathbb{P}(X = k, Y = n - k) = \sum_{k \in \mathbb{Z}} \mathbb{P}(X = k) \mathbb{P}(Y = n - k)$$

$$\text{or } \mathbb{P}(X + Y = n) = \sum_{k \in \mathbb{Z}} \mathbb{P}(X = n - k, Y = k) = \sum_{k \in \mathbb{Z}} \mathbb{P}(X = n - k) \mathbb{P}(Y = k)$$

# Sums of Independent Random Variables

## Lemma

Let  $X, Y$  be two independent random variables.

1. If  $X, Y$  are discrete r.v. with p.m.f.  $p_X, p_Y$  (defined w.l.o.g. on  $\mathbb{Z}$ ), then for  $n \in \mathbb{Z}$ ,

$$p_{X+Y}(n) = \sum_{k \in \mathbb{Z}} p_X(k)p_Y(n-k) = \sum_{k \in \mathbb{Z}} p_X(n-k)p_Y(k)$$

2. If  $X, Y$  are continuous r.v. with p.d.f.  $f_X, f_Y$ , then for  $x \in \mathbb{R}$ ,

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{+\infty} f_X(z-x)f_Y(x)dx$$

# Sums of Independent Random Variables

## Lemma

Let  $X, Y$  be two independent random variables.

1. If  $X, Y$  are discrete r.v. with p.m.f.  $p_X, p_Y$  (defined w.l.o.g. on  $\mathbb{Z}$ ), then for  $n \in \mathbb{Z}$ ,

$$p_{X+Y}(n) = \sum_{k \in \mathbb{Z}} p_X(k)p_Y(n-k) = \sum_{k \in \mathbb{Z}} p_X(n-k)p_Y(k)$$

2. If  $X, Y$  are continuous r.v. with p.d.f.  $f_X, f_Y$ , then for  $x \in \mathbb{R}$ ,

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{+\infty} f_X(z-x)f_Y(x)dx$$

## Proof

$$\begin{aligned} F_{X+Y}(z) &= \mathbb{P}(X + Y \leq z) = \int \int_{x+y \leq z} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{z-x} f_{X,Y}(x,y)dy \right) dx \\ &= \int_{-\infty}^{+\infty} \left( \int_{-\infty}^z f_X(x)f_Y(w-x)dw \right) dx = \int_{-\infty}^z \left( \int_{-\infty}^{+\infty} f_X(x)f_Y(w-x)dx \right) dw \end{aligned}$$

$$\text{Therefore } f_{X+Y}(z) = F'_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx$$

# Sums of Independent Random Variables

## Lemma

Let  $X, Y$  be two independent random variables.

1. If  $X, Y$  are discrete r.v. with p.m.f.  $p_X, p_Y$  (defined w.l.o.g. on  $\mathbb{Z}$ ), then for  $n \in \mathbb{Z}$ ,

$$p_{X+Y}(n) = \sum_{k \in \mathbb{Z}} p_X(k)p_Y(n-k) = \sum_{k \in \mathbb{Z}} p_X(n-k)p_Y(k) \triangleq p_X \star p_Y(n)$$

2. If  $X, Y$  are continuous r.v. with p.d.f.  $f_X, f_Y$ , then for  $x \in \mathbb{R}$ ,

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{+\infty} f_X(z-x)f_Y(x)dx \triangleq f_X \star f_Y(z)$$

**Note:** The  $\star$  operation is called a convolution.

So summing two random variables amount to convolve their p.m.f./p.d.f.

## Sums of Independent Random Variables

### Example (Convolution of Poisson Random Variables)

1. Let  $X \sim \text{Poisson}(\lambda)$  and  $Y \sim \text{Poisson}(\mu)$  be independent.  
What is the distribution of  $Z = X + Y$ ?
2. Suppose a factory experiences on average 1 accident per month and that this number of accidents is Poisson distributed.  
What is the proba. that during a period of 2 months, there are 3 accidents?

## Sums of Independent Random Variables

### Example (Convolution of Poisson Random Variables)

1. Let  $X \sim \text{Poisson}(\lambda)$  and  $Y \sim \text{Poisson}(\mu)$  be independent.  
What is the distribution of  $Z = X + Y$ ?
2. Suppose a factory experiences on average 1 accident per month and that this number of accidents is Poisson distributed.  
What is the proba. that during a period of 2 months, there are 3 accidents?

### Solution

1.  $Z \sim \text{Poisson}(\lambda + \mu)$

$$\begin{aligned}\mathbb{P}(Z = n) &= \sum_{k \in \mathbb{Z}} \mathbb{P}(X = k) \mathbb{P}(Y = n - k) = \sum_{k=0}^n \mathbb{P}(X = k) \mathbb{P}(Y = n - k) \\ &= \sum_{k=0}^n e^{-\lambda} \frac{\lambda^k}{k!} e^{-\mu} \frac{\mu^{n-k}}{(n-k)!} = \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \frac{n!}{k!(n-k)!} \lambda^k \mu^{n-k} \\ &= e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^n}{n!}\end{aligned}$$

## Sums of Independent Random Variables

### Example (Convolution of Poisson Random Variables)

1. Let  $X \sim \text{Poisson}(\lambda)$  and  $Y \sim \text{Poisson}(\mu)$  be independent.  
What is the distribution of  $Z = X + Y$ ?
2. Suppose a factory experiences on average 1 accident per month and that this number of accidents is Poisson distributed.  
What is the proba. that during a period of 2 months, there are 3 accidents?

### Solution

1.  $Z \sim \text{Poisson}(\lambda + \mu)$

$$\begin{aligned}\mathbb{P}(Z = n) &= \sum_{k \in \mathbb{Z}} \mathbb{P}(X = k) \mathbb{P}(Y = n - k) = \sum_{k=0}^n \mathbb{P}(X = k) \mathbb{P}(Y = n - k) \\ &= \sum_{k=0}^n e^{-\lambda} \frac{\lambda^k}{k!} e^{-\mu} \frac{\mu^{n-k}}{(n-k)!} = \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \frac{n!}{k!(n-k)!} \lambda^k \mu^{n-k} \\ &= e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^n}{n!}\end{aligned}$$

2. Number of accidents during a period of 2 months is  $Z = X_1 + X_2$  where  $X_i$  is the number of month during month  $i$ . So  $Z \sim \text{Poisson}(2)$  and

$$\mathbb{P}(Z = 3) = e^{-2} \frac{2^3}{3!} \approx 0.18$$

## Sums of Independent Random Variables

Example (Convolution of Binomial)

$X \sim \text{Bin}(m_1, p)$  and  $Y \sim \text{Bin}(m_2, p)$  independent. Distribution of  $X + Y$ ?

## Sums of Independent Random Variables

### Example (Convolution of Binomial)

$X \sim \text{Bin}(m_1, p)$  and  $Y \sim \text{Bin}(m_2, p)$  independent. Distribution of  $X + Y$ ?

#### Solution

$X = \sum_{i=1}^{m_1} B_i, Y = \sum_{j=1}^{m_2} C_j$  where  $B_i \sim \text{Ber}(p), C_i \sim \text{Ber}(p)$  are independent  
So  $X + Y \sim \text{Bin}(m_1 + m_2, p)$

## Sums of Independent Random Variables

### Example (Negative binomial)

1.  $X \sim \text{Geom}(p)$ ,  $Y \sim \text{Geom}(p)$  independent. Distribution of  $X + Y$ ?
2.  $X_i \sim \text{Geom}(p)$   $i \in \{1, \dots, p\}$  independent. Distribution of  $Z = X_1 + \dots + X_m$ ?

# Sums of Independent Random Variables

## Example (Negative binomial)

1.  $X \sim \text{Geom}(p)$ ,  $Y \sim \text{Geom}(p)$  independent. Distribution of  $X + Y$ ?
2.  $X_i \sim \text{Geom}(p)$   $i \in \{1, \dots, p\}$  independent. Distribution of  $Z = X_1 + \dots + X_m$ ?

### Solution

1.  $X(\Omega) = \{1, \dots, \}$ , same for  $Y(\Omega)$  so  $(X + Y)(\Omega) = \{2, \dots\}$

$$\begin{aligned}\mathbb{P}(X + Y = n) &= \sum_{k=-\infty}^{+\infty} \mathbb{P}(X = k) \mathbb{P}(Y = n - k) \\ &= \sum_{k=1}^{n-1} p(1-p)^{k-1} p(1-p)^{n-k-1} = (n-1)p^2(1-p)^{n-2}\end{aligned}$$

# Sums of Independent Random Variables

## Example (Negative binomial)

1.  $X \sim \text{Geom}(p)$ ,  $Y \sim \text{Geom}(p)$  independent. Distribution of  $X + Y$ ?
2.  $X_i \sim \text{Geom}(p)$   $i \in \{1, \dots, p\}$  independent. Distribution of  $Z = X_1 + \dots + X_m$ ?

### Solution

1.  $X(\Omega) = \{1, \dots, \}$ , same for  $Y(\Omega)$  so  $(X + Y)(\Omega) = \{2, \dots\}$

$$\begin{aligned}\mathbb{P}(X + Y = n) &= \sum_{k=-\infty}^{+\infty} \mathbb{P}(X = k) \mathbb{P}(Y = n - k) \\ &= \sum_{k=1}^{n-1} p(1-p)^{k-1} p(1-p)^{n-k-1} = (n-1)p^2(1-p)^{n-2}\end{aligned}$$

2. (Optional)

$$\begin{aligned}\{Z = n\} &= \{\text{"among the } n-1 \text{ first trials there were } m-1 \text{ successes"}\} \\ &\quad \cap \{\text{"the } n^{\text{th}} \text{ trial gives the } m^{\text{th}} \text{ success"}\}\end{aligned}$$

$$\text{So } \mathbb{P}(Z = n) = \binom{n-1}{m-1} p^{m-1} (1-p)^{n-m} p = \binom{n-1}{m-1} p^m (1-p)^{n-m}$$

$Z$  is called a negative binomial distribution, denoted  $Z \sim \text{Negbin}(m, p)$

## Sums of Independent Random Variables

### Lemma

Let  $X_1, \dots, X_n$  be independent Gaussian variables  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , then

$$X_1 + \dots + X_n \sim \mathcal{N}(\mu_1 + \dots + \mu_n, \sigma_1^2 + \dots + \sigma_n^2)$$

# Sums of Independent Random Variables

## Lemma

Let  $X_1, \dots, X_n$  be independent Gaussian variables  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , then

$$X_1 + \dots + X_n \sim \mathcal{N}(\mu_1 + \dots + \mu_n, \sigma_1^2 + \dots + \sigma_n^2)$$

**Solution** Suffices to prove it for  $n=2$ , for  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ ,  $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$  independent  $X_1 = \sigma_1 Z_1 + \mu_1$ ,  $X_2 = \sigma_2 Z_2 + \mu_2$ , with  $Z_1 \sim \mathcal{N}(0, 1)$ ,  $Z_2 \sim \mathcal{N}(0, 1)$

$Z_1, Z_2$  are independent as functions of independent random variables  $(Z_i = \frac{X_i - \mu_i}{\sigma_i})$

We have  $X_1 + X_2 = \sigma_1 (Z_1 + \frac{\sigma_2}{\sigma_1} Z_2) + \mu_1 + \mu_2$

Now remains to compute distribution of  $Y = Z_1 + \sigma Z_2$  with  $\sigma = \frac{\sigma_2}{\sigma_1}$

$$f_Y(y) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}} dx$$

$$x^2 + \frac{(y-x)^2}{\sigma^2} = \frac{1}{\sigma^2} ((\sigma^2 + 1)x^2 - 2xy + y^2) = \frac{(\sigma^2 + 1)}{\sigma^2} \left( x - \frac{y}{(\sigma^2 + 1)} \right)^2 + \frac{y^2}{\sigma^2 + 1}$$

$$f_Y(y) = \frac{e^{-\frac{y^2}{2(\sigma^2 + 1)}}}{2\pi\sigma} \int_{-\infty}^{+\infty} e^{-\frac{(\sigma^2 + 1)}{2\sigma^2} \left( x - \frac{y}{(\sigma^2 + 1)} \right)^2} dx = \frac{e^{-\frac{y^2}{(\sigma^2 + 1)}}}{\sqrt{2\pi(\sigma^2 + 1)}}$$

So  $Y \sim \mathcal{N}(0, \sigma^2 + 1)$ , then  $Z_1 + \frac{\sigma_2}{\sigma_1} Z_2 \sim \mathcal{N}\left(0, 1 + \left(\frac{\sigma_2}{\sigma_1}\right)^2\right)$  and

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

## Quiz for Next Lecture

### Exercise

Let  $X \sim \text{Exp}(\lambda)$  and  $Y \sim \text{Exp}(\lambda)$  for  $\lambda > 0$ .

1. Find the distribution of  $X + Y$

# Review of the course

## STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 11, April 22th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

# Overview

## Course feedback

- ▶ available at <https://uw.iassystem.org/survey/222954>
- ▶ until **Monday, 27th April**

**Lecture Notes** available for Lectures 7 to 10

## 1st Exam

- ▶ Available on **Friday 1:00pm**, due on **Saturday 1:00pm**
- ▶ Scan/take a picture of your solutions, send them through Canvas

## Review plan

1. Summary of what you need to know/be able to do
2. List of exercises
3. Ask questions for each topic, feel free to interrupt

## What you need to know

### Multivariate random variables

1. joint p.m.f., joint p.d.f.
2. marginal p.m.f., marginal p.d.f. be able to compute them
3. multinomial know definition, ready to apply
4. uniform continuous in 2d, 3d know definition, ready to apply

### Independence and joint p.m.f., joint p.d.f.

1. Characterization in terms of joint p.m.f., joint p.d.f.
2. Be able to show dependence with counterexample
3. Be able to prove independence by computing marginals

### Functions of Random Variables

1. change of distribution for  $Y = g(X)$  be able to compute it for any  $g$
2. change of distribution for  $(U, V) = g(X, Y)$  know the formula, be able to use it for simple cases
3. min, max of independent random variables know links of c.d.f., be able to manipulate them
4. sums of independent random variables know formula, be able to use it and simplify computations

## Multivariate Random Variables, Discrete Case

### Exercise

$X_1, X_2$  two independent dice of 4 faces ( $X_1, X_2 \in \{1, \dots, 4\}$ ) after a roll

Define  $S = X_1 - X_2$ ,  $Y = \min(X_1, X_2)$

1. Compute joint p.m.f. ( $S, Y$ ) as a table
2. Compute marginal p.m.f. of  $S$ , marginal p.d.f. of  $Y$
3. Are  $S, Y$  independent?

## Multivariate Random Variables, Continuous case

### Exercise

Let  $X, Y$  with joint p.d.f.

$$f_{X,Y}(x,y) = c(xy + y^2) \quad \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1$$

for a given  $c \geq 0$

1. Find the value of  $c$
2. Find the marginal p.d.f. of  $x$  and  $y$
3. Are  $X$  and  $Y$  independent?
4. Compute  $\mathbb{P}(X < Y)$

## Multinomial

### Exercise

A pizza place shop offers 8 different pizzas. Alex likes them equally. Each evening he picks one randomly independently from previous choices. Regina, Calzone, 4 cheeses are three of the different pizzas. During a week of 5 days. let  $X, Y, Z$ , the number of times he chooses Regina, Calzone and 4 cheeses respectively and denote by  $W$  the number of times he chooses something else.

1. What is the joint p.m.f. of  $(X, Y, Z, W)$ ?

# Functions of Random Variables

## Exercise

1. Let  $X \sim \text{Unif}([-\pi, 2\pi])$ , find the p.d.f. of  $Y = \sin(X)$
2. Let  $X \sim \mathcal{N}(0, 1)$  and  $Y = e^X$ . Find the p.d.f of  $Y$ .
3. Let  $X, Y$  be two independent standard Normal random variables. Let  $U = X + 2Y$ ,  $V = X - 2Y$ .
  - 3.1 Find the joint distribution of  $(U, V)$
  - 3.2 Are  $U$  and  $V$  independent?

## Minimums Maximums of Independent Random Variables

### Exercise

Let  $X_1, X_2, \dots, X_n$  be independent with  $X_i \sim \text{Exp}(\lambda_i)$  for  $i \in \{1, \dots, n\}$ ,  $\lambda_i > 0$

1. What is the p.d.f. of  $\min(Y_1, \dots, Y_n)$ ?

## Sums of random variables

### Exercise

*Let  $X, Y$  be independent with  $X \sim \text{Poisson}(\lambda)$  and  $Y \sim \text{Ber}(p)$ .*

*Find the p.m.f. of  $X + Y$ .*

# Exchangeable Random Variables

Section 7.2

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 13, April 27th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Course feedback

- ▶ available at <https://uw.iassystem.org/survey/222954>
- ▶ until tonight

### Homeworks, Exam

- ▶ Homework 1 graded, solutions of exam coming soon

### Previous lectures

- ▶ Functions of random variables
- ▶ Independence of random variables, charac. in terms of p.m.f./p.d.f.

### This lecture

- ▶ New notion of symmetry: exchangeability
- ▶ Consequences: marginals identically distributed
- ▶ Characterization in terms of p.m.f./p.d.f.
- ▶ Applications for
  - ▶ Sampling without replacement
  - ▶ Independent identically distributed random variables

## Exchangeable Random Variables

### Motivating Example

Flip over cards from shuffled deck one by one.

What is the probability that the 23rd card is a spade?

## Exchangeable Random Variables

### Motivating Example

Flip over cards from shuffled deck one by one.

What is the probability that the 23rd card is a spade?

### Intuition

*Without any additional information,*  $\mathbb{P}(\text{"23rd card is a spade"})$  should be equal to  $\mathbb{P}(\text{"1st card is a spade"})$

# Exchangeable Random Variables

## Motivating Example

Flip over cards from shuffled deck one by one.

What is the probability that the 23rd card is a spade?

## Intuition

*Without any additional information*,  $\mathbb{P}(\text{"23rd card is a spade"})$  should be equal to  $\mathbb{P}(\text{"1st card is a spade"})$

## How to formalize that?

1. Define the r.v. associated to the first 23rd cards  $X_1, \dots, X_{23}$  with  $X_i \in \{\text{heart, diamond, spade, club}\}$
  2. Write down the joint p.m.f. of  $X_1, \dots, X_{23}$
  3. Compute marginal p.m.f. of  $X_{23}$  and of  $X_1$ , should be the same
- *The joint p.m.f. must satisfy some property... and that's not independence...*

## Equalities in Distribution

### Definition (Equality in distribution)

Two random vectors  $(X_1, \dots, X_n)$ ,  $(Y_1, \dots, Y_n)$  are **equal in distribution** if

$$\mathbb{P}((X_1, \dots, X_n) \in B) = \mathbb{P}((Y_1, \dots, Y_n) \in B) \quad \text{for any } B \subset \mathbb{R}^n$$

we denote it

$$(X_1, \dots, X_n) \stackrel{d}{=} (Y_1, \dots, Y_n)$$

---

<sup>1</sup>In the continuous case, one random variable may have multiple p.d.f. (see previous lectures). Here if a marginal p.d.f. can be used to compute probabilities associated to  $X_k$  then the same p.d.f. can be used to compute probabilities associated to  $X_j$

## Equalities in Distribution

### Definition (Equality in distribution)

Two random vectors  $(X_1, \dots, X_n)$ ,  $(Y_1, \dots, Y_n)$  are **equal in distribution** if

$$\mathbb{P}((X_1, \dots, X_n) \in B) = \mathbb{P}((Y_1, \dots, Y_n) \in B) \quad \text{for any } B \subset \mathbb{R}^n$$

we denote it

$$(X_1, \dots, X_n) \stackrel{d}{=} (Y_1, \dots, Y_n)$$

### Definition (Identically distributed)

$X_1, \dots, X_n$  are **identically distributed** if for any  $k, j \in \{1, \dots, n\}$ ,

$$X_k \stackrel{d}{=} X_j$$

i.e. they have same **marginal p.m.f.** (Discrete case) or **p.d.f.** (Continuous case)<sup>1</sup>

---

<sup>1</sup>In the continuous case, one random variable may have multiple p.d.f. (see previous lectures). Here if a marginal p.d.f. can be used to compute probabilities associated to  $X_k$  then the same p.d.f. can be used to compute probabilities associated to  $X_j$

## Exchangeable Random Variables

### Definition (Exchangeability)

$X_1, \dots, X_n$  are **exchangeable** if for any permutation  $k_1, \dots, k_n$  of  $\{1, \dots, n\}$ ,

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{k_1}, \dots, X_{k_n})$$

## Exchangeable Random Variables

### Definition (Exchangeability)

$X_1, \dots, X_n$  are **exchangeable** if for any permutation  $k_1, \dots, k_n$  of  $\{1, \dots, n\}$ ,

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{k_1}, \dots, X_{k_n})$$

### Lemma (Consequences of exchangeability 1)

*Let  $(X_1, \dots, X_n)$  be exchangeable, then they are identically distributed*

## Exchangeable Random Variables

### Definition (Exchangeability)

$X_1, \dots, X_n$  are **exchangeable** if for any permutation  $k_1, \dots, k_n$  of  $\{1, \dots, n\}$ ,

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{k_1}, \dots, X_{k_n})$$

### Lemma (Consequences of exchangeability 1)

Let  $(X_1, \dots, X_n)$  be exchangeable, then they are identically distributed

**Proof** Let  $B_1 \subset \mathbb{R}$  and  $B_2 = \dots = B_n = \mathbb{R}$ ,

$$\mathbb{P}(X_1 \in B_1) = \mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_j \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_j \in B_1)$$

So  $X_1, X_j$  are identically distributed (same for any  $k, j$  in  $\{1, \dots, n\}$ )

## Exchangeable Random Variables

### Definition (Exchangeability)

$X_1, \dots, X_n$  are **exchangeable** if for any permutation  $k_1, \dots, k_n$  of  $\{1, \dots, n\}$ ,

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{k_1}, \dots, X_{k_n})$$

### Lemma (Consequences of exchangeability 1)

*Let  $(X_1, \dots, X_n)$  be exchangeable, then they are identically distributed*

### Example (Flipping 23 cards)

In our motivating example, if we show exchangeability, then

$$\mathbb{P}(X_{23} \text{ is a spade}) = \mathbb{P}(X_1 \text{ is a spade}) = 1/4$$

## Exchangeable Random Variables

### Definition (Exchangeability)

$X_1, \dots, X_n$  are **exchangeable** if for any permutation  $k_1, \dots, k_n$  of  $\{1, \dots, n\}$ ,

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{k_1}, \dots, X_{k_n})$$

### Lemma (Consequences of exchangeability 1)

Let  $(X_1, \dots, X_n)$  be exchangeable, then they are identically distributed

### Lemma (Consequences of exchangeability 2)

Let  $(X_1, \dots, X_n)$  be exchangeable, then for any  $k \in \{1, \dots, n\}$  and any permutation  $(i_1, \dots, i_k)$  of  $\{1, \dots, k\}$

$$(X_1, \dots, X_k) \stackrel{d}{=} (X_{i_1}, \dots, X_{i_k})$$

and for any  $g : \mathbb{R}^k \rightarrow \mathbb{R}$ ,  $\mathbb{E}[g(X_1, \dots, X_k)] = \mathbb{E}[g(X_{i_1}, \dots, X_{i_k})]$

## Exchangeable random variables

### Lemma (How to check for exchangeability)

Let  $(X_1, \dots, X_n)$  be random variables,

1. If  $X_1, \dots, X_n$  are discrete, they are exchangeable if and only if their joint p.m.f.  $p$  is symmetric, i.e.

$$\begin{aligned}\mathbb{P}(X_1 = k_1, \dots, X_n = k_n) &= p(k_1, \dots, k_n) \\ &= p(k_{i_1}, \dots, k_{i_n}) = \mathbb{P}(X_1 = k_{i_1}, \dots, X_n = k_{i_n})\end{aligned}$$

for  $k_1, \dots, k_n \in \mathbb{Z}$  and  $i_1, \dots, i_n$  a permutation of  $\{1, \dots, n\}$

2. If  $X_1, \dots, X_n$  are jointly continuous, they are exchangeable if and only if their joint p.d.f.  $f$  is symmetric, i.e.

$$f(x_1, \dots, x_n) = f(x_{i_1}, \dots, x_{i_n})$$

for  $x_1, \dots, x_n \in \mathbb{R}$  and  $i_1, \dots, i_n$  a permutation of  $\{1, \dots, n\}$

## Exchangeable random variables

### Example

Let  $X_1, X_2, X_3$  be jointly continuous with joint p.d.f.  $f$ , are there exchangeable if

1.  $f(x_1, x_2, x_3) = x_1 x_2 x_3 \mathbf{1}_{[0,1]^3}(x_1, x_2, x_3)$ ?
2.  $f(x_1, x_2, x_3) = (x_1 x_2 + x_3) \mathbf{1}_{[0,1]^3}(x_1, x_2, x_3)$ ?

