MATH/STAT395: Probability II Spring 2020

Review of MATH/STAT394

Chapters 1, 2, 3, Sections 4.4, 4.5, 4.6 of ASV

Instructor: Vincent Roulet *Teaching Assistant: Zhenman Yuen*

Review of probability distributions

This lecture note serves as reference about the material you should know from MATH/STAT394. Starred items are advanced topics, you don't need to know but it is preferable.

1 Probability space, conditional probability, independence

1.1 Probability space

Definition 1 (Probability space)**.** *A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ consists of*

- *A sample space Ω, 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 Ω, ($\mathcal{F} \subset 2^\Omega$ must be a σ-algebra)*
- *A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ 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 $A_1, A_2, \ldots \in \mathcal{F}$ of disjoint sets,*

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

Definition 2 (σ-algebra*)**.** *Let Ω be a set. A σ-algebra $\mathcal{F}$ on Ω is a subset of $2^\Omega = \{B \subset \Omega\}$ such that*

1. *$\Omega \in \mathcal{F}$*
2. *For any $A \in \mathcal{F}$, $A^c \triangleq \Omega \setminus A \in \mathcal{F}$*
3. *For any $A_1, A_2, \ldots \in \mathcal{F}$, $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$*

The smallest σ-algebra that contains all intervals of $\mathbb{R}^n$ is called the Borel algebra of $\mathbb{R}^n$.

1.2 Conditional probability

Definition 3 (Conditional probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ s.t. $\mathbb{P}(B) \neq 0$, the **conditional probability of** A **given** B is defined as*

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

Definition 4. *$B_1, \ldots, B_n \subset \Omega$ is a partition of Ω if $\bigcup_{i=1}^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for any $i \neq j$.*

Property 5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space*

1. *For $B \in \mathcal{F}$ s.t. $\mathbb{P}(B) \neq 0$, $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability*

2. For $A_1 \dots A_n \in \mathcal{F}$,

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_n|A_{n-1} \cap \dots \cap A_1)\, \mathbb{P}(A_{n-1}|A_{n-2} \cap \dots \cap A_1) \dots \mathbb{P}(A_1)$$

3. Let $B_1, \dots, B_n \in \mathcal{F}$ a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i, then we have

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\, \mathbb{P}(B_i)$$

Theorem 6 (Bayes Formula)**.** *Let $B_1, \dots, B_n \in \mathcal{F}$ a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i, then we have for any $k \in \{1, \dots, n\}$,*

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_k)\, \mathbb{P}(B_k)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i)\, \mathbb{P}(B_i)}$$

1.3 Independence

Definition 7 (Independence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are said to be **independent** if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\, \mathbb{P}(B)$$

*n events $A_1, \dots A_n \in \mathcal{F}$ are **independent** or **mutually independent** if for any $2 \leq k \leq n$ and $1 \leq i_1 \leq \dots \leq i_k \leq n$,*

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \dots \mathbb{P}(A_{i_k})$$

Property 8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If A, B are independent, then any pair of events $(A^*, B^*) \in \{(A, B), (A^c, B), (A, B^c), (A^c, B^c)\}$ is a pair of independent events.*

Definition 9 (Conditional independence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ s.t. $\mathbb{P}(B) \neq 0$, events $A_1, \dots A_n$ are **conditionally independent** if they are independent with respect to the probability $\mathbb{P}(\cdot|B)$.*

Definition 10. *Let $X_1, \dots X_n$ be r.v. (see definition below) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $X_1, \dots X_n$ are independent if for any[1] $B_1, \dots, B_n \subset 2^{\Omega}$,*

$$\mathbb{P}(X_1^{-1}(B_1) \cap \dots \cap X_n^{-1}(B_n)) = \prod_{i=1}^{n} \mathbb{P}(X_i^{-1}(B_i))$$

2 Random variables

2.1 Probability distribution

Definition 11 (Probability distribution of a random variable)**.** *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a (real-valued) **random variable** (r.v.) X is defined as a mapping $X : \Omega \to \mathbb{R}$ such that for any[1] 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*

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

We that "X follows a distribution $\mathbb{P}_X$" and denote it by $X \sim \mathbb{P}_X$.