## Sampling Without Replacement

### Theorem

*Let  $X_1, \dots, X_m$  denote the outcomes of successive draws uniformly at random without replacement from  $\{1, \dots, n\}$  ( $n$  distinct objects numbered from 1 to  $n$ ) with  $m \leq n$ .*

*Then  $X_1, \dots, X_m$  are exchangeable.*

## Sampling Without Replacement

### Theorem

Let  $X_1, \dots, X_m$  denote the outcomes of successive draws uniformly at random without replacement from  $\{1, \dots, n\}$  ( $n$  distinct objects numbered from 1 to  $n$ ) with  $m \leq n$ .

Then  $X_1, \dots, X_m$  are exchangeable.

**Proof** Let  $k_1, \dots, k_m$  be  $m$  elements of  $\{1, \dots, n\}$ . Then

$$\mathbb{P}(X_1 = k_1, \dots, X_n = k_n) = \frac{1}{n} \times \frac{1}{n-1} \times \dots \times \frac{1}{n-m+1}$$

which shows that the joint p.m.f. only depends on the number of draws.  
Formally, for  $i_1, \dots, i_m$  a permutation of  $\{1, \dots, m\}$ ,

$$\mathbb{P}(X_1 = k_{i_1}, \dots, X_m = k_{i_m}) = \frac{1}{n(n-1)\dots(n-m+1)} = \mathbb{P}(X_1 = k_1, \dots, X_n = k_n)$$

which shows exchangeability.

## Exchangeable Random Variables

### **Indistinct outcomes**

Previous theorem assumes that the outcomes are distinct

What about indistinct outcomes? (Like "spade", "heart",... when flipping cards)

# Exchangeable Random Variables

## Indistinct outcomes

Previous theorem assumes that the outcomes are distinct

What about indistinct outcomes? (Like "spade", "heart",... when flipping cards)

## Idea

1. Consider that you numbered the cards.
2. Assume that the index of the 54 cards are ordered such that you can define

$$g(y) = \begin{cases} \text{spade} & \text{if } y \in \{1, \dots, 13\} \\ \text{heart} & \text{if } y \in \{14, \dots, 26\} \\ \text{diamond} & \text{if } y \in \{27, \dots, 39\} \\ \text{club} & \text{if } y \in \{40, \dots, 52\} \end{cases}$$

3. Denote  $Y_1, \dots, Y_{23}$  the random index of the 23 first cards you draw.
4. Then  $X_1 = g(Y_1), \dots, X_{23} = g(Y_{23})$  are the random variables we defined,  $X_i \in \{\text{spade, heart, diamond, club}\}$ ,  $\{X_i = \text{spade}\} \Leftrightarrow \text{"the } i^{\text{th}} \text{ card is a spade"}$
5.  $Y_1, \dots, Y_{23}$  are distinct and drawn without replacement so exchangeable
6. What about  $g(Y_1), \dots, g(Y_{23})$ ?

## Function of exchangeable random variable

### Theorem

If  $Y_1, \dots, Y_n$  are exchangeable, then for any function  $g$ ,  $g(Y_1), \dots, g(Y_n)$  are exchangeable.

### Example (Flipping 23 cards)

From our previous reasoning, we get

$$\mathbb{P}(X_{23} \text{ is a spade}) = \mathbb{P}(X_1 \text{ is a spade}) = 1/4$$

## Function of exchangeable random variable

### Theorem

If  $Y_1, \dots, Y_n$  are exchangeable, then for any function  $g$ ,  $g(Y_1), \dots, g(Y_n)$  are exchangeable.

### Example (Flipping 23 cards)

From our previous reasoning, we get

$$\mathbb{P}(X_{23} \text{ is a spade}) = \mathbb{P}(X_1 \text{ is a spade}) = 1/4$$

### Example

An urn contains 5 red balls, 3 green balls. Draw 8 balls without replacement.  
What is the probability that the 3rd ball is red and the seventh a green one?

## Function of exchangeable random variable

### Theorem

If  $Y_1, \dots, Y_n$  are exchangeable, then for any function  $g$ ,  $g(Y_1), \dots, g(Y_n)$  are exchangeable.

### Example (Flipping 23 cards)

From our previous reasoning, we get

$$\mathbb{P}(X_{23} \text{ is a spade}) = \mathbb{P}(X_1 \text{ is a spade}) = 1/4$$

### Example

An urn contains 5 red balls, 3 green balls. Draw 8 balls without replacement. What is the probability that the 3rd ball is red and the seventh a green one?

**Solution** Denote  $X_1, \dots, X_8$  the colors of the balls you draw. This can be treated with the same reasoning as before (numbering the balls and write the color of the ball you draw as a function of the index of the balls) such that  $X_1, \dots, X_8$  are exchangeable, so

$$\mathbb{P}(X_3 = \text{red}, X_7 = \text{green}) = \mathbb{P}(X_1 = \text{red}, X_2 = \text{green}) = \frac{5}{8} \times \frac{3}{7} \approx 0.27$$

## Independent, Identically Distributed Random Variables

### Lemma

*n independent identically distributed (i.i.d.) r.v.  $X_1, \dots, X_n$  are exchangeable.*

## Independent, Identically Distributed Random Variables

### Lemma

*n independent identically distributed (i.i.d.) r.v.  $X_1, \dots, X_n$  are exchangeable.*

**Proof** (Discrete case) Denote  $p = p_{X_j}$  for  $j \in \{1, \dots, n\}$  (same for all  $j$ )

For any  $k_1, \dots, k_n \in \mathbb{Z}$  and any permutation  $i_1, \dots, i_n$  of  $\{1, \dots, n\}$

$$\mathbb{P}(X_1 = k_1, \dots, X_n = k_n) = p_{X_1}(k_1) \dots p_{X_n}(k_n) = p(k_1) \dots p(k_n)$$

$$\mathbb{P}(X_1 = k_{i_1}, \dots, X_n = k_{i_n}) = p_{X_1}(k_{i_1}) \dots p_{X_n}(k_{i_n}) = p(k_{i_1}) \dots p(k_{i_n}) = p(k_1) \dots p(k_n)$$

So the joint p.m.f. is symmetric therefore the random variables are exchangeable.

Continuous case can be done similarly

## Independent, Identically Distributed Random Variables

### Example (Simplification by exchangeability)

Suppose that  $X_1, X_2, X_3$  are i.i.d with  $X_i \sim \text{Unif}([0, 1])$  for  $i \in \{1, 2, 3\}$

What is the probability that  $X_1$  is the largest?

## Independent, Identically Distributed Random Variables

### Example (Simplification by exchangeability)

Suppose that  $X_1, X_2, X_3$  are i.i.d with  $X_i \sim \text{Unif}([0, 1])$  for  $i \in \{1, 2, 3\}$

What is the probability that  $X_1$  is the largest?

**Solution** Since they are exchangeable,

$$\mathbb{P}(X_1 \text{ is largest}) = \mathbb{P}(X_2 \text{ is largest}) = \mathbb{P}(X_3 \text{ is largest})$$

Moreover

$$1 = \mathbb{P}(X_1 \text{ is largest}) + \mathbb{P}(X_2 \text{ is largest}) + \mathbb{P}(X_3 \text{ is largest})$$

since the probability that they are equal is zero (they are jointly continuous).

So  $\mathbb{P}(X_1 \text{ is largest}) = 1/3$

## Quiz for Next Lecture

### Exercise

*Deal 10 cards from a standard deck (52 cards).*

*What is the probability that the 6th card is a queen, given that the 5th and the 10th ones are both queens?*

# Expectation, Variance of Independent Random Variables

Section 8.2

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 14, April 30th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Course feedback

**Thanks for the constructive comments!**

### Proposed improvements

1. Talk slower, take more time on intuitions, but stay on time

2. Basic examples at the end of each lecture with solutions

**Reminder:** slides are updated after the lecture with solutions, check the syllabus

3. Additional office hours (fill the poll)

4. Whiteboard issues:

- ▶ will write more clearly
- ▶ whiteboard is meant for you to try to solve the proof/exercise by yourself on paper, i.e., to be active
- ▶ we keep it, give me feedback if things do not improve, I'll find another way
- ▶ switching screens issues → searching for other software, ideas are welcome!

### Checking improvements

→ Feedback needed every two weeks! (through PollEverywhere)

### Answer to comments

1. More explanations of the grading

→ please come to the office hour to discuss it :)

2. "Provide typed notes to coincide with the lectures rather than handwritten ones."

→ Lecture notes are available, please send me a message to detail

# Overview

## Previous Lectures

1. Random variables, joint p.m.f., p.d.f.
  2. Independence, exchangeability, charac. in terms of p.m.f./p.d.f.
  3. Functions of random variables
- Schedule an office hour with me if things are unclear!  
especially if you are giving an actuarial exam in probability in June :)

## This Lecture

1. Sample mean
2. Variance of a sum of independent r.v.
3. Unbiased estimators
4. Coupon collector problem

## Quiz Previous Lecture

### Exercise

*Deal 10 cards from a standard deck (52 cards).*

*What is the probability that the 6th card is a queen, given that the 5th and the 10th ones are both queens?*

## Quiz Previous Lecture

### Exercise

Deal 10 cards from a standard deck (52 cards).

What is the probability that the 6th card is a queen, given that the 5th and the 10th ones are both queens?

**Solution** Let  $X_j$  be the value of the  $j^{\text{th}}$  card.

$$\begin{aligned}\mathbb{P}(X_6 = \text{queen} | X_5 = \text{queen}, X_{10} = \text{queen}) &= \frac{\mathbb{P}(X_6 = \text{queen}, X_5 = \text{queen}, X_{10} = \text{queen})}{\mathbb{P}(X_5 = \text{queen}, X_{10} = \text{queen})} \\ &= \frac{\mathbb{P}(X_1 = \text{queen}, X_2 = \text{queen}, X_3 = \text{queen})}{\mathbb{P}(X_1 = \text{queen}, X_2 = \text{queen})} \\ &= \mathbb{P}(X_3 = \text{queen} | X_1 = \text{queen}, X_2 = \text{queen}) \\ &= \frac{2}{50} \approx 0.04\end{aligned}$$

## Motivation

### **How to estimate mean and variance from a random variable?**

- ▶ You have access to a random variable  $X$  through its realizations

## Motivation

### **How to estimate mean and variance from a random variable?**

- ▶ You have access to a random variable  $X$  through its realizations
- ▶ You make  $n$  independent trials from this random variable

### How to estimate mean and variance from a random variable?

- ▶ You have access to a random variable  $X$  through its realizations
- ▶ You make  $n$  independent trials from this random variable
- ▶ These trials can be seen as  $n$  i.i.d. r.v. following the distribution of  $X$

## Motivation

### How to estimate mean and variance from a random variable?

- ▶ You have access to a random variable  $X$  through its realizations
- ▶ You make  $n$  independent trials from this random variable
- ▶ These trials can be seen as  $n$  i.i.d. r.v. following the distribution of  $X$
- ▶ Denoting these trials  $X_1, \dots, X_n$ , define the **sample mean/empirical mean**

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

### How to estimate mean and variance from a random variable?

- ▶ You have access to a random variable  $X$  through its realizations
- ▶ You make  $n$  independent trials from this random variable
- ▶ These trials can be seen as  $n$  i.i.d. r.v. following the distribution of  $X$
- ▶ Denoting these trials  $X_1, \dots, X_n$ , define the **sample mean/empirical mean**

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

- ▶ What is the expectation of  $\bar{X}_n$ ? (easy)
- ▶ What is the variance of  $\bar{X}_n$ ? (needs more tools!)

## Product of Independent Random Variables

Lemma (Expectation of product of independent random variables)

Let  $X_1, \dots, X_n$  be independent r.v.

Let  $g_1, \dots, g_n$  be  $n$  functions  $g_i : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[g_i(X_i)]$  is defined.

$$\mathbb{E}[g_1(X_1) \dots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \dots \mathbb{E}[g_n(X_n)]$$

## Product of Independent Random Variables

**Lemma (Expectation of product of independent random variables)**

Let  $X_1, \dots, X_n$  be independent r.v.

Let  $g_1, \dots, g_n$  be  $n$  functions  $g_i : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[g_i(X_i)]$  is defined.

$$\mathbb{E}[g_1(X_1) \dots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \dots \mathbb{E}[g_n(X_n)]$$

**Proof** (2 continuous r.v. case) Let  $X, Y$  be independent and continuous,  $g, h : \mathbb{R} \rightarrow \mathbb{R}$

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x)h(y)f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{+\infty} g(x)f_X(x)dx \int_{-\infty}^{+\infty} h(y)f_Y(y)dy = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]\end{aligned}$$

## Variance of independent Random Variables

### Lemma

Let  $X_1, \dots, X_n$  be  $n$  **independent** random variables with finite variance

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n)$$

# Variance of independent Random Variables

## Lemma

Let  $X_1, \dots, X_n$  be  $n$  **independent** random variables with finite variance

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n)$$

**Proof** Denote  $\mu_i = \mathbb{E}[X_i]$ , we know that  $\mathbb{E} \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n \mu_i$ .

$$\begin{aligned}\text{Var} \left( \sum_{i=1}^n X_i \right) &= \mathbb{E} \left[ \left( \sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i \right)^2 \right] = \mathbb{E} \left[ \left( \sum_{i=1}^n X_i - \mu_i \right)^2 \right] \\ &= \mathbb{E} \left[ \left( \sum_{i=1}^n X_i - \mu_i \right) \left( \sum_{j=1}^n X_j - \mu_j \right) \right] \\ &= \mathbb{E} \left[ \sum_{i=1}^n (X_i - \mu_i)^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^n (X_i - \mu_i)(X_j - \mu_j) \right] \\ &= \sum_{i=1}^n \mathbb{E}[(X_i - \mu_i)^2] + \underbrace{\sum_{\substack{i,j=1 \\ i \neq j}}^n \mathbb{E}[(X_i - \mu_i)] \mathbb{E}[(X_j - \mu_j)]}_{=0} = \sum_{i=1}^n \text{Var}(X_i)\end{aligned}$$

(Linearity of Expectation  
+ Expectation of product  
of independent r.v.)

## Variance of independent Random Variables

Example (Variance of a binomial random variable, easy computation)

Let  $X \sim \text{Bin}(n, p)$ , what is the variance of  $X$ ?

## Variance of independent Random Variables

Example (Variance of a binomial random variable, easy computation)

Let  $X \sim \text{Bin}(n, p)$ , what is the variance of  $X$ ?

**Solution** By definition,  $X = B_1 + \dots + B_n$  where  $B_i \sim \text{Ber}(p)$  are independent.

We have  $\text{Var}(B_i) = p(1 - p)$ , so  $\text{Var}(X) = \text{Var}(B_1) + \dots + \text{Var}(B_n) = np(1 - p)$

## Variance of independent Random Variables

Example (Variance of a binomial random variable, easy computation)

Let  $X \sim \text{Bin}(n, p)$ , what is the variance of  $X$ ?

**Solution** By definition,  $X = B_1 + \dots + B_n$  where  $B_i \sim \text{Ber}(p)$  are independent.

We have  $\text{Var}(B_i) = p(1 - p)$ , so  $\text{Var}(X) = \text{Var}(B_1) + \dots + \text{Var}(B_n) = np(1 - p)$

Example (Variance of negative binomial random variable)

Let  $X \sim \text{NegBin}(n, p)$ , i.e.  $X = G_1 + \dots + G_n$  with  $G_i \sim \text{Geom}(p)$  independent.

What is the expectation and variance of  $X$ ?

## Variance of independent Random Variables

**Example (Variance of a binomial random variable, easy computation)**

Let  $X \sim \text{Bin}(n, p)$ , what is the variance of  $X$ ?

**Solution** By definition,  $X = B_1 + \dots + B_n$  where  $B_i \sim \text{Ber}(p)$  are independent.

We have  $\text{Var}(B_i) = p(1 - p)$ , so  $\text{Var}(X) = \text{Var}(B_1) + \dots + \text{Var}(B_n) = np(1 - p)$

**Example (Variance of negative binomial random variable)**

Let  $X \sim \text{NegBin}(n, p)$ , i.e.  $X = G_1 + \dots + G_n$  with  $G_i \sim \text{Geom}(p)$  independent.

What is the expectation and variance of  $X$ ?

**Solution** We have  $\mathbb{E}(G_i) = \frac{1}{p}$  and  $\text{Var}(G_i) = \frac{1-p}{p^2}$

So  $\mathbb{E}(X) = \frac{n}{p}$ ,  $\text{Var}(X) = \frac{n(1-p)}{p^2}$ .

## Sample Mean/Empirical Mean

### Lemma

Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables drawn from the distribution of a random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ .

The **sample mean** or **empirical mean** of the first  $n$  observations is defined as

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

It satisfies  $\mathbb{E}(\bar{X}_n) = \mu$  and  $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ .

**Note:** The variance of the empirical mean tends to 0 as  $n \rightarrow +\infty$ .

Gives the intuition that, as  $n \rightarrow +\infty$ ,  $\bar{X}_n$  converges to the mean of  $X$ , i.e.  $\mu$   
(This will be shown properly with a proof of the law of large numbers)

## Unbiased Estimators

### Definition (Estimator)

Let  $\theta$  be a parameter of the distribution of a r.v.  $X$  (e.g.  $\theta = \mathbb{E}(X)$  or  $\theta = \text{Var}(X)$ ).

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of  $X$  seen as random variables  
(i.e.  $n$  independent r.v. all following the distribution of  $X$ )

## Unbiased Estimators

### Definition (Estimator)

Let  $\theta$  be a parameter of the distribution of a r.v.  $X$  (e.g.  $\theta = \mathbb{E}(X)$  or  $\theta = \text{Var}(X)$ ).

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of  $X$  seen as random variables

(i.e.  $n$  independent r.v. all following the distribution of  $X$ )

1. An **estimator**  $\hat{\theta}$  of  $\theta$  from  $n$  observations is a function of the  $n$  i.i.d. r.v.

## Unbiased Estimators

### Definition (Estimator)

Let  $\theta$  be a parameter of the distribution of a r.v.  $X$  (e.g.  $\theta = \mathbb{E}(X)$  or  $\theta = \text{Var}(X)$ ).

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of  $X$  seen as random variables  
(i.e.  $n$  independent r.v. all following the distribution of  $X$ )

1. An **estimator**  $\hat{\theta}$  of  $\theta$  from  $n$  observations is a function of the  $n$  i.i.d. r.v.
2. The bias of an estimator  $\hat{\theta}$  of  $\theta$  is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

## Unbiased Estimators

### Definition (Estimator)

Let  $\theta$  be a parameter of the distribution of a r.v.  $X$  (e.g.  $\theta = \mathbb{E}(X)$  or  $\theta = \text{Var}(X)$ ).

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of  $X$  seen as random variables  
(i.e.  $n$  independent r.v. all following the distribution of  $X$ )

1. An **estimator**  $\hat{\theta}$  of  $\theta$  from  $n$  observations is a function of the  $n$  i.i.d. r.v.
2. The bias of an estimator  $\hat{\theta}$  of  $\theta$  is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

3. An **unbiased estimator** is an estimator with zero bias

**Note:** An estimator is itself a r.v. as a function of r.v.

## Unbiased Estimators

### Definition (Estimator)

Let  $\theta$  be a parameter of the distribution of a r.v.  $X$  (e.g.  $\theta = \mathbb{E}(X)$  or  $\theta = \text{Var}(X)$ ).

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of  $X$  seen as random variables  
(i.e.  $n$  independent r.v. all following the distribution of  $X$ )

1. An **estimator**  $\hat{\theta}$  of  $\theta$  from  $n$  observations is a function of the  $n$  i.i.d. r.v.
2. The bias of an estimator  $\hat{\theta}$  of  $\theta$  is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

3. An **unbiased estimator** is an estimator with zero bias

**Note:** An estimator is itself a r.v. as a function of r.v.

### Example

The sample mean of the first  $n$  observations  $X_1, \dots, X_n$  of a r.v.  $X$  is an unbiased estimator of the mean of  $X$ .

Namely  $\mathbb{E}[\bar{X}_n] = \mu$  where  $\mu$  is the mean of the r.v.  $X$

## Unbiased Estimators

### Unbiased estimator of the variance

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of a r.v.  $X$ .

What would be an unbiased estimator of  $\sigma^2 = \text{Var}(X)$  from  $X_1, \dots, X_n$ ?

1. Would  $Y_n = \frac{1}{n} \sum_{i=1}^n (\bar{X}_n - X_i)^2$  work?

## Unbiased Estimators

### Unbiased estimator of the variance

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of a r.v.  $X$ .

What would be an unbiased estimator of  $\sigma^2 = \text{Var}(X)$  from  $X_1, \dots, X_n$ ?

1. Would  $Y_n = \frac{1}{n} \sum_{i=1}^n (\bar{X}_n - X_i)^2$  work?

→ No!

$$\begin{aligned}\mathbb{E}[Y_n] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\bar{X}_n - \mu + \mu - X_i)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\bar{X}_n - \mu)^2 + (X_i - \mu)^2 - 2(\bar{X}_n - \mu)(X_i - \mu)] \\ &= \frac{1}{n} \left( n \frac{\sigma^2}{n} \right) + \frac{1}{n} n \sigma^2 - 2 \mathbb{E} \left[ \sum_{i=1}^n (\bar{X}_n - \mu)(X_i - \mu) \right] \\ &= \sigma^2 \left( \frac{1}{n} + 1 \right) - 2 \mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2 \left( 1 - \frac{1}{n} \right) \neq \sigma^2\end{aligned}$$

## Unbiased Estimators

### Unbiased estimator of the variance

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of a r.v.  $X$ .

What would be an unbiased estimator of  $\sigma^2 = \text{Var}(X)$  from  $X_1, \dots, X_n$ ?

1. Would  $Y_n = \frac{1}{n} \sum_{i=1}^n (\bar{X}_n - X_i)^2$  work?

→ No!

$$\begin{aligned}\mathbb{E}[Y_n] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\bar{X}_n - \mu + \mu - X_i)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\bar{X}_n - \mu)^2 + (X_i - \mu)^2 - 2(\bar{X}_n - \mu)(X_i - \mu)] \\ &= \frac{1}{n} \left( n \frac{\sigma^2}{n} \right) + \frac{1}{n} n \sigma^2 - 2 \mathbb{E} \left[ \sum_{i=1}^n (\bar{X}_n - \mu)(X_i - \mu) \right] \\ &= \sigma^2 \left( \frac{1}{n} + 1 \right) - 2 \mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2 \left( 1 - \frac{1}{n} \right) \neq \sigma^2\end{aligned}$$

2. But  $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X}_n - X_i)^2$  is an unbiased estimator

## Unbiased Estimators

### Unbiased estimator of the variance

Let  $X_1, \dots, X_n$  be  $n$  i.i.d. observations of a r.v.  $X$ .

What would be an unbiased estimator of  $\sigma^2 = \text{Var}(X)$  from  $X_1, \dots, X_n$ ?

1. Would  $Y_n = \frac{1}{n} \sum_{i=1}^n (\bar{X}_n - X_i)^2$  work?

→ No!

$$\begin{aligned}\mathbb{E}[Y_n] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\bar{X}_n - \mu + \mu - X_i)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\bar{X}_n - \mu)^2 + (X_i - \mu)^2 - 2(\bar{X}_n - \mu)(X_i - \mu)] \\ &= \frac{1}{n} \left( n \frac{\sigma^2}{n} \right) + \frac{1}{n} n \sigma^2 - 2 \mathbb{E} \left[ \sum_{i=1}^n (\bar{X}_n - \mu)(X_i - \mu) \right] \\ &= \sigma^2 \left( \frac{1}{n} + 1 \right) - 2 \mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2 \left( 1 - \frac{1}{n} \right) \neq \sigma^2\end{aligned}$$

2. But  $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X}_n - X_i)^2$  is an unbiased estimator

**Proof**  $\hat{\sigma}_n^2 = \frac{n}{n-1} Y_n$  so  $\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2$

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Approach

1. Could write the p.m.f. to compute  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Approach

1. Could write the p.m.f. to compute  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$
2. Rather try to decompose  $T_n$  in a sum of simpler r.v.

**Note:** Same idea used to compute e.g. variance of binomial

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Solution

1. Denote  $T_k$  the number of boxes you need to buy to get  $k$  different toys among  $n$

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Solution

1. Denote  $T_k$  the number of boxes you need to buy to get  $k$  different toys among  $n$
2.  $T_1 = 1$  clearly, what about  $T_2$ ?

$T_2 - T_1$  is the nb of boxes (*think nb of trials*) bought before getting a different toy than the 1st one.

For each box the proba. of getting a different toy is  $\frac{n-1}{n}$ .

So formally  $T_2 - T_1 \sim \text{Geom}\left(\frac{n-1}{n}\right)$

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Solution

1. Denote  $T_k$  the number of boxes you need to buy to get  $k$  different toys among  $n$
2.  $T_1 = 1$  clearly, what about  $T_2$ ?

$T_2 - T_1$  is the nb of boxes (*think nb of trials*) bought before getting a different toy than the 1st one.

For each box the proba. of getting a different toy is  $\frac{n-1}{n}$ .

So formally  $T_2 - T_1 \sim \text{Geom}\left(\frac{n-1}{n}\right)$

3. Similarly  $W_k = T_{k+1} - T_k$  is the nb of boxes needed to be bought to get a different toy than first  $k$  ones

By same reasoning  $W_k \sim \text{Geom}\left(\frac{n-k}{n}\right)$

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Solution

1. Denote  $T_k$  the number of boxes you need to buy to get  $k$  different toys among  $n$
2.  $T_1 = 1$  clearly, what about  $T_2$ ?

$T_2 - T_1$  is the nb of boxes (*think nb of trials*) bought before getting a different toy than the 1st one.

For each box the proba. of getting a different toy is  $\frac{n-1}{n}$ .

So formally  $T_2 - T_1 \sim \text{Geom}\left(\frac{n-1}{n}\right)$

3. Similarly  $W_k = T_{k+1} - T_k$  is the nb of boxes needed to be bought to get a different toy than first  $k$  ones

By same reasoning  $W_k \sim \text{Geom}\left(\frac{n-k}{n}\right)$

4. Finally

$$T_n = T_1 + T_2 - T_1 + \dots + T_n - T_{n-1} = 1 + W_1 + \dots + W_{n-1}$$

So we can get  $\mathbb{E}[T_n]$  without computing the p.m.f. of  $T_n$ !

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Solution

5.  $T_n = 1 + W_1 + \dots + W_{n-1}$ ,  $W_k \sim \text{Geom}\left(\frac{n-k}{n}\right)$ , how can we compute  $\text{Var}(T_n)$ ?

## Coupon Collector Problem

### Example

Each box of a brand of cereals contains a toy. There are  $n$  different kinds of toys, each kind is equally probable to appear in a box and all boxes are independently made.

Let  $T_n$  be the number of boxes needed to be bought to collect all different toys.

What is  $\mathbb{E}[T_n]$  and  $\text{Var}(T_n)$ ?

### Solution

5.  $T_n = 1 + W_1 + \dots + W_{n-1}$ ,  $W_k \sim \text{Geom}\left(\frac{n-k}{n}\right)$ , how can we compute  $\text{Var}(T_n)$ ?  
→ Needs  $W_k$  independent!
6. Intuitively yes, why the waiting time for the  $k^{\text{th}}$  different toy should depend on the waiting time to get the first  $k - 1$  different toys?
7. Formally, for any  $k, j$ ,  $a_k, a_j > 0$  integers  $\mathbb{P}(W_k = a_k | W_j = a_j) = \mathbb{P}(W_k = a_k)$
8. Variance can then be computed as before

## Quiz for Next Lecture

### Exercise

*As you walk in a park, you pick at random 1 flower every 5min. There are 5 different species in the park.*

- 1. How long should you walk on average before you get a complete bunch with all possible flowers from the park?*
- 2. What would be the variance of the time of your walk?*

## Additional exercises

### Exercise

Let  $X \sim \text{Geom}(p)$   $Y \sim \text{Poisson}(\lambda)$  be independent with  $p = 1/3$ ,  $\lambda = 4$ . A rectangle is drawn with side lengths  $X$  and  $Y + 1$ . What is the expected value of the perimeter and of the area of the rectangle?

## Additional exercises

### Exercise

Let  $X \sim \text{Geom}(p)$   $Y \sim \text{Poisson}(\lambda)$  be independent with  $p = 1/3$ ,  $\lambda = 4$ . A rectangle is drawn with side lengths  $X$  and  $Y + 1$ . What is the expected value of the perimeter and of the area of the rectangle?

**Solution** The perimeter is  $P = X + Y + 1$  so by linearity of the expectation,

$$\mathbb{E}[P] = \frac{1}{p} + \lambda + 1 = 8$$

The area is  $A = X(Y + 1)$  so since  $X$  and  $Y + 1$  are independent,

$$\mathbb{E}(A) = \mathbb{E}(X)\mathbb{E}(Y + 1) = 21$$

## Additional exercises

### Exercise

*Our faucet is broken and a plumber has been called. The arrival of the plumber is uniformly distributed between 1pm and 7pm. Independently of when the plumber arrives, the time it takes to fix the broken faucet is exponentially distributed with mean 30 min.*

*What is the expectation and variance of the time at which the plumber has fixed the faucet?*

## Additional exercises

### Exercise

*Our faucet is broken and a plumber has been called. The arrival of the plumber is uniformly distributed between 1pm and 7pm. Independently of when the plumber arrives, the time it takes to fix the broken faucet is exponentially distributed with mean 30 min.*

*What is the expectation and variance of the time at which the plumber has fixed the faucet?*

**Solution** Let  $X \sim \text{Unif}([1, 7])$  the time at which the plumber arrives and  $Y \sim \mathbb{E}(\lambda)$  the time he needs to fix the faucet. By assumptions  $X$  and  $Y$  are independent and  $\mathbb{E}(Y) = 0.5$  (we take hours to be the unit here, btw note that we have  $\lambda = 0.5$  since  $\mathbb{E}(Y) = \lambda$ ).

The time at which the plumber has fixed the faucet is  $Z = X + Y$ .

$$\text{So } \mathbb{E}(Z) = \mathbb{E}(X) + \mathbb{E}(Y) = \frac{7+1}{2} + 0.5 = 4 : 30\text{pm}.$$

Since  $X$  and  $Y$  are independent,

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = \frac{(7-1)^2}{12} + \frac{1}{\lambda^2} = 3 + 4 = 7\text{hours}$$

## Additional exercises

### Exercise

Let  $X, Y$  be two independent r.v. such that  $\mathbb{E}(X) = 3$ ,  $\mathbb{E}(Y) = 5$ ,  $\text{Var}(X) = 2$ ,  $\text{Var}(Y) = 3$ . Compute

1.  $\mathbb{E}(3X - 2Y + 7)$
2.  $\text{Var}(3X - 2Y + 7)$
3.  $\text{Var}(XY)$

## Additional exercises

### Exercise

Let  $X, Y$  be two independent r.v. such that  $\mathbb{E}(X) = 3$ ,  $\mathbb{E}(Y) = 5$ ,  $\text{Var}(X) = 2$ ,  $\text{Var}(Y) = 3$ . Compute

1.  $\mathbb{E}(3X - 2Y + 7)$
2.  $\text{Var}(3X - 2Y + 7)$
3.  $\text{Var}(XY)$

**Solution** Results follow from

1.  $\mathbb{E}(3X - 2Y + 7) = 3\mathbb{E}(X) - 2\mathbb{E}(Y) + 7 = 6$
2.  $\text{Var}(3X - 2Y + 7) = 3^2 \text{Var}(X) + 2^2 \text{Var}(Y) = 30$
3.  $\text{Var}(XY) = \mathbb{E}[(XY)^2] - \mathbb{E}[XY]^2 = \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]\mathbb{E}[Y]^2 = 83$  using that  $\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2$  and same thing for  $Y$ .

# Covariance

Section 8.4

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 15, May 1st, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Course feedback implementations

- ▶ if you want more office hours, fill the google form sent on the announcement
- ▶ if you had problems with the lecture notes, send me a message

### Previous lecture

- ▶ Expectation of a product of independent random variables
- ▶ Computation of expectation and variances by decomposition into sums
- ▶ Application to sample mean

### This lecture

- ▶ Indicator random variables (useful for the aforementioned decompositions)
- ▶ Covariance

## Quiz previous lecture

### Exercise

*As you walk in a park, you pick at random 1 flower every 5min. There are 5 different species in the park.*

- 1. How long should you walk on average before you get a complete bunch with all possible flowers from the park?*
- 2. What would be the variance of the time of your walk?*

## Quiz previous lecture

### Exercise

As you walk in a park, you pick at random 1 flower every 5min. There are 5 different species in the park.

1. How long should you walk on average before you get a complete bunch with all possible flowers from the park?
2. What would be the variance of the time of your walk?

**Solution** This is an instance from the coupon collector's problem with  $n = 5$ . Let  $T_n$  be the number of flowers I need to pick to have a complete bunch of  $n$  different flowers. ⚠ Typo corrected on slides of previous lecture

$$T_n = 1 + W_1 + \dots + W_{n-1}$$

with  $W_k \sim \text{Geom}(p_k)$ ,  $p_k = \frac{n-k}{n}$  independents

We have  $\mathbb{E}[W_k] = \frac{1}{p_k} = \frac{n}{n-k}$ ,  $\text{Var}(W_k) = \frac{1-p_k}{p_k^2} = \frac{k/n}{(n-k)^2/n^2} = \frac{kn}{(n-k)^2}$  so

$$\mathbb{E}[T_n] = 1 + \sum_{k=1}^{n-1} \frac{n}{n-k} = n \cdot \frac{1}{n} + n \sum_{k=1}^{n-1} \frac{1}{n-k} = n \sum_{j=1}^n \frac{1}{j}$$

$$\text{Var}(T_n) = \sum_{k=1}^{n-1} \frac{kn}{(n-k)^2} = \sum_{j=1}^{n-1} \frac{n(n-j)}{j^2} = n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} - n \sum_{j=1}^{n-1} \frac{1}{j}$$

So in our case we get an average time of 57 min and a variance of 126 min. Denote  $Y = 5T_5$  the time of the walk,  $\mathbb{E}[Y] = 5\mathbb{E}[T_5] = 57$ ,  $\text{Var}(Y) = 25\text{Var}(T_5) = 629$

## Indicator Random Variables

### Definition

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , the **indicator random variable** of an event  $A \subset \Omega$  is defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Denote  $p = \mathbb{P}(A)$ , we have  $I_A \sim \text{Ber}(p)$  and  $\mathbb{E}[I_A] = \mathbb{P}(A)$

## Indicator Random Variables

### Definition

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , the **indicator random variable** of an event  $A \subset \Omega$  is defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Denote  $p = \mathbb{P}(A)$ , we have  $I_A \sim \text{Ber}(p)$  and  $\mathbb{E}[I_A] = \mathbb{P}(A)$

### Example

Every day you walk around your house, you see at least one rabbit with probability 0.1, at least one cat with probability 0.3 and at least one bird with probability 0.5.

What is the average number of different animals you will see tomorrow?

## Indicator Random Variables

### Definition

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , the **indicator random variable** of an event  $A \subset \Omega$  is defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Denote  $p = \mathbb{P}(A)$ , we have  $I_A \sim \text{Ber}(p)$  and  $\mathbb{E}[I_A] = \mathbb{P}(A)$

### Example

Every day you walk around your house, you see at least one rabbit with probability 0.1, at least one cat with probability 0.3 and at least one bird with probability 0.5.

What is the average number of different animals you will see tomorrow?

**Solution** Define  $A_1, A_2, A_3$  the events "I see at least one rabbit", "I see at least one cat", "I see at least one bird" respectively.

The number of different animals you see is given by  $X = I_{A_1} + I_{A_2} + I_{A_3}$ .

So  $\mathbb{E}[X] = \mathbb{E}[I_{A_1}] + \mathbb{E}[I_{A_2}] + \mathbb{E}[I_{A_3}] = 0.9$

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.
2. How could we measure their dependence?

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.
2. How could we measure their dependence?
3. Requires a tool that could be expressed in terms of expectation...  
(then we could estimate it by replacing the expectation by a sample mean)

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.
2. How could we measure their dependence?
3. Requires a tool that could be expressed in terms of expectation...  
(then we could estimate it by replacing the expectation by a sample mean)
4. *Proposition:*

Take a function  $h$  of  $X, Y$  and define some dependence measure as

$$\mathbb{E}[h(X, Y)]$$

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.
2. How could we measure their dependence?
3. Requires a tool that could be expressed in terms of expectation...  
(then we could estimate it by replacing the expectation by a sample mean)
4. *Proposition:*

Take a function  $h$  of  $X, Y$  and define some dependence measure as

$$\mathbb{E}[h(X, Y)]$$

5. Today we take

$$h(X, Y) = (X - \mu_X)(Y - \mu_Y)$$

with  $\mu_X = \mathbb{E}(X), \mu_Y = \mathbb{E}(Y)$  which defines the **covariance**

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.
2. How could we measure their dependence?
3. Requires a tool that could be expressed in terms of expectation...  
(then we could estimate it by replacing the expectation by a sample mean)
4. *Proposition:*

Take a function  $h$  of  $X, Y$  and define some dependence measure as

$$\mathbb{E}[h(X, Y)]$$

5. Today we take

$$h(X, Y) = (X - \mu_X)(Y - \mu_Y)$$

with  $\mu_X = \mathbb{E}(X), \mu_Y = \mathbb{E}(Y)$  which defines the **covariance**

6. Can this measure assess independence?

## Covariance

### Motivation

Let  $X, Y$  be two random variables

(e.g. take a person at random denote  $X$  their size and  $Y$  the size of their feet)

1. We saw estimators of the mean and the variance of each r.v.
2. How could we measure their dependence?
3. Requires a tool that could be expressed in terms of expectation...  
(then we could estimate it by replacing the expectation by a sample mean)
4. *Proposition:*

Take a function  $h$  of  $X, Y$  and define some dependence measure as

$$\mathbb{E}[h(X, Y)]$$

5. Today we take

$$h(X, Y) = (X - \mu_X)(Y - \mu_Y)$$

with  $\mu_X = \mathbb{E}(X), \mu_Y = \mathbb{E}(Y)$  which defines the **covariance**

6. Can this measure assess independence?

→ Intuitively, why a single choice of  $h$  would be sufficient to capture all possible dependencies between  $X$  and  $Y$  ? ...

Yet, it is still going to be informative ! :)

e.g. it can inform about linear dependence (next lecture)

## Covariance

### Definition (Covariance)

Let  $X, Y$  be two random variables defined on the same probability space with expectations  $\mu_X, \mu_Y$ . The **covariance** of  $X$  and  $Y$  is defined by

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectations on the right are defined

## Covariance

### Definition (Covariance)

Let  $X, Y$  be two random variables defined on the same probability space with expectations  $\mu_X, \mu_Y$ . The **covariance** of  $X$  and  $Y$  is defined by

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectations on the right are defined

### Interpretation

Covariance can be interpreted as

“a measure of how  $X$  and  $Y$  **jointly** deviate from their mean”

## Covariance

### Definition (Covariance)

Let  $X, Y$  be two random variables defined on the same probability space with expectations  $\mu_X, \mu_Y$ . The **covariance** of  $X$  and  $Y$  is defined by

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectations on the right are defined

### Interpretation

Covariance can be interpreted as

“a measure of how  $X$  and  $Y$  **jointly** deviate from their mean”

e.g. if on all possible values that  $X, Y$  can take,

$$(X - \mu_X)(Y - \mu_Y) > 0$$

is on average more probable, i.e. that  $X$  tends to be higher than its mean  
**when**  $Y$  is higher than its mean then  $\text{Cov}(X, Y) > 0$

## Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

## Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. Clearly  $X_4$  and  $X_6$  are not independent

## Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. Clearly  $X_4$  and  $X_6$  are not independent
2. How can we measure that the "higher is  $X_4$ , the lower should be  $X_6$ "?

## Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. Clearly  $X_4$  and  $X_6$  are not independent
  2. How can we measure that the "higher is  $X_4$ , the lower should be  $X_6$ "?
- Compute  $\text{Cov}(X_4, X_6)$ , we should get that  $\text{Cov}(X_4, X_6) < 0$ , i.e.,  
as  $X_4$  tends to be higher than its mean,  $X_6$  tends to be lower than its mean

## Covariance

### Lemma

The **covariance** of  $X$  and  $Y$  can be formulated as

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$$

## Covariance

### Lemma

The **covariance** of  $X$  and  $Y$  can be formulated as

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$$

### Proof

$$\begin{aligned}\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] &= \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mu_X\mu_Y\end{aligned}$$

## Covariance

### Lemma

The **covariance** of  $X$  and  $Y$  can be formulated as

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$$

### Remarks:

1. For  $X = Y$  we retrieve the definition of the variance of  $X$ .

# Covariance

## Lemma

The **covariance** of  $X$  and  $Y$  can be formulated as

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$$

## Remarks:

1. For  $X = Y$  we retrieve the definition of the variance of  $X$ .
2. Computation of covariance requires to have access to the joint p.m.f/p.d.f.  
If  $X, Y$  are jointly continuous,

$$\text{Cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) dx dy$$

If  $X, Y$  are discrete (and integer valued),

$$\text{Cov}(X, Y) = \sum_{k=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} (k - \mu_X)(j - \mu_Y) \mathbb{P}(X = k, Y = j)$$

# Covariance

## Terminology

We say that two r.v.  $X, Y$  are

1. positively correlated if  $\text{Cov}(X, Y) > 0$
2. negatively correlated if  $\text{Cov}(X, Y) < 0$
3. uncorrelated if  $\text{Cov}(X, Y) = 0$

# Covariance

## Terminology

We say that two r.v.  $X, Y$  are

1. positively correlated if  $\text{Cov}(X, Y) > 0$
2. negatively correlated if  $\text{Cov}(X, Y) < 0$
3. uncorrelated if  $\text{Cov}(X, Y) = 0$

## Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

Intuitively,  $X_4, X_6$  are negatively correlated

→ proof next lecture

## Covariance

### Example

Let  $(X, Y)$  be uniformly distributed on a triangle  $T$  defined by vertices  $(0, 0), (0, 1), (1, 0)$

1. Intuitively, are  $X, Y$  positively, negatively correlated or uncorrelated ?
2. Compute  $\text{Cov}(X, Y)$ .

## Covariance

### Example

Let  $(X, Y)$  be uniformly distributed on a triangle  $T$  defined by vertices  $(0, 0), (0, 1), (1, 0)$

1. Intuitively, are  $X, Y$  positively, negatively correlated or uncorrelated ?
2. Compute  $\text{Cov}(X, Y)$ .

### Solution

1. Intuitively when  $X$  gets larger than its mean,  $Y$  diminishes, so they should be negatively correlated.
2.  $f_{X,Y}(x, y) = 2$  if  $(x, y) \in T$  and 0 o.w. (do following computations by yourself)

$$\mathbb{E}[X] = \int_T \int xf_{X,Y}(x, y) dxdy = \int_0^1 \int_0^{1-y} 2x dxdy = \frac{1}{3}$$

By symmetry,  $\mathbb{E}[Y] = \frac{1}{3}$  and

$$\mathbb{E}[XY] = \int_T \int xyf_{X,Y}(x, y) dxdy = \int_0^1 \int_0^{1-y} 2xy dxdy = \frac{1}{12}$$

$$\text{So } \text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{12} - \frac{1}{3} \cdot \frac{1}{3} = -\frac{1}{36} < 0$$

## Covariance of Indicator Random Variables

### Covariance of indicator random variables

Let  $A, B$  be two events on a proba. space  $\Omega, \mathcal{F}, \mathbb{P}$ .

$$\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

$$\text{If } \mathbb{P}(B) > 0, \text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$$

## Covariance of Indicator Random Variables

### Covariance of indicator random variables

Let  $A, B$  be two events on a proba. space  $\Omega, \mathcal{F}, \mathbb{P}$ .

$$\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

If  $\mathbb{P}(B) > 0$ ,  $\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$

#### Proof

$$\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{E}[\mathbf{I}_A \mathbf{I}_B] - \mathbb{E}[\mathbf{I}_A] \mathbb{E}[\mathbf{I}_B]$$

$$(\mathbf{I}_A \mathbf{I}_B)(\omega) = \mathbf{I}_A(\omega) \mathbf{I}_B(\omega) = \begin{cases} 1 & \text{if } \omega \in A \text{ and } \omega \in B \\ 0 & \text{otherwise} \end{cases}. \quad \text{Thus } \mathbf{I}_A \mathbf{I}_B = \mathbf{I}_{A \cap B}$$

$$\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{E}[\mathbf{I}_{A \cap B}] - \mathbb{E}[\mathbf{I}_A] \mathbb{E}[\mathbf{I}_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$$

provided that  $\mathbb{P}(B) > 0$  (for the last equality).