Definition 12 (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 $\mathcal{X} = X(\Omega)$ s.t. $\sum_{k \in \mathcal{X}} \mathbb{P}(X = k) = 1$*

[1] A formal definition requires to restrict the subsets considered in the definition to belong to the Borel algebra of $\mathbb{R}$ defined above.

2.2 Probability functions

Definition 13 (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 :*

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

Definition 14 (Probability density function)**.** *Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If a function f satisfies*

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

*then f is called the **probability density function** (p.d.f.) of X. X is then called a **continuous** r.v.*

Note: From now on, in the definitions, we consider without loss of generality, that if X is a discrete random variable, then $X(\Omega) = \mathbb{Z}$, that is, we identify any countable set to the set of integers. For random variables taking values in a finite set $\mathcal{X}$, it means that we assume this set to be a set of integers and that we consider $\mathbb{P}(X = k) = 0$ for any $k \in \mathbb{Z} \setminus \mathcal{X}$. Similarly, for continuous random variables, we consider $\mathcal{X}(\Omega) = \mathbb{R}$, such that if the random variable takes values in a bounded set $\mathcal{X}$, then $f(x) = 0$ for any $x \in \mathbb{R} \setminus \mathcal{X}$.

Property 15. *Let f be a p.d.f. of a r.v. X then*

1. $\int_{-\infty}^{+\infty} f(x)dx = 1$, $f(x) \geq 0$ *for all* $x \in \mathbb{R}$
2. $\mathbb{P}(X = k) = \int_k^k f(x)dx = 0$

Definition 16 (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 17. *Let F be the c.d.f. of a r.v. then*

1. $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a)$
2. $\lim\limits_{t \to -\infty} F(t) = 0$, $\lim\limits_{t \to +\infty} F(t) = 1$
3. *If* $s \leq t$, $F(s) \leq F(t)$
4. $F(t) = \lim\limits_{s \to t^+} F(s)$

2.3 Expectation

Definition 18 (Expectation)**.** *Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

1. *(Discrete case) If X has a p.m.f p s.t. $\sum_{k \in \mathbb{Z}} |k| p(k) < \infty$, the **expectation** (or **expected value**) of X exists and reads*
$$\mathbb{E}[X] = \sum_{k \in \mathbb{Z}} k p(k)$$
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*
$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x)dx$$

Property 19 (Linearity of Expectation)**.** *Let X, Y be two (discrete/continuous) r.v. and $a \in \mathbb{R}$,*

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

Proof. If X, Y are two discrete continuous random variables the result comes from the linearity of the sum. If X, Y are two continuous random variables, the result comes from the linearity of the integral. $\square$

Property 20 (Expectation of a function of a random variable)**.** *Let X be a r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $g : X(\Omega) \to \mathbb{R}$. Then $g(X)$ is a r.v. and*

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

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

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

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

Property 21. *Let X be a r.v. with probability distribution $\mathbb{P}_X$ and c.d.f. F_X, then*

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

where $\mathbf{1}_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise} \end{cases}$

2.4 Moments, Variance

Definition 22 (Moment)**.** *For a r.v. X and $m \in \mathbb{N}$, if $\mathbb{E}[|X|^m] < +\infty$, then*

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

Definition 23 (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*

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

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

Definition 24 (Degenerate random variable)**.** *A r.v. X is said to be degenerate if $\exists a \in \mathbb{R}$ s.t. $\mathbb{P}(X = a) = 1$.*

Property 25. *If X is a degenerate r.v. as defined in Def. 24, then $\mathbb{E}[X] = a$. Moreover, for any r.v. X we have $\mathrm{Var}(X) = 0 \Leftrightarrow X$ is degenerate.*

Remark 26. *In the course, for any $b \in \mathbb{R}$, we define e.g. $\mathbb{E}[b]$ by identifying b to the associated degenerate r.v.* $X : \begin{cases} \Omega & \to \mathbb{R} \\ \omega & \mapsto b \end{cases}$

Property 27. *For any r.v. X and $a, b \in \mathbb{R}$,*

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

2.5 Common discrete random variables

In the following we emphasize the set of values that can take the random variable as $X(\Omega) = \{k \in \mathbb{Z} : \mathbb{P}(X = k) \neq 0\}$.

2.5.1 Bernoulli

Model Models the success of a trial (1 for success, 0 for fail)
Example Can model that the flip of a coin will be tail.
Range $X(\Omega) = \{0, 1\}$
Parameters $p \in [0, 1]$
Notation $X \sim \text{Ber}(p)$
Probability mass function $\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p$
Expectation, Variance $\mathbb{E}[X] = p$, $\text{Var}(X) = p(1 - p)$

2.5.2 Binomial

Model Model the number of success among n trials, each trial being independent and identically distributed as a Bernoulli r.v. with parameter p
Example Models the number of tails among n flips of a coin
Range $X(\Omega) = \{0, \ldots, n\}$
Parameters $n \in \mathbb{N}, p \in [0, 1]$
Notation $X \sim \text{Bin}(n, p)$
Probability mass function $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)} \quad \text{for } k \in \{0, \ldots n\}$
Expectation, Variance $\mathbb{E}[X] = np, \text{Var}[X] = np(1 - p)$

Proof. Proof done for expectation. For the variance the proof can be found in the book page 115. We will provide a much simpler proof later. □

Remark Can be written as $X = \sum_{i=1}^{n} B_i$, where $B_i \sim \text{Ber}(p)$ are n independent Bernoulli r.v.

2.5.3 Poisson

Model Models the number of success among an infinite number of trials, with an average number of success λ
Example Models the number of typos in an infinite document
Range $X(\Omega) = \mathbb{N}$
Parameters $\lambda > 0$
Notation $X \sim \text{Poisson}(\lambda)$
Probability mass function $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } k \in \mathbb{N}.$
Expectation, Variance $\mathbb{E}[X] = \lambda$, $\text{Var}[X] = \lambda$.
Remark Consider a sequence of binomial random variables $B_n \sim \text{Bin}(n, \lambda/n)$ defined for $n > \lambda$, such that the average number of success of all those random variables is independent of n, then this sequence of random variables converge in distribution to a Poisson distribution as n goes to infinity. That is we retrieve the model of a Poisson distribution as the number of successes among an infinite number of trials.

2.5.4 Geometric

Model Models the number of trials of Bernoulli random variable with proba of success p before getting one success
Example Number of times you play an armed bandit before getting some money
Range $X(\Omega) = \mathbb{N}$
Parameters $p \in [0, 1]$
Notation $X \sim \text{Geom}(p)$
Probability mass function $\mathbb{P}(X = k) = (1 - p)^{k-1} p$
Expectation, Variance $\mathbb{E}(X) = \frac{1}{p}, \text{Var}(X) = \frac{1-p}{p^2}$

2.5.5 Hypergeometric*

Model Models sampling without replacement with order not mattering. Specifically denote K the number of items A in a total number of items N and assume we draw n items from this set. The random variable $X =$" number of items A in the n items that we sampled from the set" is distributed as a hypergeometric random variable

Range $X(\Omega) = \{0, \ldots K\}$

Parameters $K, N, n \in \mathbb{N}$ with $1 \leq n \leq N$ and $1 \leq K \leq N$

Notation $X \sim \text{Hypergeom}(N, K, n)$

Probability mass function $\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Expectation $\mathbb{E}[X] = n\frac{K}{N}$

2.6 Common continuous random variables

In the following we emphasize the set of values that can take the random variable as $X(\Omega) = \{x \in \mathbb{R} : f(x) \neq 0\}$.

2.6.1 Uniform

Model Uniform probability on an interval $[a, b]$, with $a < b$

Example Models the reaching point of a bowling ball

Range $X(\Omega) = [a, b]$

Parameters $a, b \in \mathbb{R}$, $a < b$

Notation $X \sim \text{Unif}([a, b])$

Probability density function $f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

Expectation, Variance $\mathbb{E}(X) = \frac{a+b}{2}, \text{Var}(X) = \frac{(b-a)^2}{12}$

2.6.2 Gaussian

Model Standard continuous distribution to model a continuous random variable centered around a point μ with variance σ^2

Range $X(\Omega) = \mathbb{R}$

Parameters μ, σ^2

Notation $X \sim \mathcal{N}(\mu, \sigma^2)$

Probability density function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Expectation, Variance $\mathbb{E}(X) = \mu$, $\text{Var}(X) = \sigma^2$

Remark Appears as the asymptotic behavior of the empirical mean of independent and identically distributed random variables, see central limit theorem studied later in the course.