## Covariance of Indicator Random Variables

### Covariance of indicator random variables

Let  $A, B$  be two events on a proba. space  $\Omega, \mathcal{F}, \mathbb{P}$ .

$$\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

If  $\mathbb{P}(B) > 0$ ,  $\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$

### Interpretation of covariance for indicator random variables

1.  $\mathbf{I}_A, \mathbf{I}_B$  are *positively correlated* ( $\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) > 0$ )  $\Leftrightarrow \mathbb{P}(A|B) - \mathbb{P}(A) > 0$   
 $\rightarrow$  the occurrence of  $B$  *increases* the chances of  $A$ .

## Covariance of Indicator Random Variables

### Covariance of indicator random variables

Let  $A, B$  be two events on a proba. space  $\Omega, \mathcal{F}, \mathbb{P}$ .

$$\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

If  $\mathbb{P}(B) > 0$ ,  $\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$

### Interpretation of covariance for indicator random variables

1.  $\mathbf{I}_A, \mathbf{I}_B$  are *negatively correlated* ( $\text{Cov}(\mathbf{I}_A, \mathbf{I}_B) < 0$ )  $\Leftrightarrow \mathbb{P}(A|B) - \mathbb{P}(A) < 0$   
 $\rightarrow$  the occurrence of  $B$  *decreases* the chances of  $A$ .

## Covariance of Indicator Random Variables

### Covariance of indicator random variables

Let  $A, B$  be two events on a proba. space  $\Omega, \mathcal{F}, \mathbb{P}$ .

$$\text{Cov}(I_A, I_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

If  $\mathbb{P}(B) > 0$ ,  $\text{Cov}(I_A, I_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$

### Interpretation of covariance for indicator random variables

1.  $I_A, I_B$  are *negatively correlated* ( $\text{Cov}(I_A, I_B) < 0$ )  $\Leftrightarrow \mathbb{P}(A|B) - \mathbb{P}(A) < 0$   
 $\rightarrow$  the occurrence of  $B$  *decreases* the chances of  $A$ .
2.  $I_A, I_B$  are *uncorrelated* ( $\text{Cov}(I_A, I_B) = 0$ )  $\Leftrightarrow A, B$  are independent.

## Covariance of Indicator Random Variables

### Covariance of indicator random variables

Let  $A, B$  be two events on a proba. space  $\Omega, \mathcal{F}, \mathbb{P}$ .

$$\text{Cov}(I_A, I_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)$$

If  $\mathbb{P}(B) > 0$ ,  $\text{Cov}(I_A, I_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$

### Interpretation of covariance for indicator random variables

- $I_A, I_B$  are *negatively correlated* ( $\text{Cov}(I_A, I_B) < 0$ )  $\Leftrightarrow \mathbb{P}(A|B) - \mathbb{P}(A) < 0$   
 $\rightarrow$  the occurrence of  $B$  *decreases* the chances of  $A$ .
- $I_A, I_B$  are *uncorrelated* ( $\text{Cov}(I_A, I_B) = 0$ )  $\Leftrightarrow A, B$  are independent.

The covariance of **indicator random variables** is a measure of the independence of the corresponding events.

## Covariance of Indicator Random Variables

### Example

Let  $S$  be the sum of two fair dice  $X_1$  and  $X_2$ .

Are  $I_{\{S>10\}}, I_{\{X_2=6\}}$  positively, negatively correlated or uncorrelated?

## Covariance of Indicator Random Variables

### Example

Let  $S$  be the sum of two fair dice  $X_1$  and  $X_2$ .

Are  $I_{\{S>10\}}, I_{\{X_2=6\}}$  positively, negatively correlated or uncorrelated?

### Solution

$$\text{Cov}(I_{\{S>10\}}, I_{\{X_2=6\}}) = \mathbb{P}(S > 10, X_2 = 6) - \mathbb{P}(S > 10) \mathbb{P}(X_2 = 6) = \frac{2}{36} - \frac{3}{36} \frac{1}{6} > 0$$

So  $I_{\{S>10\}}, I_{\{X_2=6\}}$  are positively correlated.

## Covariance and Independence

### Theorem

Let  $X, Y$  be two random variables,

$$X, Y \text{ are independent} \Rightarrow \text{Cov}(X, Y) = 0$$

but the converse **does not hold in general**

# Covariance and Independence

## Theorem

Let  $X, Y$  be two random variables,

$$X, Y \text{ are independent} \Rightarrow \text{Cov}(X, Y) = 0$$

but the converse **does not hold in general**

## Proof

1. Let  $X, Y$  be two independent r.v., then from previous lecture

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

2. Counter example: take  $X \sim \text{Unif}(\{-1, 0, 1\})$  and  $Y = X^2$ , then

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2]$$

We have  $X^3 = X$  and  $\mathbb{E}[X] = 0$  therefore  $\text{Cov}(X, Y) = 0$

Yet,

$$\mathbb{P}(X = 1, Y = 0) = 0 \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 0) = \frac{1}{3}\frac{1}{3}$$

that is  $X, Y$  are not independent.

## Covariance and Independence

### Theorem

Let  $X, Y$  be two random variables,

$$X, Y \text{ are independent} \Rightarrow \text{Cov}(X, Y) = 0$$

but the converse **does not hold in general**

### Intuition:

1. Joint deviation from the means does not capture all possible interactions
2. For indicator r.v. it is sufficient because they describe only one event.
- General random variables describe much more than one event, we need to have more information than this simple covariance
3. Can still be used to potentially assess linear dependence (see next lecture)

**Theoretical explanation** see next lecture

No quiz for this lecture

Have a good week-end!

# Covariance Properties Variance of a Sum, Correlation

Section 8.4

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 16, May 4th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

# Overview

## Previous lecture

- ▶ Covariance, definition, properties, computations

## This lecture

- ▶ Covariance properties and simplified computations
- ▶ Variance of a sum
- ▶ Correlation

## Properties of Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. How can we compute  $\text{Cov}(X_4, X_6)$  without using the joint p.m.f. ?  
(here that would be a multinomial)

## Properties of Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. How can we compute  $\text{Cov}(X_4, X_6)$  without using the joint p.m.f. ?  
(here that would be a multinomial)
2. Can we use that the multinomial can be decomposed in simple r.v.?

## Properties of Covariance

### Example

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. How can we compute  $\text{Cov}(X_4, X_6)$  without using the joint p.m.f. ?  
(here that would be a multinomial)
  2. Can we use that the multinomial can be decomposed in simple r.v.?
- Needs more properties of covariance

## Properties of Covariance

### Lemma (Properties of covariance 1)

*Provided that the covariances defined below are well defined,*

1.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2.  $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$  for any  $a, b \in \mathbb{R}$

## Properties of Covariance

### Lemma (Properties of covariance 1)

*Provided that the covariances defined below are well defined,*

1.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2.  $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$  for any  $a, b \in \mathbb{R}$

### Proof

1. clear from definition
- 2.

$$\text{Cov}(aX + b, Y) = \mathbb{E}[(aX + b)Y] - \mathbb{E}[aX + b]\mathbb{E}[Y]$$

## Properties of Covariance

### Lemma (Properties of covariance 1)

*Provided that the covariances defined below are well defined,*

1.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2.  $\text{Cov}(aX + b, Y) = a \text{Cov}(X, Y)$  for any  $a, b \in \mathbb{R}$

### Proof

1. clear from definition
- 2.

$$\begin{aligned}\text{Cov}(aX + b, Y) &= \mathbb{E}[(aX + b)Y] - \mathbb{E}[aX + b]\mathbb{E}[Y] \\ &= a\mathbb{E}[XY] + b\mathbb{E}[Y] - a\mathbb{E}[X]\mathbb{E}[Y] - b\mathbb{E}[Y]\end{aligned}$$

## Properties of Covariance

### Lemma (Properties of covariance 1)

*Provided that the covariances defined below are well defined,*

1.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2.  $\text{Cov}(aX + b, Y) = a \text{Cov}(X, Y)$  for any  $a, b \in \mathbb{R}$

### Proof

1. clear from definition
- 2.

$$\begin{aligned}\text{Cov}(aX + b, Y) &= \mathbb{E}[(aX + b)Y] - \mathbb{E}[aX + b]\mathbb{E}[Y] \\ &= a\mathbb{E}[XY] + b\mathbb{E}[Y] - a\mathbb{E}[X]\mathbb{E}[Y] - b\mathbb{E}[Y] \\ &= a(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) = a\text{Cov}(X, Y)\end{aligned}$$

## Properties of Covariance

### Lemma (Bilinearity of covariance)

*Provided that the covariances defined below are well defined,*

*For  $X_1, \dots, X_m$ ,  $Y_1, \dots, Y_n$  r.v. and  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbb{R}$ ,*

$$\text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

## Properties of Covariance

### Lemma (Bilinearity of covariance)

Provided that the covariances defined below are well defined,  
For  $X_1, \dots, X_m, Y_1, \dots, Y_n$  r.v. and  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbb{R}$ ,

$$\text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

#### Proof

$$\mu_{X_i} = \mathbb{E}[X_i], \mu_{Y_i} = \mathbb{E}[Y_i] \text{ so } \mathbb{E} \left[ \sum_{i=1}^m X_i \right] = \sum_{i=1}^m \mu_{X_i}, \mathbb{E} \left[ \sum_{j=1}^n Y_j \right] = \sum_{j=1}^n \mu_{Y_j}$$

$$\text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \mathbb{E} \left[ \left( \sum_{i=1}^m a_i X_i - \sum_{i=1}^m a_i \mu_{X_i} \right) \left( \sum_{j=1}^n b_j Y_j - \sum_{j=1}^n b_j \mu_{Y_j} \right) \right]$$

## Properties of Covariance

### Lemma (Bilinearity of covariance)

Provided that the covariances defined below are well defined,  
For  $X_1, \dots, X_m, Y_1, \dots, Y_n$  r.v. and  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbb{R}$ ,

$$\text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

#### Proof

$$\mu_{X_i} = \mathbb{E}[X_i], \mu_{Y_i} = \mathbb{E}[Y_i] \text{ so } \mathbb{E} \left[ \sum_{i=1}^m X_i \right] = \sum_{i=1}^m \mu_{X_i}, \mathbb{E} \left[ \sum_{j=1}^n Y_j \right] = \sum_{j=1}^n \mu_{Y_j}$$

$$\begin{aligned} \text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) &= \mathbb{E} \left[ \left( \sum_{i=1}^m a_i X_i - \sum_{i=1}^m a_i \mu_{X_i} \right) \left( \sum_{j=1}^n b_j Y_j - \sum_{j=1}^n b_j \mu_{Y_j} \right) \right] \\ &= \mathbb{E} \left[ \left( \sum_{i=1}^m a_i (X_i - \mu_{X_i}) \right) \left( \sum_{j=1}^n b_j (Y_j - \mu_{Y_j}) \right) \right] \end{aligned}$$

## Properties of Covariance

### Lemma (Bilinearity of covariance)

Provided that the covariances defined below are well defined,  
For  $X_1, \dots, X_m, Y_1, \dots, Y_n$  r.v. and  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbb{R}$ ,

$$\text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

#### Proof

$$\mu_{X_i} = \mathbb{E}[X_i], \mu_{Y_i} = \mathbb{E}[Y_i] \text{ so } \mathbb{E} \left[ \sum_{i=1}^m X_i \right] = \sum_{i=1}^m \mu_{X_i}, \mathbb{E} \left[ \sum_{j=1}^n Y_j \right] = \sum_{j=1}^n \mu_{Y_j}$$

$$\begin{aligned} \text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) &= \mathbb{E} \left[ \left( \sum_{i=1}^m a_i X_i - \sum_{i=1}^m a_i \mu_{X_i} \right) \left( \sum_{j=1}^n b_j Y_j - \sum_{j=1}^n b_j \mu_{Y_j} \right) \right] \\ &= \mathbb{E} \left[ \left( \sum_{i=1}^m a_i (X_i - \mu_{X_i}) \right) \left( \sum_{j=1}^n b_j (Y_j - \mu_{Y_j}) \right) \right] \\ &= \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^n a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}) \right] \end{aligned}$$

## Properties of Covariance

### Lemma (Bilinearity of covariance)

Provided that the covariances defined below are well defined,  
For  $X_1, \dots, X_m, Y_1, \dots, Y_n$  r.v. and  $a_1, \dots, a_m, b_1, \dots, b_n \in \mathbb{R}$ ,

$$\text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j)$$

#### Proof

$$\mu_{X_i} = \mathbb{E}[X_i], \mu_{Y_i} = \mathbb{E}[Y_i] \text{ so } \mathbb{E} \left[ \sum_{i=1}^m X_i \right] = \sum_{i=1}^m \mu_{X_i}, \mathbb{E} \left[ \sum_{j=1}^n Y_j \right] = \sum_{j=1}^n \mu_{Y_j}$$

$$\begin{aligned} \text{Cov} \left( \sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j \right) &= \mathbb{E} \left[ \left( \sum_{i=1}^m a_i X_i - \sum_{i=1}^m a_i \mu_{X_i} \right) \left( \sum_{j=1}^n b_j Y_j - \sum_{j=1}^n b_j \mu_{Y_j} \right) \right] \\ &= \mathbb{E} \left[ \left( \sum_{i=1}^m a_i (X_i - \mu_{X_i}) \right) \left( \sum_{j=1}^n b_j (Y_j - \mu_{Y_j}) \right) \right] \\ &= \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^n a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}) \right] \\ &= \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathbb{E} [(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})] = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \text{Cov}(X_i, Y_j) \end{aligned}$$

## Covariance of Multinomial

### Example

Let  $(X_1, \dots, X_r) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  with  $p_1 + \dots + p_r = 1$ , i.e.

1. An experiment has  $r$  outcomes
2. The  $i^{\text{th}}$  outcome has proba  $p_i$
3.  $X_i$  is the number of times outcome  $i$  occurs when performing  $n$  independent trials

Find  $\text{Cov}(X_i, X_j)$  for  $i, j \in \{1, \dots, n\}$

## Covariance of Multinomial

### Example

Let  $(X_1, \dots, X_r) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  with  $p_1 + \dots + p_r = 1$ , i.e.

1. An experiment has  $r$  outcomes
2. The  $i^{\text{th}}$  outcome has proba  $p_i$
3.  $X_i$  is the number of times outcome  $i$  occurs when performing  $n$  independent trials

Find  $\text{Cov}(X_i, X_j)$  for  $i, j \in \{1, \dots, n\}$

### Solution

1. **Idea:** Decompose  $X_i$  and  $X_j$  as sum of simple r.v., i.e.  $X_i = \sum_{k=1}^n I_{k,i}$  where

$$I_{k,i} = \begin{cases} 1 & \text{if trial } k \text{ gives outcome } i \\ 0 & \text{if trial } k \text{ gives an outcome other than } i \end{cases}$$

## Covariance of Multinomial

### Example

Let  $(X_1, \dots, X_r) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  with  $p_1 + \dots + p_r = 1$ , i.e.

1. An experiment has  $r$  outcomes
2. The  $i^{\text{th}}$  outcome has proba  $p_i$
3.  $X_i$  is the number of times outcome  $i$  occurs when performing  $n$  independent trials

Find  $\text{Cov}(X_i, X_j)$  for  $i, j \in \{1, \dots, n\}$

### Solution

1. **Idea:** Decompose  $X_i$  and  $X_j$  as sum of simple r.v., i.e.  $X_i = \sum_{k=1}^n I_{k,i}$  where

$$I_{k,i} = \begin{cases} 1 & \text{if trial } k \text{ gives outcome } i \\ 0 & \text{if trial } k \text{ gives an outcome other than } i \end{cases}$$

2.  $I_{k,i} \sim \text{Ber}(p_i)$  so  $X_i \sim \text{Bin}(n, p_i)$  and  $\text{Cov}(X_i, X_j) = \text{Var}(X_i) = np_i(1 - p_i)$

## Covariance of Multinomial

### Example

Let  $(X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  with  $p_1 + \dots + p_r = 1$ , i.e.

1. An experiment has  $r$  outcomes
2. The  $i^{\text{th}}$  outcome has proba  $p_i$
3.  $X_i$  is the number of times outcome  $i$  occurs when performing  $n$  trials

Find  $\text{Cov}(X_i, X_j)$  for  $i, j \in \{1, \dots, n\}$

### Solution

3. For  $i \neq j$ , by bilinearity of the covariance,

$$\text{Cov}(X_i, X_j) = \text{Cov} \left( \sum_{k=1}^n I_{k,i}, \sum_{\ell=1}^n I_{\ell,j} \right) = \sum_{k=1}^n \sum_{\ell=1}^n \text{Cov}(I_{k,i}, I_{\ell,j}) = \sum_{k=1}^n \text{Cov}(I_{k,i}, I_{k,j})$$

using that if  $k \neq l$ ,  $I_{k,i}, I_{\ell,j}$  are independent by definition.

# Covariance of Multinomial

## Example

Let  $(X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  with  $p_1 + \dots + p_r = 1$ , i.e.

1. An experiment has  $r$  outcomes
2. The  $i^{\text{th}}$  outcome has proba  $p_i$
3.  $X_i$  is the number of times outcome  $i$  occurs when performing  $n$  trials

Find  $\text{Cov}(X_i, X_j)$  for  $i, j \in \{1, \dots, n\}$

## Solution

3. For  $i \neq j$ , by bilinearity of the covariance,

$$\text{Cov}(X_i, X_j) = \text{Cov} \left( \sum_{k=1}^n I_{k,i}, \sum_{\ell=1}^n I_{\ell,j} \right) = \sum_{k=1}^n \sum_{\ell=1}^n \text{Cov}(I_{k,i}, I_{\ell,j}) = \sum_{k=1}^n \text{Cov}(I_{k,i}, I_{k,j})$$

using that if  $k \neq \ell$ ,  $I_{k,i}, I_{\ell,j}$  are independent by definition.

Since  $i \neq j$ ,  $I_{k,i} I_{k,j} = 0$ , because on trial  $k$  both outcomes cannot occur

$$\text{Cov}(I_{k,i}, I_{k,j}) = \mathbb{E}[I_{k,i} I_{k,j}] - \mathbb{E}[I_{k,i}] \mathbb{E}[I_{k,j}] = 0 - p_i p_j$$

Therefore  $\text{Cov}(X_i, X_j) = -np_i p_j < 0$ ,

→ the more often  $i$  occurs, the fewer opportunities for outcome  $j$

# Variance of a Sum of Random Variables

## Motivation

- ▶ We have seen how to compute the variance of a sum of **independent** r.v.
- ▶ What about a sum of non-independent r.v. ?
- Needs to take into account the interactions btw the elements of the sum,  
i.e. their covariance!

## Variance of a Sum of Random Variables

### Corollary

Let  $X_1, \dots, X_n$  be  $n$  r.v. with finite variance and covariances (between each pair)

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

## Variance of a Sum of Random Variables

### Corollary

Let  $X_1, \dots, X_n$  be  $n$  r.v. with finite variance and covariances (between each pair)

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

**Proof**  $\text{Var}\left(\sum_{i=1}^n X_i\right) = \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j)$

Then identify  $\text{Var}(X_i) = \text{Cov}(X_i, X_i)$  in the sum and simplify the rest of the sum.  
Namely

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) &= \sum_{i=1}^n \text{Var}(X_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^n \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j) \end{aligned}$$

where we used in the last equality that  $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ .

## Variance of a Sum of Random Variables

### Corollary

Let  $X_1, \dots, X_n$  be  $n$  r.v. with finite variance and covariances (between each pair)

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

### Corollary

Let  $X_1, \dots, X_n$  be  $n$  uncorrelated r.v. ( $\text{Cov}(X_i, X_j) = 0$  for  $i, j \in \{1, \dots, n\}, i \neq j$ )

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i)$$

## Variance of a Sum of Random Variables

### Corollary

Let  $X_1, \dots, X_n$  be  $n$  r.v. with finite variance and covariances (between each pair)

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j)$$

### Corollary

Let  $X_1, \dots, X_n$  be  $n$  uncorrelated r.v. ( $\text{Cov}(X_i, X_j) = 0$  for  $i, j \in \{1, \dots, n\}, i \neq j$ )

$$\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i)$$

### Remark:

We retrieve result from last lectures that the

"variance of sum of independent r.v. is the sum of their variance"

Note however that the uncorrelated assumption is weaker than independence

## Variance of a Sum of Random Variables

### Example

Each morning a person eats bread with probability 0.5 (event  $A$ ), they eat oatmeal with probability 0.2 (event  $B$ ) and they eat both with probability 0.1 (event  $A \cap B$ ). (They can also eat nothing)

Let  $X = I_A + I_B$  be the random variables that counts how many of the events  $A$  and  $B$  occurs, i.e. how many different meals that person eats every morning.

Find  $\text{Var}(X)$ .

## Variance of a Sum of Random Variables

### Example

Each morning a person eats bread with probability 0.5 (event  $A$ ), they eat oatmeal with probability 0.2 (event  $B$ ) and they eat both with probability 0.1 (event  $A \cap B$ ). (They can also eat nothing)

Let  $X = I_A + I_B$  be the random variables that counts how many of the events  $A$  and  $B$  occurs, i.e. how many different meals that person eats every morning.  
Find  $\text{Var}(X)$ .

**Solution**  $I_A \sim \text{Ber}(p_A)$  with  $p_A = 0.5$ ,  $I_B \sim \text{Ber}(p_B)$  with  $p_B = 0.2$

## Variance of a Sum of Random Variables

### Example

Each morning a person eats bread with probability 0.5 (event  $A$ ), they eat oatmeal with probability 0.2 (event  $B$ ) and they eat both with probability 0.1 (event  $A \cap B$ ). (They can also eat nothing)

Let  $X = I_A + I_B$  be the random variables that counts how many of the events  $A$  and  $B$  occurs, i.e. how many different meals that person eats every morning.  
Find  $\text{Var}(X)$ .

**Solution**  $I_A \sim \text{Ber}(p_A)$  with  $p_A = 0.5$ ,  $I_B \sim \text{Ber}(p_B)$  with  $p_B = 0.2$

Using

$$\text{Cov}(I_A, I_B) = \mathbb{E}[I_A I_B] - \mathbb{E}[I_A] \mathbb{E}[I_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A) \mathbb{P}(B)$$

## Variance of a Sum of Random Variables

### Example

Each morning a person eats bread with probability 0.5 (event  $A$ ), they eat oatmeal with probability 0.2 (event  $B$ ) and they eat both with probability 0.1 (event  $A \cap B$ ). (They can also eat nothing)

Let  $X = I_A + I_B$  be the random variables that counts how many of the events  $A$  and  $B$  occurs, i.e. how many different meals that person eats every morning.  
Find  $\text{Var}(X)$ .