2.6.3 Exponential

Model Models the waiting time before an event occurs, with an average of waiting time λ

Example Waiting time for a phone call

Range $X(\Omega) = [0, +\infty)$

Parameters $\lambda > 0$

Notation $X \sim \text{Exp}(\lambda)$

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

Expectation, Variance $\mathbb{E}(X) = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2}$

Remark Can be seen as the continuous time counterpart of the geometric distribution see lecture 4

2.6.4 Gamma distribution*

Model Versatile family of distribution that can model for example the time needed for a n^{th} phone call

Range $X(\Omega) = [0, +\infty)$

Parameters $\lambda > 0, r > 0$

Notation $X \sim \mathrm{Gamma}(r, \lambda)$

Probability density function $f(x) = \begin{cases} \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$ where $\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx$

Expectation, Variance $\mathbb{E}(X) = \frac{r}{\lambda}, \mathrm{Var}(X) = \frac{r}{\lambda^2}$

MATH/STAT395: Probability II	Spring 2020

Joint Probability Distributions, Independence

Sections 6.1, 6.2, 6.3 of ASV

Instructor: Vincent Roulet — *Teaching Assistant: Zhenman Yuen*

1 Multivariate random variables

Definition 1 (Multivariate random variable/Random vector). *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a* ***multivariate random variable*** *or* ***random vector*** *is a vector $X = (X_1, ..., X_n)$, whose components are real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.*

Example 2 (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*

1.1 Discrete case

1.1.1 Joint probability mass function

Definition 3 (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*

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

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

Example 4.

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	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	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

Lemma 5. *Let $g : \mathbb{R}^n \to \mathbb{R}$ and let $X_1, \ldots, X_n$ be discrete r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with joint probability mass function p, then*

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

Example 6. *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. The average score reads

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

□

1.1.2 Marginal probability mass function

Definition 7. *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 Ω y definition of a discrete random variable such that

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

□

Definition 8. *Let p be the joint probability mass function of n discrete r.v. $X_1, \ldots X_n$. The probability mass function of X_j for $j \in \{1, \ldots, n\}$ is given by for any $k \in X_j(\Omega)$,*

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

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

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

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

Example 9. *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.*

Solution. Sum the columns of $p(s, y)$. So you get $\mathbb{P}(Y = 1) = 4/16$ and $\mathbb{P}(Y = 0) = 12/16$ □

1.1.3 Multinomial distribution

Motivation Consider a trial with r possible outcomes, labeled $1, \ldots, r$. Denote p_j the probability of the out come j such that $p_1 + \ldots + p_r = 1$. 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 $(X_1, \ldots, X_r)$?

Derivation

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

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

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

Definition 10 (Multinomial distribution)**.** *Let $n, r \in \mathbb{N}_*$, let $p_1, \ldots, p_r \in (0, 1)$ s.t. $p_1 + \ldots + p_r = 1$, then a r.v. X has a **multinomial distribution** with parameters $n, r, p_1, \ldots, p_r$ if it is defined for any $k_1, \ldots k_r \in \mathbb{N}$ s.t. $k_1 + \ldots + k_r = n$ with probability*

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

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

1.2 Continuous Random Variables

1.2.1 Joint probability density function

Definition 11 (Joint probability density function)**.** *Random variables $X_1, \ldots, X_n$ are **jointly continuous** if there exists a **joint probability density function** $f : \mathbb{R}^n \to \mathbb{R}$ such that for any*[1] *$B \subset \mathbb{R}^n$,*

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

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

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

Lemma 12. *Let $X_1, \ldots, X_n$ 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* $\dim(E) = m < n$,

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

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

Example 13 (Synthetic)**.** *Assume X, Y have a joint p.d.f.*

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

1. *Check that it is a valid joint p.d.f*
2. *Compute $\mathbb{P}(X < Y)$*

Solution. 1. We have $f(x,y) \ge 0$ and

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

2.

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

□

1.2.2 Uniform continuous random variables

Definition 14 (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 $(X, Y, Z) \sim \text{Unif}(D)$.