**Solution**  $I_A \sim \text{Ber}(p_A)$  with  $p_A = 0.5$ ,  $I_B \sim \text{Ber}(p_B)$  with  $p_B = 0.2$

Using

$$\text{Cov}(I_A, I_B) = \mathbb{E}[I_A I_B] - \mathbb{E}[I_A] \mathbb{E}[I_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A) \mathbb{P}(B)$$

We get

$$\begin{aligned}\text{Var}(X) &= \text{Var}(I_A) + \text{Var}(I_B) + 2 \text{Cov}(I_A, I_B) \\ &= p_A(1 - p_A) + p_B(1 - p_B) + 2(\mathbb{P}(A \cap B) - \mathbb{P}(A) \mathbb{P}(B)) \\ &= 0.25 + 0.16 + 2(0.1 - 0.1) = 0.41\end{aligned}$$

## Motivation

- ▶ We said that  $\text{Cov}(X, Y)$  could be a good proxy of dependence
- ▶ Yet, by bilinearity,  $\text{Cov}(10X, 7Y) = 70 \text{Cov}(X, Y)$
- ▶ So a huge covariance can simply be the result of a scaling of the r.v. and not signify something about their dependence
- needs a scaling invariant measure: correlation!

## Correlation

### Definition (Correlation)

The **correlation** (or **correlation coefficient**) of two r.v.  $X, Y$  with positive finite variances is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}$$

It is sometimes denoted  $\rho(X, Y)$  or  $\rho_{X,Y}$ .

## Correlation

### Definition (Correlation)

The **correlation** (or **correlation coefficient**) of two r.v.  $X, Y$  with positive finite variances is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}$$

It is sometimes denoted  $\rho(X, Y)$  or  $\rho_{X,Y}$ .

### Lemma (Scaling invariance)

Let  $X, Y$  be two r.v. with positive finite variances and  $a, b \in \mathbb{R}$ ,  $a \neq 0$

$$\text{Corr}(aX + b, Y) = \frac{a}{|a|} \text{Corr}(X, Y)$$

## Correlation

### Definition (Correlation)

The **correlation** (or **correlation coefficient**) of two r.v.  $X, Y$  with positive finite variances is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}$$

It is sometimes denoted  $\rho(X, Y)$  or  $\rho_{X,Y}$ .

### Lemma (Scaling invariance)

Let  $X, Y$  be two r.v. with positive finite variances and  $a, b \in \mathbb{R}$ ,  $a \neq 0$

$$\text{Corr}(aX + b, Y) = \frac{a}{|a|} \text{Corr}(X, Y)$$

**Proof**  $\text{Corr}(aX + b, Y) = \frac{\text{Cov}(aX+b, Y)}{\sqrt{\text{Var}(aX+b)} \sqrt{\text{Var}(Y)}} = \frac{a \text{Cov}(X, Y)}{\sqrt{a^2 \text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{a}{|a|} \text{Corr}(X, Y)$

## Correlation

### Lemma (Properties of correlation 1)

*Let  $X, Y$  be two r.v. with positive finite variances. Then  $-1 \leq \text{Corr}(X, Y) \leq 1$*

## Correlation

### Lemma (Properties of correlation 1)

Let  $X, Y$  be two r.v. with positive finite variances. Then  $-1 \leq \text{Corr}(X, Y) \leq 1$

**Idea** Use **standardized** r.v. i.e. centered & normalized by standard deviation

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X} \quad \tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}$$

where  $\mu_X = \mathbb{E}[X]$ ,  $\sigma_X^2 = \text{Var}(X)$ ,  $\mu_Y = \mathbb{E}[Y]$ ,  $\sigma_Y^2 = \text{Var}(Y)$ , s.t.

$$\mathbb{E}[\tilde{X}] = 0, \quad \text{Var}(\tilde{X}) = \mathbb{E}[\tilde{X}^2] = \mathbb{E}\left[\frac{(X - \mu_X)^2}{\sigma_X^2}\right] = 1$$

Same for  $\tilde{Y}$  and finally

$$\mathbb{E}[\tilde{X}\tilde{Y}] = \mathbb{E}\left[\frac{X - \mu_X}{\sigma_X} \frac{Y - \mu_Y}{\sigma_Y}\right] = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \text{Corr}(X, Y)$$

## Correlation

### Lemma (Properties of correlation 1)

Let  $X, Y$  be two r.v. with positive finite variances. Then  $-1 \leq \text{Corr}(X, Y) \leq 1$

**Idea** Use **standardized** r.v. i.e. centered & normalized by standard deviation

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X} \quad \tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}$$

where  $\mu_X = \mathbb{E}[X], \sigma_X^2 = \text{Var}(X), \mu_Y = \mathbb{E}[Y], \sigma_Y^2 = \text{Var}(Y)$ , s.t.

$$\mathbb{E}[\tilde{X}] = 0, \quad \text{Var}(\tilde{X}) = \mathbb{E}[\tilde{X}^2] = \mathbb{E}\left[\frac{(X - \mu_X)^2}{\sigma_X^2}\right] = 1$$

Same for  $\tilde{Y}$  and finally

$$\mathbb{E}[\tilde{X}\tilde{Y}] = \mathbb{E}\left[\frac{X - \mu_X}{\sigma_X} \frac{Y - \mu_Y}{\sigma_Y}\right] = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \text{Corr}(X, Y)$$

**Proof** of the lemma

$$0 \leq \mathbb{E}[(\tilde{X} - \tilde{Y})^2] = \mathbb{E}[\tilde{X}^2] + \mathbb{E}[\tilde{Y}^2] - 2\mathbb{E}[\tilde{X}\tilde{Y}] = 2(1 - \text{Corr}(X, Y))$$

Therefore  $1 - \text{Corr}(X, Y) \geq 0$ , i.e.  $\text{Corr}(X, Y) \leq 1$ .

Similarly  $0 \leq \mathbb{E}[(\tilde{X} + \tilde{Y})^2] = 2(1 + \text{Corr}(X, Y))$  so  $\text{Corr}(X, Y) \geq -1$

## Correlation

### Lemma (Properties of correlation 2)

Let  $X, Y$  be two r.v. with positive finite variances.

1.  $\text{Corr}(X, Y) = 1 \iff \exists a > 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$
2.  $\text{Corr}(X, Y) = -1 \iff \exists a < 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$

and naturally  $\text{Corr}(X, X) = 1$

## Correlation

### Lemma (Properties of correlation 2)

Let  $X, Y$  be two r.v. with positive finite variances.

1.  $\text{Corr}(X, Y) = 1 \iff \exists a > 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$
2.  $\text{Corr}(X, Y) = -1 \iff \exists a < 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$

and naturally  $\text{Corr}(X, X) = 1$

**Proof** of 1. (proof of 2. is analogous)

If  $Y = aX + b$  then by scaling invariance  $\text{Corr}(X, Y) = \frac{a}{|a|} \text{Corr}(X, X) = 1$

Assume  $\text{Corr}(X, Y) = 1$ , denote  $\tilde{X} = \frac{X - \mu_X}{\sigma_X}$ ,  $\tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y}$  and  $Z = \tilde{X} - \tilde{Y}$

$$\mathbb{E}[Z] = 0, \quad \text{Var}(Z) = \mathbb{E}[(\tilde{X} - \tilde{Y})^2] = 2(1 - \text{Corr}(X, Y)) = 0$$

So  $Z = 0$ , i.e.,  $\tilde{X} = \tilde{Y}$ , i.e.,

$$Y = \frac{\sigma_Y}{\sigma_X}X + \mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X = aX + b$$

with  $a = \frac{\sigma_Y}{\sigma_X} > 0$ .

## Correlation

### Lemma (Properties of correlation 2)

Let  $X, Y$  be two r.v. with positive finite variances.

1.  $\text{Corr}(X, Y) = 1 \iff \exists a > 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$
2.  $\text{Corr}(X, Y) = -1 \iff \exists a < 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$

and naturally  $\text{Corr}(X, X) = 1$

### Remark

Let  $X, Y$  be two r.v. such that  $\text{Corr}(X, Y) = 1$ .

Can  $X, Y$  be jointly continuous?

## Correlation

### Lemma (Properties of correlation 2)

Let  $X, Y$  be two r.v. with positive finite variances.

1.  $\text{Corr}(X, Y) = 1 \iff \exists a > 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$
2.  $\text{Corr}(X, Y) = -1 \iff \exists a < 0, b \in \mathbb{R}, \text{ s.t. } Y = aX + b$

and naturally  $\text{Corr}(X, X) = 1$

### Remark

Let  $X, Y$  be two r.v. such that  $\text{Corr}(X, Y) = 1$ .

Can  $X, Y$  be jointly continuous?

**Solution** No! Indeed,  $Y = aX + b$  with  $a > 0, b \in \mathbb{R}$ , so  $\mathbb{P}(Y = aX + b) = 1$

If they were jointly continuous, we would have

$$\mathbb{P}(Y = aX + b) = \int_{-\infty}^{+\infty} \int_{ax+b}^{ax+b} f_{X,Y}(x, y) dy dx = 0$$

In this case we would say that the random vector  $(X, Y)$  is **degenerated**  
(similarly as when  $\text{Var}(X) = 0$  for a single r.v.)

## Quiz next lecture

### Exercise

1. Roll a die 10 times, denote  $X_1, X_2$  the number of 1 and 2 that you get.
  - 1.1 Compute  $\text{Corr}(X_1, X_2)$
2. Flip a coin 10 times, denote  $X_1, X_2$  the number of tails and heads respectively.
  - 2.1 Compute  $\text{Corr}(X_1, X_2)$
  - 2.2 How could you have found it ?

## Characterization of Independence\*

The key lemma to show that  $\text{Independence} \Rightarrow \text{Cov}(X, Y) = 0$  is

### Lemma

If  $X, Y$  are two independent r.v. then for any  $g, h : \mathbb{R} \rightarrow \mathbb{R}$  s.t.  $\mathbb{E}[g(X)]$ ,  $\mathbb{E}[h(Y)]$  are finite,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

## Characterization of Independence\*

The key lemma to show that  $\text{Independence} \Rightarrow \text{Cov}(X, Y) = 0$  is

### Lemma

If  $X, Y$  are two independent r.v. then for any  $g, h : \mathbb{R} \rightarrow \mathbb{R}$  s.t.  $\mathbb{E}[g(X)]$ ,  $\mathbb{E}[h(Y)]$  are finite,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

Conversely we have the following theorem

### Theorem

Let  $X, Y$  be two r.v.. If for any  $g, h$  bounded s.t.  $\mathbb{E}[g(X)], \mathbb{E}[h(Y)]$  are finite, the following holds

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

then  $X, Y$  are independent.

## Characterization of Independence\*

The key lemma to show that  $\text{Independence} \Rightarrow \text{Cov}(X, Y) = 0$  is

### Lemma

If  $X, Y$  are two independent r.v. then for any  $g, h : \mathbb{R} \rightarrow \mathbb{R}$  s.t.  $\mathbb{E}[g(X)]$ ,  $\mathbb{E}[h(Y)]$  are finite,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

Conversely we have the following theorem

### Theorem

Let  $X, Y$  be two r.v.. If for any  $g, h$  bounded s.t.  $\mathbb{E}[g(X)]$ ,  $\mathbb{E}[h(Y)]$  are finite, the following holds

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

then  $X, Y$  are independent.

**Proof** Take  $g, h$  two be any indicator functions of Borel sets, you get the definition of independence.

## Characterization of Independence\*

The key lemma to show that  $\text{Independence} \Rightarrow \text{Cov}(X, Y) = 0$  is

### Lemma

If  $X, Y$  are two independent r.v. then for any  $g, h : \mathbb{R} \rightarrow \mathbb{R}$  s.t.  $\mathbb{E}[g(X)]$ ,  $\mathbb{E}[h(Y)]$  are finite,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

Conversely we have the following theorem

### Theorem

Let  $X, Y$  be two r.v.. If for any  $g, h$  bounded s.t.  $\mathbb{E}[g(X)]$ ,  $\mathbb{E}[h(Y)]$  are finite, the following holds

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

then  $X, Y$  are independent.

Covariance only checks for one particular choice of  $h$  and  $g$ .  
It is not sufficient.

## Additional exercises

### Exercise

*Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get*

1. Compute  $\text{Cov}(X_4, X_6)$

## Additional exercises

### Exercise

Roll a die 10 times, denote  $X_4, X_6$  the number of 4 and 6 resp. that you get

1. Compute  $\text{Cov}(X_4, X_6)$

**Solution** Denote  $Y$  the number of any other outcomes than a 4 or a 6.

We have  $X_4, X_6, Y \sim \text{Multinom}(10, 3, 1/6, 1/6, 2/3)$ .

$$\text{So } \text{Cov}(X_4, X_6) = -10/36 \approx -0.28$$

## Additional Exercises

### Exercise

Let  $X, Y$  be two r.v. s.t.  $\mathbb{E}[X] = 1$ ,  $\mathbb{E}[X^2] = 3$ ,  $\mathbb{E}[XY] = -4$ ,  $\mathbb{E}[Y] = 2$ . Find  $\text{Cov}(X, 2X + Y - 3)$

## Additional Exercises

### Exercise

Let  $X, Y$  be two r.v. s.t.  $\mathbb{E}[X] = 1$ ,  $\mathbb{E}[X^2] = 3$ ,  $\mathbb{E}[XY] = -4$ ,  $\mathbb{E}[Y] = 2$ . Find  $\text{Cov}(X, 2X + Y - 3)$

### Solution

$$\text{Cov}(X, 2X + Y - 3) = 2 \text{Var}(X) + \text{Cov}(X, Y) = 2(\mathbb{E}[X^2] - \mathbb{E}[X]^2) + \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = -2$$

# Multivariate Normal Distribution Moment Generating Function

Sections 8.5 8.6 5.2

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 16, May 4th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

**Additional office hour** Friday 2:30 to 3:30

### **Lecture notes**

Will be available at the end of the week to cover what has been seen

**Reminder** Additional exercises available at the end of the lectures

### **Previous Lectures**

- ▶ Covariance
- ▶ Variance of a sum
- ▶ Correlation

### **This lecture**

- ▶ Multivariate normal distribution
- ▶ Moment generating function definition

## Quiz Previous Lecture

### Exercise

1. Roll a die 10 times, denote  $X_1, X_2$  the number of 1 and 2 that you get.
  - 1.1 Compute  $\text{Corr}(X_1, X_2)$
2. Flip a coin 10 times, denote  $X_1, X_2$  the number of tails and heads respectively.
  - 2.1 Compute  $\text{Corr}(X_1, X_2)$
  - 2.2 How could you have found it ?

## Quiz Previous Lecture

### Exercise

1. Roll a die 10 times, denote  $X_1, X_2$  the number of 1 and 2 that you get.
  - 1.1 Compute  $\text{Corr}(X_1, X_2)$
2. Flip a coin 10 times, denote  $X_1, X_2$  the number of tails and heads respectively.
  - 2.1 Compute  $\text{Corr}(X_1, X_2)$
  - 2.2 How could you have found it ?

### Solution

1. We give directly the correlation of  $(X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  (here  $n = 10$ ,  $r = 6$ ,  $p_i = 1/6$ )

We saw that

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j \end{cases}$$

## Quiz Previous Lecture

### Exercise

1. Roll a die 10 times, denote  $X_1, X_2$  the number of 1 and 2 that you get.
  - 1.1 Compute  $\text{Corr}(X_1, X_2)$
2. Flip a coin 10 times, denote  $X_1, X_2$  the number of tails and heads respectively.
  - 2.1 Compute  $\text{Corr}(X_1, X_2)$
  - 2.2 How could you have found it ?

### Solution

1. We give directly the correlation of  $(X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  (here  $n = 10$ ,  $r = 6$ ,  $p_i = 1/6$ )

We saw that

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j \end{cases}$$

So for  $i \neq j$

$$\text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)} \sqrt{\text{Var}(X_j)}} = \frac{-np_i p_j}{\sqrt{np_i(1 - p_i)} \sqrt{np_j(1 - p_j)}} = -\sqrt{\frac{p_i p_j}{(1 - p_i)(1 - p_j)}}$$

$$\text{So here } \text{Corr}(X_i, X_j) = -\sqrt{1/25} = -1/5 = -0.2$$

## Quiz Previous Lecture

### Exercise

1. Roll a die 10 times, denote  $X_1, X_2$  the number of 1 and 2 that you get.
  - 1.1 Compute  $\text{Corr}(X_1, X_2)$
2. Flip a coin 10 times, denote  $X_1, X_2$  the number of tails and heads respectively.
  - 2.1 Compute  $\text{Corr}(X_1, X_2)$
  - 2.2 How could you have found it ?

### Solution

1. We give directly the correlation of  $(X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$  (here  $n = 10$ ,  $r = 6$ ,  $p_i = 1/6$ )

We saw that

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j \end{cases}$$

So for  $i \neq j$

$$\text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)} \sqrt{\text{Var}(X_j)}} = \frac{-np_i p_j}{\sqrt{np_i(1 - p_i)} \sqrt{np_j(1 - p_j)}} = -\sqrt{\frac{p_i p_j}{(1 - p_i)(1 - p_j)}}$$

$$\text{So here } \text{Corr}(X_i, X_j) = -\sqrt{1/25} = -1/5 = -0.2$$

2. In the case  $r = 2$  s.t.  $p_1 = 1 - p_2$ ,

$$\text{Corr}(X_1, X_2) = -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}} = -1$$

That reflects that  $X_2 = n - X_1$  for binomial.

# Multivariate Normal Distribution

## Motivation

1. The standard normal distribution plays a central role for r.v.
2. What about its generalization for  $n$  random variables?

# Multivariate Normal Distribution

## Motivation

1. The standard normal distribution plays a central role for r.v.
2. What about its generalization for  $n$  random variables?

## Idea

1. We saw that mean and variance entirely characterize the normal distribution
2. Same for multivariate, except that one needs to incorporate covariance between the variables!

## Mean Vector

### Definition (Random vector (reminder))

A **multivariate random variable** or **random vector** is a vector

$\mathbf{X} = (X_1, \dots, X_n)^\top$  whose components are r.v. on the same proba. space

## Mean Vector

### Definition (Random vector (reminder))

A **multivariate random variable** or **random vector** is a vector

$\mathbf{X} = (X_1, \dots, X_n)^\top$  whose components are r.v. on the same proba. space

### Definition (Mean vector)

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector, its mean vector is defined as

$$\mu_{\mathbf{X}} \triangleq \mathbb{E}[\mathbf{X}] \triangleq \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

## Mean Vector

### Definition (Random vector (reminder))

A **multivariate random variable** or **random vector** is a vector

$\mathbf{X} = (X_1, \dots, X_n)^\top$  whose components are r.v. on the same proba. space

### Definition (Mean vector)

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector, its mean vector is defined as

$$\mu_{\mathbf{X}} \triangleq \mathbb{E}[\mathbf{X}] \triangleq \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

### Example

Let  $\mathbf{X} = (X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$ , what is  $\mu_{\mathbf{X}}$ ?

## Mean Vector

### Definition (Random vector (reminder))

A **multivariate random variable** or **random vector** is a vector

$\mathbf{X} = (X_1, \dots, X_n)^\top$  whose components are r.v. on the same proba. space

### Definition (Mean vector)

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector, its mean vector is defined as

$$\mu_{\mathbf{X}} \triangleq \mathbb{E}[\mathbf{X}] \triangleq \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

### Example

Let  $\mathbf{X} = (X_1, \dots, X_n) \sim \text{Multinom}(n, r, p_1, \dots, p_r)$ , what is  $\mu_{\mathbf{X}}$ ?

**Solution**  $X_i \sim \text{Bin}(n, p_i)$  (use decomposition seen in previous lecture)

$$\mu_{\mathbf{X}} = \begin{pmatrix} np_1 \\ \vdots \\ np_n \end{pmatrix}$$

## Mean Vector

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$ ,  $A = (A_{ij})_{\substack{i=1, \dots, p \\ j=1, \dots, n}} \in \mathbb{R}^{p \times n}$  and  $b = (b_i)_{i=1}^p \in \mathbb{R}^p$ , then

$$\mathbb{E}[A\mathbf{X} + b] = A\mathbb{E}[\mathbf{X}] + b$$

## Mean Vector

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$ ,  $A = (A_{ij})_{\substack{i=1, \dots, p \\ j=1, \dots, n}} \in \mathbb{R}^{p \times n}$  and  $b = (b_i)_{i=1}^p \in \mathbb{R}^p$ , then

$$\mathbb{E}[A\mathbf{X} + b] = A\mathbb{E}[\mathbf{X}] + b$$

**Proof** Denote  $\mathbf{Y} = A\mathbf{X} + b = (Y_1, \dots, Y_p)$ ,

$$Y_i = \sum_{j=1}^n A_{ij} X_j + b_i$$

$$\mathbb{E}[Y_i] = \sum_{j=1}^n A_{ij} \mathbb{E}[X_j] + b_i$$

So  $\mathbb{E}[A\mathbf{X} + b] = A\mathbb{E}[\mathbf{X}] + b$

## Covariance matrix

### Definition

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector its **covariance matrix** is defined as

$$S_{\mathbf{X}} = \begin{pmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{pmatrix} = (\text{Cov}(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

1.  $S_{\mathbf{X}}$  is symmetric, i.e.  $(S_{\mathbf{X}})_{ij} = (S_{\mathbf{X}})_{ji} = \text{Cov}(X_i, X_j)$
2. The diagonal of  $S_{\mathbf{X}}$  represents the variances  $(S_{\mathbf{X}})_{ii} = \text{Var}(X_i)$

## Covariance matrix

### Definition

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector its **covariance matrix** is defined as

$$S_{\mathbf{X}} = \begin{pmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Cov}(X_n, X_n) \end{pmatrix} = (\text{Cov}(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

1.  $S_{\mathbf{X}}$  is symmetric, i.e.  $(S_{\mathbf{X}})_{ij} = (S_{\mathbf{X}})_{ji} = \text{Cov}(X_i, X_j)$
2. The diagonal of  $S_{\mathbf{X}}$  represents the variances  $(S_{\mathbf{X}})_{ii} = \text{Var}(X_i)$

### Example

Let  $X, Y$  be two random variables, their covariance matrix is

$$S = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{pmatrix}$$

## Covariance Matrix

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random vector with covariance matrix  $S_{\mathbf{X}}$ .

Let  $A \in \mathbb{R}^{p \times n}$ ,  $b \in \mathbb{R}^p$ , the covariance of the random vector  $\mathbf{Y} = A\mathbf{X} + b \in \mathbb{R}^p$  is

$$S_{\mathbf{Y}} = AS_{\mathbf{X}}A^{\top} \in \mathbb{R}^{p \times p}$$

where  $A^{\top}$  is the transpose of  $A$ , i.e.,  $(A^{\top})_{ij} = A_{ji}$

## Covariance Matrix

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random vector with covariance matrix  $S_{\mathbf{X}}$ .