Lemma 15. *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.

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

□

1.2.3 Marginal Probability Density Function

Definition 16 (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

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

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

Definition 17 (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.

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

Example 18. *Consider a disk of radius r, $D_r = \{(x,y) : x^2 + y^2 \leq r\}$ and $(X,Y) \sim \mathrm{Unif}(D_r)$.*

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 $|x| \leq r$,

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

□

1.3 Joint cumulative distribution

Definition 19 (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 20. *1. If (X,Y) are jointly continuous with joint p.d.f. f,*

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

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

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

2 Joint Probability Distributions and Independence

2.1 Independence of random variables

Definition 21 (Independent random variables)**.** *Random variables $X_1, \dots, X_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are* ***independent*** *if for any*[2] *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

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

2.2 Discrete case

Lemma 22. *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

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

□

Example 23.

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	*5*	*1/4*	*0*
	6	*1/8*	*1/16*
	7	*1/8*	*0*
	8	*0*	*1/16*

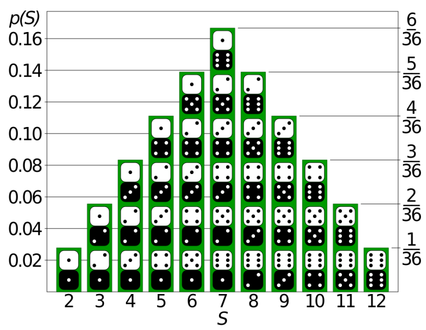
Solution. Check for example $\mathbb{P}(S = 2, Y = 0) = 0 \neq \mathbb{P}(S = 2)\, \mathbb{P}(Y = 0) > 0$

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 □

Example 24.

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?*



[2] Again a formal definition requires these subsets to be Borel subsets of $\mathbb{R}^n$

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

□

2.3 Continuous case

Lemma 25. *Let $X_1, \ldots, X_n$ be n r.v. on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that for $j \in \{1, \ldots, n\}$, the rv. X_j has p.d.f. f_{X_j}.*

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

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

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

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

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

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

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

Conversely, if X, Y are independent

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

□

Example 26. *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 $W{=}\{(x,y) : -1{\le}x{\le}1, 0{\le}y{\le}1\}$.*

1. *Are X, Y independent?*

2. *What consequences it had when computing the probability to get the target $T{=}\{(x,y){:}{-0.1}{\le}x{\le}0.1, 0.4{\le}y{\le}0.6\}$?*

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.

□

3 Borel Algebra*

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 σ-algebra $\mathcal{B}$

Definition 27 (σ-algebra)**.** *Let Ω be a set, a σ-algebra $\mathcal{F}$ on Ω 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 $A_1, A_2, \ldots \in \mathcal{F}$, $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$*

Why introducing σ-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.

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

which leads to $1 \leq \sum_{k=1}^{+\infty} \mathbb{P}(V_k) \leq 3$ which is impossible
Formally, we restrict our focus on the Borel algebra of $\mathbb{R}^n$

Definition 28 (Borel algebra in $\mathbb{R}^n$)**.** *The Borel algebra in $\mathbb{R}^n$, denoted $\mathcal{B}_n$, is the smallest σ-algebra (in terms of inclusion) that contains*

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

or equivalently defined as the smallest σ-algebra that contains

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

Consequence

1. If we can measure all intervals of the form $(-\infty, a_1] \times \ldots \times (-\infty, a_n]$ for $a_i \in \mathbb{R}$, then we can measure all subsets of interests, i.e. all $B \in \mathcal{B}_n$,

   $\rightarrow$ 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.

MATH/STAT395: Probability II	Spring 2020

Functions of Random Variables

Sections 5.1, 6.3, 6.4, 7.1 of ASV

Instructor: Vincent Roulet *Teaching Assistant: Zhenman Yuen*

1 Functions of Random Variables

Consider either

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

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

1.1 Classical approach

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

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

Proof. For continuous case it comes from the definition. For the discrete case, we use that $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$ with $B^c = \Omega \setminus B$ for $A = \{Y \leq k\}$ and $B = \{Y \leq k-1\}$

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

Similarly if we have access to $\bar{F}_Y(k) = 1 - F_Y(k) = \mathbb{P}(Y > k)$,

$$\mathbb{P}(Y = k) = \mathbb{P}(k-1 < Y \leq k) = \mathbb{P}(k-1 < Y) - \mathbb{P}(k < Y, k-1 < Y) = \mathbb{P}(k-1 < Y) - \mathbb{P}(k < Y) = \bar{F}_Y(k-1) - \bar{F}_Y(k)$$

□

Example 1. *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.