Let  $A \in \mathbb{R}^{p \times n}$ ,  $b \in \mathbb{R}^p$ , the covariance of the random vector  $\mathbf{Y} = A\mathbf{X} + b \in \mathbb{R}^p$  is

$$S_{\mathbf{Y}} = AS_{\mathbf{X}}A^{\top} \in \mathbb{R}^{p \times p}$$

where  $A^{\top}$  is the transpose of  $A$ , i.e.,  $(A^{\top})_{ij} = A_{ji}$

**Proof** (See additional slides at the end of the lecture )

## Multivariate Normal Random Vector

Definition (Standard normal random vector)

A random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is a **standard normal random vector** if  $X_1, \dots, X_n$  are i.i.d. standard normal r.v. ( $X_i \sim \mathcal{N}(0, 1)$ ) s.t.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x_1^2 + \dots + x_n^2)}$$

## Multivariate Normal Random Vector

### Definition (Standard normal random vector)

A random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is a **standard normal random vector** if  $X_1, \dots, X_n$  are i.i.d. standard normal r.v. ( $X_i \sim \mathcal{N}(0, 1)$ ) s.t.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x_1^2 + \dots + x_n^2)}$$

### Question

What are the mean and covariance matrix of a standard normal random vector?

## Multivariate Normal Random Vector

Definition (Standard normal random vector)

A random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is a **standard normal random vector** if  $X_1, \dots, X_n$  are i.i.d. standard normal r.v. ( $X_i \sim \mathcal{N}(0, 1)$ ) s.t.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x_1^2 + \dots + x_n^2)}$$

Property

A standard normal random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  satisfies

$$\mu_{\mathbf{X}} = \mathbf{0}_n \triangleq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad S_{\mathbf{X}} = \mathbf{I}_n \triangleq \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

It is denoted  $\mathbf{X} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$

## Multivariate Normal Random Vector

### Definition (Multivariate Normal Distribution )

A random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is a **normal random vector** if there exist  $\mu \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, \mathbf{Z} \sim \mathcal{N}(0_n, I_n)$  s.t.

$$\mathbf{X} = A\mathbf{Z} + \mu$$

## Multivariate Normal Random Vector

Definition (Multivariate Normal Distribution )

A random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is a **normal random vector** if there exist  $\mu \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, \mathbf{Z} \sim \mathcal{N}(0_n, I_n)$  s.t.

$$\mathbf{X} = A\mathbf{Z} + \mu$$

Question

What are the mean and covariance matrix of a standard normal random vector?

## Multivariate Normal Random Vector

### Definition (Multivariate Normal Distribution )

A random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top$  is a **normal random vector** if there exist  $\mu \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, \mathbf{Z} \sim \mathcal{N}(0_n, I_n)$  s.t.

$$\mathbf{X} = A\mathbf{Z} + \mu$$

### Property

A *normal random vector*  $\mathbf{X} = (X_1, \dots, X_n)^\top$  satisfies as defined above

$$\mu_{\mathbf{X}} = \mu \quad S_{\mathbf{X}} = AA^\top$$

It is denoted  $\mathbf{X} \sim \mathcal{N}(\mu, S)$  with  $S = S_{\mathbf{X}}$ .

## Multivariate Normal Distribution

### Definition

A normal random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with **invertible covariance matrix** has a joint p.d.f.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

## Multivariate Normal Distribution

### Definition

A normal random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with **invertible covariance matrix** has a joint p.d.f.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

**Proof** (See additional slides for a sketch of proof)

## Multivariate Normal Distribution

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with finite positive marginal variances ( $0 < \text{Var}(X_i) < +\infty$ ),

$\text{Cov}(X_i, X_j) = 0$  for all  $i \neq j \iff X_1, \dots, X_n$  are independent

## Multivariate Normal Distribution

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with finite positive marginal variances ( $0 < \text{Var}(X_i) < +\infty$ ),

$$\text{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \iff X_1, \dots, X_n \text{ are independent}$$

**Proof** If  $\text{Cov}(X_i, X_j) = 0$  for all  $i \neq j$ , then

$$S = \begin{pmatrix} \sigma_{X_1}^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{X_n}^2 \end{pmatrix} \quad S^{-1} = \begin{pmatrix} \sigma_{X_1}^{-2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{X_n}^{-2} \end{pmatrix}$$

## Multivariate Normal Distribution

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with finite positive marginal variances ( $0 < \text{Var}(X_i) < +\infty$ ),

$$\text{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \iff X_1, \dots, X_n \text{ are independent}$$

**Proof** If  $\text{Cov}(X_i, X_j) = 0$  for all  $i \neq j$ , then

$$S = \begin{pmatrix} \sigma_{X_1}^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{X_n}^2 \end{pmatrix} \quad S^{-1} = \begin{pmatrix} \sigma_{X_1}^{-2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{X_n}^{-2} \end{pmatrix}$$

So

$$\begin{aligned} f_{\mathbf{X}}(x_1, \dots, x_n) &= \frac{1}{(2\pi)^{n/2} \sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\ &= \frac{1}{(2\pi)^{n/2} \sigma_{X_1} \dots \sigma_{X_n}} e^{-\left(\frac{(x_1-\mu_1)^2}{2\sigma_{X_1}^2} + \dots + \frac{(x_n-\mu_n)^2}{2\sigma_{X_n}^2}\right)} = f_1(x_1) \dots f_n(x_n) \end{aligned}$$

## Multivariate Normal Distribution

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with finite positive marginal variances ( $0 < \text{Var}(X_i) < +\infty$ ),

$$\text{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \iff X_1, \dots, X_n \text{ are independent}$$

**Proof** If  $\text{Cov}(X_i, X_j) = 0$  for all  $i \neq j$ , then

$$S = \begin{pmatrix} \sigma_{X_1}^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{X_n}^2 \end{pmatrix} \quad S^{-1} = \begin{pmatrix} \sigma_{X_1}^{-2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{X_n}^{-2} \end{pmatrix}$$

So

$$\begin{aligned} f_{\mathbf{X}}(x_1, \dots, x_n) &= \frac{1}{(2\pi)^{n/2} \sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\ &= \frac{1}{(2\pi)^{n/2} \sigma_{X_1} \dots \sigma_{X_n}} e^{-\left(\frac{(x_1-\mu_1)^2}{2\sigma_{X_1}^2} + \dots + \frac{(x_n-\mu_n)^2}{2\sigma_{X_n}^2}\right)} = f_1(x_1) \dots f_n(x_n) \end{aligned}$$

The joint p.d.f. factorizes in functions of each r.v. (that can be shown to be the marginals) so  $(X_1, \dots, X_n)$  are independent.

## Multivariate Normal Distribution

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with finite positive marginal variances ( $0 < \text{Var}(X_i) < +\infty$ ),

$$\text{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \iff X_1, \dots, X_n \text{ are independent}$$

### Intuition:

- ▶ Mean and covariance matrix entirely define a normal random vector.
- ▶ No need to capture more information than covariance on the random variables to assess their independence

## Moment Generating Function/Characteristic Function

### Motivation

1. We saw that for normal r.v. or random vectors, knowing first and second moments are sufficient

# Moment Generating Function/Characteristic Function

## Motivation

1. We saw that for normal r.v. or random vectors, knowing first and second moments are sufficient
2. Is there a way to describe a r.v. only through its moments?

## Moment Generating Function/Characteristic Function

### Motivation

1. We saw that for normal r.v. or random vectors, knowing first and second moments are sufficient
2. Is there a way to describe a r.v. only through its moments?
3. The moment generating function and the characteristic functions are alternative ways to describe a r.v.  
(rather than using p.m.f/p.d.f or c.d.f.)

## Moment Generating Function/Characteristic Function

### Definition

The **moment generating function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{R}$  defined by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

## Moment Generating Function/Characteristic Function

### Definition

The **moment generating function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{R}$  defined by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The **characteristic function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{C}$  defined by

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

## Moment Generating Function/Characteristic Function

### Definition

The **moment generating function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{R}$  defined by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The **characteristic function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{C}$  defined by

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

### Theoretical intuitions\*

If  $X$  is continuous with p.d.f.  $f$ , then

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x) dx = \mathcal{L}(f)(-t) \quad \phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx = \mathcal{F}(f)(-t)$$

where  $\mathcal{L}(f), \mathcal{F}(f)$  are the *Laplace* and *Fourier* transforms of  $f$

## Moment Generating Function/Characteristic Function

### Definition

The **moment generating function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{R}$  defined by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The **characteristic function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{C}$  defined by

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

### Theoretical intuitions\*

If  $X$  is continuous with p.d.f.  $f$ , then

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x) dx = \mathcal{L}(f)(-t) \quad \phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx = \mathcal{F}(f)(-t)$$

where  $\mathcal{L}(f), \mathcal{F}(f)$  are the *Laplace* and *Fourier* transforms of  $f$

→ As for e.g. sounds, these transforms can provide alternative descriptions.

## Moment Generating Function/Characteristic Function

### Definition

The **moment generating function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{R}$  defined by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The **characteristic function** of a r.v.  $X$  is a function from  $\mathbb{R}$  to  $\mathbb{C}$  defined by

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

### Theoretical intuitions\*

If  $X$  is continuous with p.d.f.  $f$ , then

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x) dx = \mathcal{L}(f)(-t) \quad \phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx = \mathcal{F}(f)(-t)$$

where  $\mathcal{L}(f), \mathcal{F}(f)$  are the *Laplace* and *Fourier* transforms of  $f$

→ As for e.g. sounds, these transforms can provide alternative descriptions.

**Note:** We focus on the moment generating function  
(see additional slides for the characteristic function)

## Moment Generating Function

### Example

Let  $X \sim \text{Poisson}(\lambda)$ , for  $\lambda > 0$ . Compute  $M_X(t)$ .

## Moment Generating Function

### Example

Let  $X \sim \text{Poisson}(\lambda)$ , for  $\lambda > 0$ . Compute  $M_X(t)$ .

### Solution

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{+\infty} e^{tk} \mathbb{P}(X = k) = \sum_{k=0}^{+\infty} e^{tk} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{(e^t \lambda)^k}{k!} = e^{-\lambda} e^{\lambda e^t}$$

$$M_X(t) = e^{\lambda(e^t - 1)}$$

## Quiz Next Lecture

### Exercise

1. Let  $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(\mu, S)$  with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Compute  $\text{Corr}(X_1, X_2)$ .

2. Let  $X \sim \mathcal{N}(0, 1)$ . Compute  $M_X(t)$
3. Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Compute  $M_X(t)$

## Covariance Matrix\*

### Definition

A **random matrix**  $M \in \mathbb{R}^{p \times n}$  is a matrix whose coefficients  $M_{ij}$  are r.v. defined on the same probability space.

For a random matrix  $M$  we denote

$$\mathbb{E}(M) = (\mathbb{E}(M_{ij}))_{\substack{i=1, \dots, p \\ j=1, \dots, n}} \in \mathbb{R}^{p \times n}$$

### Lemma

For a random matrices  $M$ , and two real matrices  $A, B$  (with appropriate sizes)

$$\mathbb{E}[AM + B] = A\mathbb{E}[M] + B$$

## Covariance Matrix\*

### Definition

A **random matrix**  $M \in \mathbb{R}^{p \times n}$  is a matrix whose coefficients  $M_{ij}$  are r.v. defined on the same probability space.

For a random matrix  $M$  we denote

$$\mathbb{E}(M) = (\mathbb{E}(M_{ij}))_{\substack{i=1, \dots, p \\ j=1, \dots, n}} \in \mathbb{R}^{p \times n}$$

### Lemma

For a random matrices  $M$ , and two real matrices  $A, B$  (with appropriate sizes)

$$\mathbb{E}[AM + B] = A\mathbb{E}[M] + B$$

**Proof** Follows from the linearity of the expectation applied for each coefficient

## Covariance Matrix\*

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector its **covariance matrix** reads

$$S_{\mathbf{X}} = \mathbb{E} [(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$$

## Covariance Matrix\*

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)^\top$  be a random vector its **covariance matrix** reads

$$S_{\mathbf{X}} = \mathbb{E} [(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$$

**Proof** Denote  $\tilde{\mathbf{X}} = \mathbf{X} - \mathbb{E}[\mathbf{X}]$ ,  $\tilde{\mathbf{X}} = (X_1 - \mathbb{E}[X_1], \dots, X_n - \mathbb{E}[X_n])^\top$

Then

$$(\tilde{\mathbf{X}} \tilde{\mathbf{X}}^\top)_{ij} = \tilde{X}_i \tilde{X}_j = (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])$$

So

$$(\mathbb{E}[\tilde{\mathbf{X}} \tilde{\mathbf{X}}^\top])_{ij} = \text{Cov}(X_i, X_j)$$

which gives the result.

## Covariance Matrix\*

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random vector with covariance matrix  $S_{\mathbf{X}}$ .

Let  $A \in \mathbb{R}^{p \times n}$ ,  $b \in \mathbb{R}^p$ , the covariance of the random vector  $\mathbf{Y} = A\mathbf{X} + b \in \mathbb{R}^q$  is

$$S_{\mathbf{Y}} = AS_{\mathbf{X}}A^{\top} \in \mathbb{R}^{p \times p}$$

where  $A^{\top}$  is the transpose of  $A$ , i.e.,  $(A^{\top})_{ij} = A_{ji}$

## Covariance Matrix\*

### Lemma

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random vector with covariance matrix  $S_{\mathbf{X}}$ .

Let  $A \in \mathbb{R}^{p \times n}$ ,  $b \in \mathbb{R}^p$ , the covariance of the random vector  $\mathbf{Y} = A\mathbf{X} + b \in \mathbb{R}^q$  is

$$S_{\mathbf{Y}} = AS_{\mathbf{X}}A^T \in \mathbb{R}^{p \times p}$$

where  $A^T$  is the transpose of  $A$ , i.e.,  $(A^T)_{ij} = A_{ji}$

### Proof

$$\begin{aligned} S_{\mathbf{Y}} &= \mathbb{E}[(A\mathbf{X} + b - (A\mathbb{E}[\mathbf{X}] + b))(A\mathbf{X} + b - (A\mathbb{E}[\mathbf{X}] + b))^T] \\ &= \mathbb{E}[(A(\mathbf{X} - \mathbb{E}[\mathbf{X}])(A(\mathbf{X} - \mathbb{E}[\mathbf{X}]))^T] \\ &= A\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]A^T \\ &= AS_{\mathbf{X}}A^T \end{aligned}$$

## Multivariate Normal Distribution\*

### Definition

A normal random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with **invertible covariance matrix** has a joint p.d.f.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

## Multivariate Normal Distribution\*

### Definition

A normal random vector  $\mathbf{X} = (X_1, \dots, X_n)^\top \sim \mathcal{N}(\mu, S)$  with **invertible covariance matrix** has a joint p.d.f.

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

**Proof** (Sketch for  $\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$  with  $A \in \mathbb{R}^{n \times n}$  invertible)

Generalize the formula for change of random variables for  $n$  dimensions.

The inverse mapping is given by  $Z = A^{-1}(X - b)$

The Jacobian is  $A^{-1}$  the absolute value of its determinant is then

$|\det(A^{-1})| = 1/|\det(A)| = 1/\sqrt{\det(S)}$  where  $S = AA^\top$

## Additional Exercises

### Exercise

Let  $X$  be a discrete r.v. with  $\mathbb{P}(X = -1) = \frac{1}{3}$ ,  $\mathbb{P}(X = 4) = \frac{1}{6}$ ,  $\mathbb{P}(X = 9) = \frac{1}{2}$ .  
What is  $M_X(t)$ ?

## Additional Exercises

### Exercise

Let  $X$  be a discrete r.v. with  $\mathbb{P}(X = -1) = \frac{1}{3}$ ,  $\mathbb{P}(X = 4) = \frac{1}{6}$ ,  $\mathbb{P}(X = 9) = \frac{1}{2}$ . What is  $M_X(t)$ ?

### Solution

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k \in \{-1, 4, 9\}} e^{tk} \mathbb{P}(X = k) = \frac{1}{3}e^{-t} + \frac{1}{6}e^{4t} + \frac{1}{2}e^{9t}$$

## Additional Exercises

### Exercise

Let  $X \sim \mathbb{E}(\lambda)$ ,  $\lambda > 0$ . Compute  $M_X(t)$ .

## Additional Exercises

### Exercise

Let  $X \sim \mathbb{E}(\lambda)$ ,  $\lambda > 0$ . Compute  $M_X(t)$ .

### Solution

$$\mathbb{E}[e^{tX}] = \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} dx = \lim_{b \rightarrow +\infty} \lambda \int_0^b e^{(t-\lambda)x} dx$$

The integral is not necessarily defined, that depends on  $t$ . The proper way to analyze the result is to consider the integral as a limit as written above.

1. if  $t = \lambda$ ,  $\mathbb{E}[e^{tX}] = \lambda \lim_{b \rightarrow +\infty} \int_0^b dx = \lambda \lim_{b \rightarrow +\infty} b = +\infty$
2. if  $t \neq \lambda$ ,

$$\mathbb{E}[e^{tX}] = \lambda \lim_{b \rightarrow +\infty} \frac{e^{(t-\lambda)b} - 1}{t - \lambda} = \begin{cases} +\infty & \text{if } t > \lambda \\ \frac{\lambda}{\lambda - t} & \text{otherwise} \end{cases}$$

So

$$M_X(t) == \begin{cases} +\infty & \text{if } t \geq \lambda \\ \frac{\lambda}{\lambda - t} & \text{otherwise} \end{cases}$$

# Moment Generating Function Concentration Inequalities

Sections 5.1 8.3 9.1

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 18, May 8th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

## Overview

### Feedback:

- ▶ New anonymous form provided, available until **Monday 11th 11:59 pm**
- ▶ In the online context, feedback is essential for the teacher, so please fill it!

### Previous lecture

- ▶ Moment generating function, definition

### This lecture

- ▶ Moments form moment generating function
- ▶ Characterization of a r.v. from its moments
- ▶ Moment of a sum of independent r.v.
- ▶ Concentration inequalities (from moments to proba)

## Quiz Previous Lecture

### Exercise

1. Let  $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(\mu, S)$  with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Compute  $\text{Corr}(X_1, X_2)$ .

2. Let  $X \sim \mathcal{N}(0, 1)$ . Compute  $M_X(t)$
3. Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Compute  $M_X(t)$

## Quiz Previous Lecture

### Exercise

1. Let  $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(\mu, S)$  with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Compute  $\text{Corr}(X_1, X_2)$ .

2. Let  $X \sim \mathcal{N}(0, 1)$ . Compute  $M_X(t)$
3. Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Compute  $M_X(t)$

### Solution

1.  $\text{Corr}(X_1, X_2) = 1$  (Note that  $S$  is not invertible)

## Quiz Previous Lecture

### Exercise

1. Let  $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(\mu, S)$  with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Compute  $\text{Corr}(X_1, X_2)$ .

2. Let  $X \sim \mathcal{N}(0, 1)$ . Compute  $M_X(t)$
3. Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Compute  $M_X(t)$

### Solution

1.  $\text{Corr}(X_1, X_2) = 1$  (Note that  $S$  is not invertible)
- 2.

$$\begin{aligned}\mathbb{E}[e^{tX}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx} e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{t^2/2} e^{-(x-t)^2/2} \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-t)^2/2} = e^{t^2/2}\end{aligned}$$

## Quiz Previous Lecture

### Exercise

1. Let  $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(\mu, S)$  with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Compute  $\text{Corr}(X_1, X_2)$ .

2. Let  $X \sim \mathcal{N}(0, 1)$ . Compute  $M_X(t)$
3. Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ . Compute  $M_X(t)$

### Solution

1.  $\text{Corr}(X_1, X_2) = 1$  (Note that  $S$  is not invertible)
- 2.

$$\begin{aligned}\mathbb{E}[e^{tX}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx} e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{t^2/2} e^{-(x-t)^2/2} \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-t)^2/2} = e^{t^2/2}\end{aligned}$$

3.  $X = \sigma Z + \mu$  for  $Z \sim \mathcal{N}(0, 1)$

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(\sigma Z + \mu)}] = e^{t\mu} \mathbb{E}[e^{t\sigma Z}] = e^{\mu t + \sigma^2 t^2/2}$$

## Moments from Moment Generating Function

**Why is it called moment generating function?**

Let  $X$  be discrete r.v. that takes a finite number of values ( $X(\Omega)$  is finite)

$$M_X(t) = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k),$$

## Moments from Moment Generating Function

**Why is it called moment generating function?**

Let  $X$  be discrete r.v. that takes a finite number of values ( $X(\Omega)$  is finite)

$$M_X(t) = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k), \quad M'_X(t) = \sum_{k \in X(\Omega)} k e^{tk} \mathbb{P}(X = k)$$

## Moments from Moment Generating Function

**Why is it called moment generating function?**

Let  $X$  be discrete r.v. that takes a finite number of values ( $X(\Omega)$  is finite)

$$M_X(t) = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k), \quad M'_X(t) = \sum_{k \in X(\Omega)} k e^{tk} \mathbb{P}(X = k)$$

$$M'_X(0) = \sum_{k \in X(\Omega)} k \mathbb{P}(X = k) = \mathbb{E}[X]$$

## Moments from Moment Generating Function

### Why is it called moment generating function?

Let  $X$  be discrete r.v. that takes a finite number of values ( $X(\Omega)$  is finite)

$$M_X(t) = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k), \quad M'_X(t) = \sum_{k \in X(\Omega)} k e^{tk} \mathbb{P}(X = k)$$

$$M'_X(0) = \sum_{k \in X(\Omega)} k \mathbb{P}(X = k) = \mathbb{E}[X]$$

More generally,  $M'_X(t) = \frac{d}{dt} \mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d}{dt} e^{tX}\right] = \mathbb{E}[X e^{tX}]$ , s.t.  $M'_X(0) = \mathbb{E}[X]$ .

## Moments from Moment Generating Function

### Why is it called moment generating function?

Let  $X$  be discrete r.v. that takes a finite number of values ( $X(\Omega)$  is finite)

$$M_X(t) = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k), \quad M'_X(t) = \sum_{k \in X(\Omega)} ke^{tk} \mathbb{P}(X = k)$$
$$M'_X(0) = \sum_{k \in X(\Omega)} k \mathbb{P}(X = k) = \mathbb{E}[X]$$

More generally,  $M'_X(t) = \frac{d}{dt} \mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d}{dt} e^{tX}\right] = \mathbb{E}[X e^{tX}]$ , s.t.  $M'_X(0) = \mathbb{E}[X]$ .