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

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

□

1.2 Change of p.d.f. of one random variable

Lemma 2. *Let X be a continuous r.v. with p.d.f. f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and strictly monotonic with inverse denoted $\gamma = g^{-1}$, then the p.d.f of $Y = g(X)$ exists*[1] *and it reads*

[1] We admit that fact

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

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 $f_Y(t) = 0$ and since the probability on a point does not matter we can define $f_Y(a) = 0$
2. If $t < b$, $F_Y(t) = \mathbb{P}(g(X) \leq b) = 1$ so $f_Y(t) = 0$ and since the probability on a point does not matter we can define $f_Y(b) = 0$

Now

1. if g is strictly increasing, for $t \in (a, b)$ s.t. $g^{-1}(t)$ is defined,
$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(g(X) \leq t) = \mathbb{P}(X \leq g^{-1}(t)) = F_X(\gamma(t))$$
$$\text{so} \quad f_Y(t) = \gamma'(t) f_X(\gamma(t))$$
2. if g is strictly decreasing, for $t \in (a, b)$ s.t. $g^{-1}(t)$ is defined,
$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(g(X) \leq t) = \mathbb{P}(X \geq g^{-1}(t)) = 1 - F_X(\gamma(t))$$
$$\text{so} \quad f_Y(t) = -\gamma'(t) f_X(\gamma(t))$$

Finally 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. □

1.3 General method

Practical Method

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

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

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

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 $$f_Y(y) = F_Y'(t) = \phi(y)$$

Example 3. *Let X be a continuous r.v., $g : x \to x^2$, $Y = g(x)$. What is the p.d.f. of Y?*

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

[2] This ensures that Y is also a continuous r.v.

1.4 Change of joint p.d.f. for two random variables

Theorem 4. *Let (X, Y) jointly continuous with p.d.f. $f_{X,Y}$, denote $S=\{(x,y){:}f_{X,Y}(x,y)>0\}$*
Let $g : \mathbb{R}^2 \to \mathbb{R}^2$ such that

1. *g is invertible on S with inverse $\gamma(u,v) = (\alpha(u,v), \beta(u,v))$ ($\alpha : \mathbb{R}^2 \to \mathbb{R}, \beta : \mathbb{R}^2 \to \mathbb{R}$)*
2. *γ is continuously differentiable on $g(S)$ (partial derivatives are continuous)*
3. *The determinant of the Jacobian $J_\gamma(u,v)$ of γ 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

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

□

Example 5. *Let X, Y be two independent standard* Exp(λ) *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 $f^0_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)}\, \mathbf{1}_{[0,+\infty)^2}(x,y)$
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.
$g : (x,y) \to (x+y, \frac{x}{x+y})$ 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

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

□

2 Functions of Independent Random Variables

2.1 General result

Lemma 6. *Let $X_1, \ldots, X_{m+n}$ be $m + n$ independent r.v. (discrete or continuous).*
Let $g : \mathbb{R}^m \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$.
Then $Y = g(X_1, \ldots, X_m)$ and $Z = h(X_{m+1}, \ldots, X_{m+n})$ are independent.

2.2 Maximum, Minimum of Independent Random Variables

Lemma 7. *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

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

Proof.

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

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 □

Example 8. *Let $X_1, \dots X_n$ be n independent r.v. following $X_i \sim \mathrm{Geom}(p_i)$, $p_i \in (0, 1)$*
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 $Y \sim \mathrm{Geom}(r)$. □

Example 9. *Let $X \sim \mathrm{Exp}(\lambda)$, $Y \sim \mathrm{Exp}(\mu)$, 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., $Z \sim \mathrm{Exp}(\lambda + \mu)$ □

2.3 Sums of Independent Random Variables

Lemma 10. *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 $x \in \mathbb{R}$,*

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

The $\star$ operation is called a convolution. So summing two random variables amount to convolve their p.m.f./p.d.f.