### Lemma

Let  $X$  be a r.v. If there exists  $\delta > 0$ , s.t. for all  $t \in (-\delta, \delta)$ ,  $M_X(t) < +\infty$ , then for  $n \in \mathbb{N}$ ,  $n > 0$

$$\mathbb{E}[X^n] = M_X^{(n)}(0)$$

i.e.,

if the m.g.f. is finite on an open interval around 0

the non-centered moments of  $X$  are given

by the  $n^{\text{th}}$  derivative of the moment generating function on 0

## Moments from Moment Generating Function

### Example

Let  $X \sim \text{Ber}(p)$  for  $p \in (0, 1)$ . Compute  $\mathbb{E}[X^n]$  for  $n \in \mathbb{N}$ ,  $n > 0$

## Moments from Moment Generating Function

### Example

Let  $X \sim \text{Ber}(p)$  for  $p \in (0, 1)$ . Compute  $\mathbb{E}[X^n]$  for  $n \in \mathbb{N}$ ,  $n > 0$

### Solution

1. (Using previous lemma) We have  $M_X(t) = pe^t + (1 - p)$ , clearly finite on an open interval around 0

Therefore  $M_X^{(n)}(t) = pe^t$  and  $\mathbb{E}[X^n] = M_X^{(n)}(0) = p$ .

2. (More quickly)  $X^n = X$  so  $\mathbb{E}[X^n] = \mathbb{E}[X] = p$

## Moments from Moment Generating Function

### Example

Let  $X \sim \text{Ber}(p)$  for  $p \in (0, 1)$ . Compute  $\mathbb{E}[X^n]$  for  $n \in \mathbb{N}$ ,  $n > 0$

### Solution

1. (Using previous lemma) We have  $M_X(t) = pe^t + (1 - p)$ , clearly finite on an open interval around 0

Therefore  $M_X^{(n)}(t) = pe^t$  and  $\mathbb{E}[X^n] = M_X^{(n)}(0) = p$ .

2. (More quickly)  $X^n = X$  so  $\mathbb{E}[X^n] = \mathbb{E}[X] = p$

### Example

Let  $X \sim \text{Exp}(\lambda)$ ,  $\lambda > 0$ , compute  $\mathbb{E}[X^n]$  for  $n \in \mathbb{N}$ ,  $n > 0$

*Hint:* From the additional exercise of previous lecture,

$$M_X(t) = \begin{cases} \frac{\lambda}{\lambda-t} & \text{if } t < \lambda \\ +\infty & \text{if } t \geq \lambda \end{cases}$$

## Moments from Moment Generating Function

### Example

Let  $X \sim \text{Ber}(p)$  for  $p \in (0, 1)$ . Compute  $\mathbb{E}[X^n]$  for  $n \in \mathbb{N}$ ,  $n > 0$

### Solution

1. (Using previous lemma) We have  $M_X(t) = pe^t + (1 - p)$ , clearly finite on an open interval around 0

Therefore  $M_X^{(n)}(t) = pe^t$  and  $\mathbb{E}[X^n] = M_X^{(n)}(0) = p$ .

2. (More quickly)  $X^n = X$  so  $\mathbb{E}[X^n] = \mathbb{E}[X] = p$

### Example

Let  $X \sim \text{Exp}(\lambda)$ ,  $\lambda > 0$ , compute  $\mathbb{E}[X^n]$  for  $n \in \mathbb{N}$ ,  $n > 0$

*Hint:* From the additional exercise of previous lecture,

$$M_X(t) = \begin{cases} \frac{\lambda}{\lambda-t} & \text{if } t < \lambda \\ +\infty & \text{if } t \geq \lambda \end{cases}$$

**Solution** For  $\lambda > 0$ ,  $M(t)$  is finite on the open interval  $(a, \lambda)$  for any  $a < 0$ , i.e. an open interval around 0. We can compute for  $t < \lambda$

$$M'_X(t) = \lambda(\lambda - t)^{-2}, M''_X(t) = 2\lambda(\lambda - t)^3, \dots, M_X^{(n)}(t) = n!\lambda(\lambda - t)^{-n-1}$$

$$\text{So } \mathbb{E}[X^n] = M_X^{(n)}(0) = n!\lambda^{-n}$$

## Characterization of r.v. by Moment Generating Functions

### Definition (Equality in distribution (Reminder))

Two r.v.  $X, Y$  are **equal in distribution**, denoted  $X \stackrel{d}{=} Y$  if

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \quad \text{for any } B \subset \mathbb{R}$$

## Characterization of r.v. by Moment Generating Functions

### Definition (Equality in distribution (Reminder))

Two r.v.  $X, Y$  are **equal in distribution**, denoted  $X \stackrel{d}{=} Y$  if

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \quad \text{for any } B \subset \mathbb{R}$$

### Theorem

Let  $X, Y$  be two r.v. If there exists  $\delta > 0$  such that for all  $t \in (-\delta, \delta)$   $M_X(t)$  and  $M_Y(t)$  are finite and  $M_X(t) = M_Y(t)$  then  $X \stackrel{d}{=} Y$ ,  
i.e.

if the moment generating functions of  $X, Y$  are finite on an open interval around 0  
and that they coincide on this interval  
then  $X, Y$  have the same distribution

## Characterization of r.v. by Moment Generating Functions

### Definition (Equality in distribution (Reminder))

Two r.v.  $X, Y$  are **equal in distribution**, denoted  $X \stackrel{d}{=} Y$  if

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \quad \text{for any } B \subset \mathbb{R}$$

### Theorem

Let  $X, Y$  be two r.v. If there exists  $\delta > 0$  such that for all  $t \in (-\delta, \delta)$   $M_X(t)$  and  $M_Y(t)$  are finite and  $M_X(t) = M_Y(t)$  then  $X \stackrel{d}{=} Y$ ,  
i.e.

if the moment generating functions of  $X, Y$  are finite on an open interval around 0  
and that they coincide on this interval  
then  $X, Y$  have the same distribution

### Theoretical intuition\*

If  $X$  is continuous, then  $M_X$  is the Laplace transform of  $f_X$ ,

The Laplace transform is injective: if  $f, g$  have same Laplace transform,  $f = g$

Here if  $X, Y$  are continuous then  $M_X = M_Y$  imply  $f_X = f_Y$ , so  $X \stackrel{d}{=} Y$

## Characterization of r.v. by Moment Generating Functions

### Example

Let  $X$  be a r.v. s.t.  $M_X(t) = \frac{1}{5}e^{-17t} + \frac{1}{4} + \frac{11}{20}e^{2t}$ .  
What is the distribution of  $X$ ?

# Characterization of r.v. by Moment Generating Functions

## Example

Let  $X$  be a r.v. s.t.  $M_X(t) = \frac{1}{5}e^{-17t} + \frac{1}{4} + \frac{11}{20}e^{2t}$ .

What is the distribution of  $X$ ?

## Solution

**Intuition** The moment generating function for a discrete r.v. reads

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k)$$

So here we recognize  $\mathbb{P}(X = -17) = \frac{1}{5}$ ,  $\mathbb{P}(X = 0) = \frac{1}{4}$ ,  $\mathbb{P}(X = 2) = 11/20$ .

# Characterization of r.v. by Moment Generating Functions

## Example

Let  $X$  be a r.v. s.t.  $M_X(t) = \frac{1}{5}e^{-17t} + \frac{1}{4} + \frac{11}{20}e^{2t}$ .

What is the distribution of  $X$ ?

## Solution

**Intuition** The moment generating function for a discrete r.v. reads

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k \in X(\Omega)} e^{tk} \mathbb{P}(X = k)$$

So here we recognize  $\mathbb{P}(X = -17) = \frac{1}{5}$ ,  $\mathbb{P}(X = 0) = \frac{1}{4}$ ,  $\mathbb{P}(X = 2) = 11/20$ .

**Formally** Let  $Y$  be a r.v. s.t.  $\mathbb{P}(Y = -17) = \frac{1}{5}$ ,  $\mathbb{P}(Y = 0) = \frac{1}{4}$ ,  $\mathbb{P}(Y = 2) = 11/20$ , then for any  $t \in \mathbb{R}$ ,

$$M_Y(t) = M_X(t)$$

Therefore  $X \stackrel{d}{=} Y$ , i.e.  $X$  has the same distribution as  $Y$ .

# Moment of a Sum of Independent Random Variables

## Motivation

- ▶ The moment generating function could be very useful
- ▶ As for expectation, variance, etc... isn't there a quicker way to compute m.g.f.?

# Moment of a Sum of Independent Random Variables

## Motivation

- ▶ The moment generating function could be very useful
- ▶ As for expectation, variance, etc... isn't there a quicker way to compute m.g.f.?

## Lemma

Let  $X_1, \dots, X_n$  be independent r.v. then for any  $t \in \mathbb{R}$ ,

$$M_{X_1+\dots+X_n}(t) = M_{X_1}(t) \dots M_{X_n}(t)$$

# Moment of a Sum of Independent Random Variables

## Motivation

- ▶ The moment generating function could be very useful
- ▶ As for expectation, variance, etc... isn't there a quicker way to compute m.g.f.?

## Lemma

Let  $X_1, \dots, X_n$  be independent r.v. then for any  $t \in \mathbb{R}$ ,

$$M_{X_1+\dots+X_n}(t) = M_{X_1}(t) \dots M_{X_n}(t)$$

## Proof

$$M_{X_1+\dots+X_n}(t) = \mathbb{E}[e^{t(X_1+\dots+X_n)}] = \mathbb{E}[e^{tX_1} \dots e^{tX_n}] = \mathbb{E}[e^{tX_1}] \dots \mathbb{E}[e^{tX_n}] = M_{X_1}(t) \dots M_{X_n}(t)$$

## Moment of a Sum of Independent Random Variables

### Example

Let  $X \sim \text{Poisson}(\lambda)$ ,  $Y \sim \text{Poisson}(\mu)$  independent, (recall that  $M_X(t) = e^{\lambda(e^t - 1)}$ )

1. Compute  $M_{X+Y}(t)$
2. What can you conclude about the distribution of  $X + Y$ ?

## Moment of a Sum of Independent Random Variables

### Example

Let  $X \sim \text{Poisson}(\lambda)$ ,  $Y \sim \text{Poisson}(\mu)$  independent, (recall that  $M_X(t) = e^{\lambda(e^t - 1)}$ )

1. Compute  $M_{X+Y}(t)$
2. What can you conclude about the distribution of  $X + Y$ ?

### Solution

1.  $M_{X+Y}(t) = M_X(t)M_Y(t) = e^{(\lambda+\mu)(e^t - 1)}$

2. Let  $Z \sim \text{Poisson}(\lambda + \mu)$  s.t.  $M_Z(t) = e^{(\lambda+\mu)(e^t - 1)}$  so  $X + Y \stackrel{d}{=} Z \sim \text{Poisson}(\lambda + \mu)$

## Moment of a Sum of Independent Random Variables

### Example

Let  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ ,  $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$  independent (recall that  $M_X(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}}$ )

1. Compute  $M_{X+Y}(t)$
2. What can you conclude about the distribution of  $X + Y$ ?

## Moment of a Sum of Independent Random Variables

### Example

Let  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ ,  $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$  independent (recall that  $M_X(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}}$ )

1. Compute  $M_{X+Y}(t)$
2. What can you conclude about the distribution of  $X + Y$ ?

### Solution

1.  $M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}} e^{\mu_2 t + \frac{\sigma_2^2 t^2}} = e^{(\mu_1 + \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}}$
2.  $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

## Concentration Inequalities

### Motivation:

1. From the moment generating function, we know a probability distribution
2. What if we only have access to some of the moments?
3. Can we say something about the probability distribution?

## Concentration Inequalities

### Theorem (Monotonicity of Expectation)

If two r.v.  $X, Y$  defined on the same proba. space  $(\Omega, \mathcal{F}, \mathbb{P})$  have finite means and satisfy that  $\mathbb{P}(X \leq Y) = 1$  then  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ .

## Concentration Inequalities

### Theorem (Monotonicity of Expectation)

If two r.v.  $X, Y$  defined on the same proba. space  $(\Omega, \mathcal{F}, \mathbb{P})$  have finite means and satisfy that  $\mathbb{P}(X \leq Y) = 1$  then  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ .

**Proof** Denote  $Z = Y - X$  s.t.  $\mathbb{P}(Z \geq 0) = 1$

1. (Discrete case) If  $Z$  is discrete, for any  $k < 0$ ,  $0 \leq \mathbb{P}(Z = k) \leq \mathbb{P}(Z < 0) = 0$  so

$$\mathbb{E}[Z] = \sum_{k \in Z(\Omega)} k \mathbb{P}(Z = k) \geq 0$$

## Concentration Inequalities

### Theorem (Monotonicity of Expectation)

If two r.v.  $X, Y$  defined on the same proba. space  $(\Omega, \mathcal{F}, \mathbb{P})$  have finite means and satisfy that  $\mathbb{P}(X \leq Y) = 1$  then  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ .

**Proof** Denote  $Z = Y - X$  s.t.  $\mathbb{P}(Z \geq 0) = 1$

1. (Discrete case) If  $Z$  is discrete, for any  $k < 0$ ,  $0 \leq \mathbb{P}(Z = k) \leq \mathbb{P}(Z < 0) = 0$  so

$$\mathbb{E}[Z] = \sum_{k \in Z(\Omega)} k \mathbb{P}(Z = k) \geq 0$$

2. (Continuous case) If  $Z$  is continuous, then (as in exercise 4.2 of homework 1)

$$\int_{-\infty}^0 z f_Z(z) dz = - \int_{-\infty}^0 \int_z^0 f_Z(z) dt dz = - \iint_{z \leq t \leq 0, z \leq 0} f_Z(z) dt dz = - \int_{-\infty}^0 \int_{-\infty}^t f_Z(z) dz dt$$

# Concentration Inequalities

## Theorem (Monotonicity of Expectation)

If two r.v.  $X, Y$  defined on the same proba. space  $(\Omega, \mathcal{F}, \mathbb{P})$  have finite means and satisfy that  $\mathbb{P}(X \leq Y) = 1$  then  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ .

**Proof** Denote  $Z = Y - X$  s.t.  $\mathbb{P}(Z \geq 0) = 1$

1. (Discrete case) If  $Z$  is discrete, for any  $k < 0$ ,  $0 \leq \mathbb{P}(Z = k) \leq \mathbb{P}(Z < 0) = 0$  so

$$\mathbb{E}[Z] = \sum_{k \in Z(\Omega)} k \mathbb{P}(Z = k) \geq 0$$

2. (Continuous case) If  $Z$  is continuous, then (as in exercise 4.2 of homework 1)

$$\int_{-\infty}^0 z f_Z(z) dz = - \int_{-\infty}^0 \int_z^0 f_Z(z) dt dz = - \int_{z \leq t \leq 0, z \leq 0} \int f_Z(z) dt dz = - \int_{-\infty}^0 \int_{-\infty}^t f_Z(z) dz dt$$

So  $\int_{-\infty}^0 z f_Z(z) dz = - \int_{-\infty}^0 \mathbb{P}(Z \leq t) = 0$  since  $0 \leq \mathbb{P}(Z \leq t) \leq \mathbb{P}(Z \leq 0) = 0$  for all  $t \leq 0$ .

$$\text{Therefore } \mathbb{E}[Z] = \int_{-\infty}^0 z f_Z(z) dz + \int_0^{+\infty} z f_Z(z) dz \geq 0$$

3. So in both cases  $\mathbb{E}[Z] = \mathbb{E}[Y - X] \geq 0$ , i.e.  $\mathbb{E}[X] \leq \mathbb{E}[Y]$

## Markov Inequality

**Question:** What can be said about the proba. of  $X$  if we know  $\mathbb{E}[X]$ ?

Theorem (Markov inequality)

*Let  $X$  be a non-negative r.v. with finite mean then for any  $c > 0$ ,*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}$$

## Markov Inequality

**Question:** What can be said about the proba. of  $X$  if we know  $\mathbb{E}[X]$ ?

**Theorem (Markov inequality)**

Let  $X$  be a non-negative r.v. with finite mean then for any  $c > 0$ ,

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}$$

**Proof** Define the indicator random variable  $I_{X \geq c}$ . We have

$$X \geq c I_{X \geq c}$$

1. when  $X \geq c$  the inequality reads  $X \geq c$ ,
2. when  $X \leq c$  the inequality reads  $X \geq 0$ , true by assumption

Now applying previous theorem,

$$\mathbb{E}[X] \geq c \mathbb{E}[I_{X \geq c}] = c \mathbb{P}(X \geq c)$$

## Quiz Next Lecture

### Exercise

*A donut vendor sells on average 1000 donuts per day.*

*Could he sell more than 1400 donuts tomorrow with proba. greater than 0.8?*

### Exercise

*Let  $X \sim \text{Ber}(p)$ ,  $p \in (0, 1)$*

- What is  $\mathbb{P}(X \geq 0.01)$ ?*
- What gives Markov inequality?*

## Additional Exercises

### Exercise

Assume that  $X$  has moment generating function

$$M_X(t) = \frac{1}{2} + \frac{1}{3}e^{-4t} + \frac{1}{6}e^{5t}$$

1. Compute  $\mathbb{E}[X]$ ,  $\text{Var}(X)$  from  $M_X(t)$
2. Find the p.m.f. of  $X$  and use it to check your computations of  $\mathbb{E}[X]$ ,  $\text{Var}(X)$

## Additional Exercises

### Exercise

Assume that  $X$  has moment generating function

$$M_X(t) = \frac{1}{2} + \frac{1}{3}e^{-4t} + \frac{1}{6}e^{5t}$$

1. Compute  $\mathbb{E}[X]$ ,  $\text{Var}(X)$  from  $M_X(t)$
2. Find the p.m.f. of  $X$  and use it to check your computations of  $\mathbb{E}[X]$ ,  $\text{Var}(X)$

### Solution

1.  $M'_X(t) = \frac{-4}{3}e^{-4t} + \frac{5}{6}e^{5t}$ ,  $M''_X(t) = \frac{16}{3}e^{-4t} + \frac{25}{6}e^{5t}$  so  
 $\mathbb{E}[X] = M'(0) = \frac{-4}{3} + \frac{5}{6} = -0.5$ ,  $\mathbb{E}[X^2] = \frac{16}{3} + \frac{25}{6} = 9.5$ ,  
 $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 9.25$
2.  $\mathbb{P}(X = 0) = \frac{1}{2}$ ,  $\mathbb{P}(X = -4) = \frac{1}{3}$ ,  $\mathbb{P}(X = 5) = \frac{1}{6}$ , so  $\mathbb{E}[X] = \frac{-4}{3} + \frac{5}{6} = -0.5$ ,  
 $\mathbb{E}[X^2] = 4^2 \cdot \frac{1}{3} + 5^2 \cdot \frac{1}{6} = 9.5$  so yes we got the same result.

## Additional exercises

### Exercise

In the following use the information given to determine the distribution of the r.v., or show that the information is not sufficient by describing at least two different r.v. with different distributions that satisfy the given condition.

1.  $X$  r.v. s.t.  $M_X(t) = e^{6t^2}$  for  $|t| < 2$
2.  $Y$  r.v. s.t.  $M_Y(t) = \frac{2}{2-t}$  for  $t < 0.5$
3.  $X$  r.v. s.t.  $M_Z(t) = +\infty$  for  $t \geq 5$
4.  $W$  r.v. s.t.  $M_W(2) = 2$

Note: each time the information on  $t$  is not about the domain of definition of the m.g.f., it is just some range of  $t$  for which we know the value

## Additional exercises

### Exercise

In the following use the information given to determine the distribution of the r.v., or show that the information is not sufficient by describing at least two different r.v. with different distributions that satisfy the given condition.

1.  $X$  r.v. s.t.  $M_X(t) = e^{6t^2}$  for  $|t| < 2$
2.  $Y$  r.v. s.t.  $M_Y(t) = \frac{2}{2-t}$  for  $t < 0.5$
3.  $X$  r.v. s.t.  $M_Z(t) = +\infty$  for  $t \geq 5$
4.  $W$  r.v. s.t.  $M_W(2) = 2$

Note: each time the information on  $t$  is not about the domain of definition of the m.g.f., it is just some range of  $t$  for which we know the value

### Solution

1. Yes cause we have the information on an open interval around 0.  
Precisely we recognize  $X \sim \mathcal{N}(0, 12)$  (see solution of the previous quiz above)
2. Yes cause we have the information on an open interval around 0.  
Precisely we recognize  $Y \sim \text{Exp}(2)$  (see additional exercise last lecture)
3. No, we miss the information on an open interval around 0.  
For example,  $Z_1 \sim \text{Exp}(3)$  and  $Z_2 \sim \text{Exp}(1)$  both satisfy  $M_Z(t) = +\infty$  for  $t \geq 5$ ,  
yet  $Z_1 \neq Z_2$
4. No, we miss the information on an open interval around 0.  
For example  $W_1 \sim \text{Exp}(4)$  and  $W_2 \sim \mathcal{N}(0, 2/\log(2))$  both satisfy  $M_W(2) = 2$

## Additional exercises

### Exercise

Let  $Z$  be s.t.  $M_Z(t) = (\frac{1}{2}e^{-t} + \frac{2}{5} + \frac{1}{10}e^{t/2})^{36}$  Express  $Z$  as a sum of independent r.v. with precised distribution

## Additional exercises

### Exercise

Let  $Z$  be s.t.  $M_Z(t) = (\frac{1}{2}e^{-t} + \frac{2}{5} + \frac{1}{10}e^{t/2})^{36}$  Express  $Z$  as a sum of independent r.v. with precised distribution

**Solution** Define  $X$  s.t  $\mathbb{P}(X = -1) = \frac{1}{2}, \mathbb{P}(X = 0) = \frac{2}{5}, \mathbb{P}(X = 1/2) = \frac{1}{10}$ .

Let  $X_1, \dots, X_{36}$  be i.i.d. r.v. with same distribution as  $X$  then

$$M_{X_1+\dots+X_{36}}(t) = M_Z(t) \text{ so } X_1 + \dots + X_{36} \stackrel{d}{=} Z.$$

# Convergence of Random Variables, Law of Large Numbers, Central Limit Theorem

Sections 9.2 9.3

STAT/MATH 395 Spring 2020

Vincent Roulet

Lecture 19, May 11th, 2020

Ask questions via [chat on Zoom](#)

Answer quiz via [PollEverywhere](#) (username: vincentroulet)

# Overview

## Feedback:

- ▶ Form available until **tonight 11:59 pm**
- ▶ In the online context, feedback is essential for the teacher, so please fill it!

## 2nd exam content

- ▶ Everything since the 1st exam, i.e.,
  - ▶ exchangeability, i.i.d. random variables
  - ▶ covariance, variance of a sum,
  - ▶ moment generating function,
  - ▶ concentration inequalities
- ▶ **main tool:** decompose r.v. in sums and use key lemmas: linearity of expectation, covariance of a sum, expectation of a product of independent r.v.
- ▶ No convergence theorems will be asked in the exam
- ▶ Take home exam (more time to give you more flexibility)
  - ▶ Available on Friday 8:00 am
  - ▶ Due Saturday 11:59 pm

## 2nd exam review

- ▶ **Fill the google form** to know what you want to be covered in particular for the exam

## Overview

**Lecture notes:** available, cover everything to know since the first exam

### Previous lecture

- ▶ Concentration inequalities

### This lecture

- ▶ Different convergences of r.v.
- ▶ Law of large numbers
- ▶ Central limit theorem

## Quiz Previous Lecture

### Exercise

*A donut vendor sells on average 1000 donuts per day.*

*Could he sell more than 1400 donuts tomorrow with proba. greater than 0.8?*

## Quiz Previous Lecture

### Exercise

A donut vendor sells on average 1000 donuts per day.

Could he sell more than 1400 donuts tomorrow with proba. greater than 0.8?

**Solution** Denote  $X$  the number of donuts sold per day. Clearly  $X$  is non-negative.

$$\mathbb{P}(X \geq 1400) \leq \frac{\mathbb{E}[X]}{1400} = \frac{1000}{1400} = 5/7 \approx 0.71 < 0.8 \quad \rightarrow \text{so the answer is no}$$

## Quiz Previous Lecture

### Exercise

A donut vendor sells on average 1000 donuts per day.

Could he sell more than 1400 donuts tomorrow with proba. greater than 0.8?

**Solution** Denote  $X$  the number of donuts sold per day. Clearly  $X$  is non-negative.

$$\mathbb{P}(X \geq 1400) \leq \frac{\mathbb{E}[X]}{1400} = \frac{1000}{1400} = 5/7 \approx 0.71 < 0.8 \quad \rightarrow \text{so the answer is no}$$

### Exercise

Let  $X \sim \text{Ber}(p)$ ,  $p \in (0, 1)$

1. What is  $\mathbb{P}(X \geq 0.01)$ ?
2. What gives Markov inequality?

## Quiz Previous Lecture

### Exercise

A donut vendor sells on average 1000 donuts per day.

Could he sell more than 1400 donuts tomorrow with proba. greater than 0.8?

**Solution** Denote  $X$  the number of donuts sold per day. Clearly  $X$  is non-negative.

$$\mathbb{P}(X \geq 1400) \leq \frac{\mathbb{E}[X]}{1400} = \frac{1000}{1400} = 5/7 \approx 0.71 < 0.8 \quad \rightarrow \text{so the answer is no}$$

### Exercise

Let  $X \sim \text{Ber}(p)$ ,  $p \in (0, 1)$

1. What is  $\mathbb{P}(X \geq 0.01)$ ?
2. What gives Markov inequality?

### Solution

1. Clearly  $\mathbb{P}(X \geq 0.01) = \mathbb{P}(X = 1) = p$

2. Markov's inequality gives

$$\mathbb{P}(X \geq 0.01) \leq \frac{\mathbb{E}[X]}{0.01} = 100p$$

Here Markov's inequality is useless (we may even have  $100p \geq 1$  s.t. it is even less informative than knowing that  $\mathbb{P}(X \geq 0.01) \leq 1$ )

## Chebyshev's inequality

**Question:** What can be said about the proba. of  $X$  if we know  $\mathbb{E}[X]$  and  $\text{Var}(X)$ ?

### Theorem (Chebyshev's Inequality)

*Let  $X$  be a r.v. with finite mean  $\mu$  and finite variance  $\sigma^2$ , then for any  $c > 0$ ,*

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

## Chebyshev's inequality

**Question:** What can be said about the proba. of  $X$  if we know  $\mathbb{E}[X]$  and  $\text{Var}(X)$ ?

### Theorem (Chebyshev's Inequality)

Let  $X$  be a r.v. with finite mean  $\mu$  and finite variance  $\sigma^2$ , then for any  $c > 0$ ,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

**Proof** Define  $Z = (X - \mu)^2$ ,  $Z$  is non-negative, has finite mean (since  $X$  has finite variance)  
Using Markov's inequality on  $Z$  we get

$$\mathbb{P}(|X - \mu| \geq c) = \mathbb{P}(Z \geq c^2) \leq \frac{\mathbb{E}[Z]}{c^2} = \frac{\mathbb{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

## Chebyshev's inequality

**Question:** What can be said about the proba. of  $X$  if we know  $\mathbb{E}[X]$  and  $\text{Var}(X)$ ?

### Theorem (Chebyshev's Inequality)

Let  $X$  be a r.v. with finite mean  $\mu$  and finite variance  $\sigma^2$ , then for any  $c > 0$ ,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

**Proof** Define  $Z = (X - \mu)^2$ ,  $Z$  is non-negative, has finite mean (since  $X$  has finite variance)  
Using Markov's inequality on  $Z$  we get

$$\mathbb{P}(|X - \mu| \geq c) = \mathbb{P}(Z \geq c^2) \leq \frac{\mathbb{E}[Z]}{c^2} = \frac{\mathbb{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

### Note:

The event  $\{|X - \mu| \geq c\}$  contains the events  $\{X \geq \mu + c\}$  and  $\{X \leq \mu - c\}$   
So we naturally have a bound on  $\mathbb{P}(X \geq \mu + c)$ ,  $\mathbb{P}(X \leq \mu - c)$

## Chebyshev's inequality

### Example

A donut vendor sells on average 1000 donuts per day with a standard deviation of  $\sqrt{200}$ . Provide a bound on

1. the proba. that there will be between 950 and 1050 donuts sold tomorrow
2. the proba. that there will be at least 1400 donuts sold tomorrow

## Chebyshev's inequality

### Example

A donut vendor sells on average 1000 donuts per day with a standard deviation of  $\sqrt{200}$ . Provide a bound on

1. the proba. that there will be between 950 and 1050 donuts sold tomorrow
2. the proba. that there will be at least 1400 donuts sold tomorrow

### Solution

1.  $\mathbb{P}(950 < X < 1050) = \mathbb{P}(|X - 1000| < 50) = 1 - \mathbb{P}(|X - 1000| \geq 50)$

By Chebyshev's inequality,

$$\mathbb{P}(|X - 1000| \geq 50) = \mathbb{P}(|X - \mathbb{E}[X]| \geq 50) \leq \frac{\text{Var}(X)}{50^2} = \frac{200}{50^2} = \frac{2}{25} = 0.08$$

So  $\mathbb{P}(950 < X < 1050) \geq 1 - 0.08 = 0.92$

2.  $\mathbb{P}(X \geq 1400) = \mathbb{P}(X - 1000 \geq 400) \leq \frac{200}{400^2} = \frac{1}{800} = 0.00125$

## Law of Large Numbers

### Motivation

- ▶ Consider having access to a r.v.  $X$  (e.g. flip of biased coin) only through observations: first flip you get  $x_1 = \text{'heads'}$ , second flip you get  $x_2 = \text{'heads'}$  and so on...
- ▶ What can you say about e.g. the moments of  $X$  from these observations?  
(That would get us some info about the proba using concentration inequalities)

# Law of Large Numbers

## Motivation

- ▶ Consider having access to a r.v.  $X$  (e.g. flip of biased coin) only through observations: first flip you get  $x_1 = \text{'heads'}$ , second flip you get  $x_2 = \text{'heads'}$  and so on...
- ▶ What can you say about e.g. the moments of  $X$  from these observations? (That would get us some info about the proba using concentration inequalities)
- ▶ Treat these observations as a sequence of i.i.d. r.v.  $X_1, X_2, \dots$

# Law of Large Numbers

## Motivation

- ▶ Consider having access to a r.v.  $X$  (e.g. flip of biased coin) only through observations: first flip you get  $x_1 = \text{'heads'}$ , second flip you get  $x_2 = \text{'heads'}$  and so on...
- ▶ What can you say about e.g. the moments of  $X$  from these observations? (That would get us some info about the proba using concentration inequalities)
- ▶ Treat these observations as a sequence of i.i.d. r.v.  $X_1, X_2, \dots$
- ▶ To get e.g. the mean consider the sample mean of the first  $n$  observations

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

# Law of Large Numbers

## Motivation

- ▶ Consider having access to a r.v.  $X$  (e.g. flip of biased coin) only through observations: first flip you get  $x_1 = \text{'heads'}$ , second flip you get  $x_2 = \text{'heads'}$  and so on...
- ▶ What can you say about e.g. the moments of  $X$  from these observations? (That would get us some info about the proba using concentration inequalities)
- ▶ Treat these observations as a sequence of i.i.d. r.v.  $X_1, X_2, \dots$
- ▶ To get e.g. the mean consider the sample mean of the first  $n$  observations

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

How the empirical mean  $\bar{X}_n$  approaches the mean  $\mathbb{E}(X)$  as  $n \rightarrow +\infty$ ?

# Law of Large Numbers

## Motivation

- ▶ Consider having access to a r.v.  $X$  (e.g. flip of biased coin) only through observations: first flip you get  $x_1 = \text{'heads'}$ , second flip you get  $x_2 = \text{'heads'}$  and so on...
- ▶ What can you say about e.g. the moments of  $X$  from these observations? (That would get us some info about the proba using concentration inequalities)
- ▶ Treat these observations as a sequence of i.i.d. r.v.  $X_1, X_2, \dots$
- ▶ To get e.g. the mean consider the sample mean of the first  $n$  observations

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

How the empirical mean  $\bar{X}_n$  approaches the mean  $\mathbb{E}(X)$  as  $n \rightarrow +\infty$ ?

## Problems

1. In what sense is the convergence of r.v. is defined?

# Law of Large Numbers

## Motivation

- ▶ Consider having access to a r.v.  $X$  (e.g. flip of biased coin) only through observations: first flip you get  $x_1 = \text{'heads'}$ , second flip you get  $x_2 = \text{'heads'}$  and so on...
- ▶ What can you say about e.g. the moments of  $X$  from these observations? (That would get us some info about the proba using concentration inequalities)
- ▶ Treat these observations as a sequence of i.i.d. r.v.  $X_1, X_2, \dots$
- ▶ To get e.g. the mean consider the sample mean of the first  $n$  observations

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

How the empirical mean  $\bar{X}_n$  approaches the mean  $\mathbb{E}(X)$  as  $n \rightarrow +\infty$ ?

## Problems

1. In what sense is the convergence of r.v. is defined?
2. Good to know the limit, but how does it behave asymptotically, i.e., how  $\bar{X}_n - \mu$  behaves as  $n \rightarrow +\infty$ ?

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?  
In that case, **how do we measure the convergence of the mappings?**

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?  
In that case, **how do we measure the convergence of the mappings?**
  - 1.1 Do we require that  $X_n(\omega)$  converges to  $X_*(\omega)$  almost surely?

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?  
In that case, **how do we measure the convergence of the mappings?**
  - 1.1 Do we require that  $X_n(\omega)$  converges to  $X_*(\omega)$  almost surely?  
 $\rightarrow$  convergence almost surely

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?  
In that case, **how do we measure the convergence of the mappings?**
  - 1.1 Do we require that  $X_n(\omega)$  converges to  $X_*(\omega)$  almost surely?  
 $\rightarrow$  convergence almost surely
  - 1.2 Or do we simply require that the proba. that  $X_n(\omega)$  differs from  $X_*(\omega)$  becomes smaller and smaller?

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?  
In that case, **how do we measure the convergence of the mappings?**
  - 1.1 Do we require that  $X_n(\omega)$  converges to  $X_*(\omega)$  almost surely?  
 $\rightarrow$  **convergence almost surely**
  - 1.2 Or do we simply require that the proba. that  $X_n(\omega)$  differs from  $X_*(\omega)$  becomes smaller and smaller?  $\rightarrow$  **convergence in probability**
2. Or do we only care about the measures  $\mathbb{P}_{X_n}$  and  $\mathbb{P}_{X_*}$  that  $X_n$  and  $X_*$  define on  $\mathbb{R}$ ?

# Convergence of Random Variables

## Convergence definition

1. Consider a sequence of r.v.  $(X_n)_{n=1}^{+\infty} = (X_1, X_2, \dots)$  defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$
2. How can we define a notion of convergence to a r.v.  $X_*$ ?

## What do we care about?

1. Do we care about the actual mappings  $X_n : \omega \rightarrow X_n(\omega)$ ?  
In that case, **how do we measure the convergence of the mappings?**
  - 1.1 Do we require that  $X_n(\omega)$  converges to  $X_*(\omega)$  almost surely?  
 $\rightarrow$  **convergence almost surely**
  - 1.2 Or do we simply require that the proba. that  $X_n(\omega)$  differs from  $X_*(\omega)$  becomes smaller and smaller?  $\rightarrow$  **convergence in probability**
2. Or do we only care about the measures  $\mathbb{P}_{X_n}$  and  $\mathbb{P}_{X_*}$  that  $X_n$  and  $X_*$  define on  $\mathbb{R}$ ?  $\rightarrow$  **convergence in distribution**

# Convergence of Random Variables



## Convergence almost surely

The events for which  $(X_n)_{n=1}^{+\infty}$  does not converge to  $X_*$  have probability 0

## Convergence in probability

The probability that  $X_n$  differs from  $X_*$  by any  $\varepsilon$  tends to 0 as  $n \rightarrow +\infty$

## Convergence in distribution

The probability measures  $\mathbb{P}_{X_n}$  tend to  $\mathbb{P}_{X_*}$  as  $n \rightarrow +\infty$

## Almost-sure convergence

### Definition

A sequence of r.v.  $(X_n)_{n=1}^{+\infty}$  converges **almost surely** to a r.v.  $X_*$ , all defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  if

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} X_n = X_*\right) = 1$$

## Almost-sure convergence

### Definition

A sequence of r.v.  $(X_n)_{n=1}^{+\infty}$  converges **almost surely** to a r.v.  $X_*$ , all defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  if

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} X_n = X_*\right) = 1$$

### Interpretation

$(X_n)_{n=1}^{+\infty}$  converges **almost surely** to  $X_*$   
if the events for which  $X_n$  does not converge to  $X_*$  have probability 0

## Almost-sure convergence

### Definition

A sequence of r.v.  $(X_n)_{n=1}^{+\infty}$  converges **almost surely** to a r.v.  $X_*$ , all defined on the same probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  if

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} X_n = X_*\right) = 1$$

### Interpretation

$(X_n)_{n=1}^{+\infty}$  converges **almost surely** to  $X_*$   
if the events for which  $X_n$  does not converge to  $X_*$  have probability 0

### Complete mathematical definition

$(X_n)_{n=1}^{+\infty}$  converges **almost surely** to  $X_*$  if

$$\{\omega : \forall \varepsilon > 0, \exists n_0 \in \mathbb{N}, \text{s.t. } \forall n \geq n_0, |X_n(\omega) - X_*(\omega)| \leq \varepsilon\} \in \mathcal{F}$$

is an event with probability 1

## Strong Law of Large Numbers

Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

## Strong Law of Large Numbers

Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

Note:

1. Convergence almost-surely is the strongest type of convergence,  
→ with this lemma we are done: all other types of convergence apply.

## Strong Law of Large Numbers

### Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

### Note:

1. Convergence almost-surely is the strongest type of convergence,  
→ with this lemma we are done: all other types of convergence apply.
2. Yet, the proof is complex.

## Strong Law of Large Numbers

### Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

### Note:

1. Convergence almost-surely is the strongest type of convergence,  
→ with this lemma we are done: all other types of convergence apply.
2. Yet, the proof is complex.
3. Convergence in probability, although weaker is easier to prove

No Quiz

## Generalizations of Markov's inequality\*

**Question:** What if we know more moments? How can proba of  $X$  be bounded?

### Lemma

Let  $X$  be a r.v. and  $f$  be positive and strictly increasing s.t.  $\mathbb{E}[f(X)]$  is finite.

$$\mathbb{P}(X \geq c) = \mathbb{P}(f(X) \geq f(c)) \leq \frac{\mathbb{E}[f(X)]}{f(c)}$$

## Generalizations of Markov's inequality\*

**Question:** What if we know more moments? How can proba of  $X$  be bounded?

**Lemma**

Let  $X$  be a r.v. and  $f$  be positive and strictly increasing s.t.  $\mathbb{E}[f(X)]$  is finite.

$$\mathbb{P}(X \geq c) = \mathbb{P}(f(X) \geq f(c)) \leq \frac{\mathbb{E}[f(X)]}{f(c)}$$

**Proof** First equality comes from  $f$  strictly increasing, second inequality is Markov.

## Generalizations of Markov's inequality\*

**Question:** What if we know more moments? How can proba of  $X$  be bounded?

### Lemma

Let  $X$  be a r.v. and  $f$  be positive and strictly increasing s.t.  $\mathbb{E}[f(X)]$  is finite.

$$\mathbb{P}(X \geq c) = \mathbb{P}(f(X) \geq f(c)) \leq \frac{\mathbb{E}[f(X)]}{f(c)}$$

**Proof** First equality comes from  $f$  strictly increasing, second inequality is Markov.

### Corollary (Chernoff's bound)

Let  $X$  be a r.v. s.t.  $M_X(t) = \mathbb{E}[e^{tX}]$  is finite for  $t \in (0, \theta]$ , then

$$\mathbb{P}(X \geq c) \leq e^{-tc} \mathbb{E}[e^{tX}] \quad \text{for all } t \in (0, \theta]$$

## Generalizations of Markov's inequality\*

**Question:** What if we know more moments? How can proba of  $X$  be bounded?

### Lemma

Let  $X$  be a r.v. and  $f$  be positive and strictly increasing s.t.  $\mathbb{E}[f(X)]$  is finite.

$$\mathbb{P}(X \geq c) = \mathbb{P}(f(X) \geq f(c)) \leq \frac{\mathbb{E}[f(X)]}{f(c)}$$

**Proof** First equality comes from  $f$  strictly increasing, second inequality is Markov.

### Corollary (Chernoff's bound)

Let  $X$  be a r.v. s.t.  $M_X(t) = \mathbb{E}[e^{tX}]$  is finite for  $t \in (0, \theta]$ , then

$$\mathbb{P}(X \geq c) \leq e^{-tc} \mathbb{E}[e^{tX}] \quad \text{for all } t \in (0, \theta]$$

**Proof** Apply above lemma for  $f(x) = e^{tx}$

## Generalizations of Markov's inequality\*

### Example

Let  $X \sim \mathcal{N}(0, 1)$ . What is the best possible Chernoff's bound we can get ?

## Generalizations of Markov's inequality\*

### Example

Let  $X \sim \mathcal{N}(0, 1)$ . What is the best possible Chernoff's bound we can get ?

**Solution** The m.g.f. of  $X$  is defined for any  $t$ , so we can search for the best  $t$  that gives the lowest bound. Applying Chernoff's bound, for fixed  $c$  and for any  $t \in \mathbb{R}$ ,

$$\mathbb{P}(X \geq c) \leq e^{-tc} \mathbb{E}[e^{tX}] = e^{-tc} e^{t^2/2} = e^{(t-c)^2/2} e^{-c^2/2}$$

The minimum is obtained for  $t = c$  and we get

$$\mathbb{P}(X \geq c) \leq e^{-c^2/2}$$

## Generalizations of Markov's inequality\*

### Question:

- ▶ Can we define a class of r.v. that behave similarly as normal standard r.v.?
- ▶ Namely that they share the same sharp concentration inequality

### Definition (Sub-Gaussian distribution)

The proba. distribution of a r.v.  $X$  is called **sub-Gaussian** if there are positive constants  $C, \nu$ , s.t. for every  $c > 0$

$$\mathbb{P}(X \geq c) \leq Ce^{-\nu c^2/2}$$

**Idea:** "The tail of the probability distribution decreases very fast"

Namely if  $X$  is continuous  $f_X(x)$  decreases so fast as  $x \rightarrow +\infty$

that  $\int_c^{+\infty} f_X(x) dx \leq Ce^{-\nu c^2/2}$

### Why introducing sub-Gaussian distributions?

Allow a common treatment (in terms of e.g. probability inequalities) of numerous proba distributions

### Examples:

$X \sim \mathbb{N}(0, 1)$ ,  $X$  continuous with bounded support  $\{x : f_X(x) > 0\}$  is finite, ...

## Strong law of large numbers\*

Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

## Strong law of large numbers\*

### Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

**Proof** (Assume in addition  $\mathbb{E}[X_i^4]$  is finite for all  $i$ )

1. Show that  $\mathbb{E}\left[\sum_{n=1}^{+\infty} (\bar{X}_n - \mu)^4\right]$  is finite (see next slide)
  2. This implies that  $\sum_{n=1}^{+\infty} (\bar{X}_n - \mu)^4$  must be finite with probability one
  3. If  $\sum_{n=1}^{+\infty} (\bar{X}_n(\omega) - \mu)^4$  is finite, then necessarily  $(\bar{X}_n(\omega) - \mu)^4 \xrightarrow[n \rightarrow +\infty]{} 0$
- so  $(\bar{X}_n - \mu) \xrightarrow[n \rightarrow +\infty]{} 0$  with probability one

## Strong law of large numbers\*

### Theorem (Strong law of large numbers)

Let  $(X_n)_{n=1}^{+\infty}$  be a sequence of i.i.d. random variables with finite mean  $\mu$ . The empirical mean  $\bar{X}_n = (X_1 + \dots + X_n)/n$  converges almost-surely to  $\mu$ , i.e.,

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1$$

**Proof** (Assume in addition  $\mathbb{E}[X_i^4]$  is finite for all  $i$ )

Show that  $\mathbb{E}\left[\sum_{n=1}^{+\infty} (\bar{X}_n - \mu)^4\right]$  is finite

1.  $(\bar{X}_n - \mu)^4 = \frac{1}{n^4} \tilde{S}_n^4$  where  $\tilde{S}_n = \sum_{i=1}^n \tilde{X}_i$  and  $\tilde{X}_i = X_i - \mu$
2.  $\tilde{S}_n^4 = (X_1 + \dots + X_n)^4$  decomposes in terms  
 $\tilde{X}_i^4 \quad \tilde{X}_i^2 \tilde{X}_j^2 \quad \tilde{X}_i^3 \tilde{X}_j \quad \tilde{X}_i^2 \tilde{X}_j \tilde{X}_k \quad \tilde{X}_i \tilde{X}_j \tilde{X}_k \tilde{X}_\ell$
3. Only the terms  $\tilde{X}_i^4$  and  $\tilde{X}_i^2 \tilde{X}_j^2$  have non zero expectation.
4. Denoting  $\mathbb{E}[X_i^4] = \alpha \in \mathbb{R}$ ,  $\mathbb{E}[X_i^2 X_j^2] = (\mathbb{E}[X_i^2])^2 = \beta$ , we have

$$\mathbb{E}[\tilde{S}_n^4] = n \mathbb{E}[X_1^4] + 6 \binom{n}{2} (\mathbb{E}[X_1^2])^2 = n\alpha + 3n(n-1)\beta \leq n\alpha + n^2\beta$$

5. Switching expectation and infinite sum (see Beppo Levi's monotone convergence theorem) yields

$$\mathbb{E}\left[\sum_{n=1}^{+\infty} (\bar{X}_n - \mu)^4\right] = \mathbb{E}\left[\sum_{n=1}^{+\infty} \frac{\tilde{S}_n^4}{n^4}\right] = \sum_{n=1}^{+\infty} \mathbb{E}\left[\frac{\tilde{S}_n^4}{n^4}\right] \leq \alpha \sum_{n=1}^{+\infty} n^{-3} + \beta \sum_{n=1}^{+\infty} n^{-2} < +\infty$$