Proof. 1.

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

2.

$$F_{X+Y}(z) = \mathbb{P}(X+Y \le z) = \iint\limits_{x+y\le z} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{z-x} f_{X,Y}(x)f_Y(y)dy\right)dx$$
$$= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{z} f_X(x)f_Y(w-x)dw\right)dx = \int_{-\infty}^{z}\left(\int_{-\infty}^{+\infty} f_X(x)f_Y(w-x)dx\right)dw$$

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

□

Example 11 (Sums 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

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

□

Example 12 (Sums of Binomial). *$X \sim \text{Bin}(m_1,p)$ and $Y \sim \text{Bin}(m_2,p)$ independent. Distribution of $X+Y$?*

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

So $X+Y \sim \text{Bin}(m_1+m_2,p)$ □

Example 13 (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+...+X_m?

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 to know)

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

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 $Z \sim \text{Negbin}(m,p)$

Lemma 14. *Let $X_1,\ldots,X_n$ be independent Gaussian variables $X_i \sim \mathcal{N}(\mu_i,\sigma_i^2)$, then*

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

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 $\left(Z_i = \frac{X_i-\mu_i}{\sigma_i}\right)$
We have $X_1+X_2 = \sigma_1\left(Z_1 + \frac{\sigma_2}{\sigma_1}Z_2\right) + \mu_1+\mu_2$
Now remains to compute distribution of $Y = Z_1 + \sigma Z_2$ with $\sigma = \frac{\sigma_2}{\sigma_1}$

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

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

MATH/STAT395: Probability II	Spring 2020

Exchangeability, i.i.d. r.v., mean, variance computations

Sections 7.2 8.1 8.2 of ASV

Instructor: Vincent Roulet *Teaching Assistant: Zhenman Yuen*

1 Exchangeability

1.1 Motivating example

Flip over a cards from shuffled deck one by one. What is the probability that the 23rd card is a spade?

1. (Intuition)

   Without any additional information, $\mathbb{P}$("23rd card is a spade") should be equal to $\mathbb{P}$("1st card is a spade")

2. (How to formalize that?)

   (a) Define the r.v. associated to the first 23rd cards $X_1, \ldots, X_{23}$ with $X_i \in \{\text{heart}, \text{diamond}, \text{spade}, \text{club}\}$

   (b) Write down the joint p.m.f. of $X_1, \ldots, X_{23}$

   (c) Compute marginal p.m.f. of X_{23} and of X_1, should be the same

   $\rightarrow$ *The joint p.m.f. must satisfy some property... and that's not independence...*

Applications

1. Independent identically distributed (i.i.d.) random variables

2. Sample without replacement

1.2 Identically distributed, exchangeable r.v.

Definition 1 (Equality in distribution)**.** *Two random vectors $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$ are **equal in distribution** if*

$$\mathbb{P}((X_1, \ldots, X_n) \in B) = \mathbb{P}((Y_1, \ldots, Y_n) \in B) \quad \text{for any } B \subset \mathbb{R}^n$$

we denote it

$$(X_1, \ldots, X_n) \stackrel{d}{=} (Y_1, \ldots, Y_n)$$

Definition 2 (Identically distributed)**.** *$X_1, \ldots X_n$ are **identically distributed** if for any $k, j \in \{1, \ldots n\}$,*

$$X_k \stackrel{d}{=} X_j$$

*i.e. they have same **marginal** p.m.f. (Discrete case) or p.d.f. (Continuous case)*[1]

Definition 3 (Exchangeability)**.** *$X_1, \ldots X_n$ are **exchangeable** if for any permutation $k_1, \ldots, k_n$ of $\{1, \ldots, n\}$,*

$$(X_1, \ldots X_n) \stackrel{d}{=} (X_{k_1}, \ldots, X_{k_n})$$

Lemma 4 (Consequences of exchangeability 1)**.** *Let $(X_1, \ldots, X_n)$ be exchangeable, then they are identically distributed*

[1] 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

Proof. Let $B_1 \subset \mathbb{R}$ and $B_2 = \ldots B_n = \mathbb{R}$,

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

So X_1, X_j are identically distributed (same for any k, j in $\{1, \ldots n\}$) $\square$

Example 5 (Flipping 23 cards)**.** *In our motivating example, if we show exchangeability, then*

$$\mathbb{P}(X_{23} \text{ is a spade}) = \mathbb{P}(X_1 \text{ is a spade}) = 1/4$$

Lemma 6 (Consequences of exchangeability 2)**.** *Let $(X_1, \ldots, X_n)$ be exchangeable, then for any $k \in \{1, \ldots n\}$ and any permutation $(i_1, \ldots i_k)$ of $\{1, \ldots k\}$*

$$(X_1, \ldots, X_k) \stackrel{d}{=} (X_{i_1}, \ldots X_{i_k})$$

and for any $g : \mathbb{R}^k \to \mathbb{R}$, $\mathbb{E}[g(X_1, \ldots, X_k)] = \mathbb{E}[g(X_{i_1}, \ldots, X_{i_k})]$

Proof. Follows from same reasoning as th proof for Consequences of exchangeability 1 $\square$

Lemma 7 (How to check for exchangeability)**.** *Let $(X_1, \ldots, X_n)$ be random variables,*

1. *If $X_1, \ldots X_n$ are discrete, they are exchangeable if and only if their joint p.m.f. p is symmetric, i.e.*

$$\begin{aligned}\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n) &= p(k_1, \ldots, k_n) \\ &= p(k_{i_1}, \ldots, k_{i_n}) = \mathbb{P}(X_1 = k_{i_1}, \ldots, X_n = k_{i_n})\end{aligned}$$

   for $k_1, \ldots, k_n \in \mathbb{Z}$ and $i_1, \ldots, i_n$ a permutation of $\{1, \ldots n\}$

2. *If $X_1, \ldots X_n$ are jointly continuous, they are exchangeable if and only if their joint p.d.f. f is symmetric, i.e.*

$$f(x_1, \ldots, x_n) = f(x_{i_1}, \ldots, x_{i_n})$$

   for $x_1, \ldots, x_n \in \mathbb{R}$ and $i_1, \ldots, i_n$ a permutation of $\{1, \ldots n\}$

Example 8. *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. *$f(x_1, x_2, x_3) = (x_1 x_2 + x_3) \, \mathbf{1}_{[0,1]^3}(x_1, x_2, x_3)$?*

Solution

1. *Yes, one can try to permute the values of x_1, x_2, x_3, the p.d.f. will still have the same value*
2. *No, take $x = (x_1, x_2, x_3)$ with $x_1 = 1, x_2 = 0.5, x_3 = 0$, define $y(y_1, y_2, y_3) = (x_3, x_2, x_1)$ which is a permutation of (x_1, x_2, x_3), $f(x_1, x_2, x_3) = 0.5 \neq f(y_1, y_2, y_3) = 1$.*

1.3 Sampling without replacement

Theorem 9. *Let $X_1, \ldots, X_m$ denote the outcomes of successive draws uniformly at random without replacement from $\{1, \ldots, n\}$ (n distinct objects numbered from 1 to n) with $m \leq n$.*

Then $X_1, \ldots, X_m$ are exchangeable.

Proof. Let $k_1, \ldots k_m$ be m elements of $\{1, \ldots, n\}$. Then

$$\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n) = \frac{1}{n} \times \frac{1}{n-1} \times \ldots \times \frac{1}{n-m+1}$$

which shows that the joint p.m.f. only depends on the number of draws.

Formally, for $i_1, \ldots, i_m$ a permutation of $\{1, \ldots, m\}$,

$$\mathbb{P}(X_1 = k_{i_1}, \ldots, X_m = k_{i_m}) = \frac{1}{n(n-1)\ldots(n-m+1)} = \mathbb{P}(X_1 = k_1, \ldots, X_n = k_n)$$

which shows exchangeability.

$\square$

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,
   $X_i \in \{\text{spade, heart, diamond, club}\}$, $\{X_i\text{=spade}\} \Leftrightarrow$ "the i^{th} card is a spade"
5. $Y_1, \dots, Y_{23}$ are distinct and drawn without replacement so exchangeable
6. What about $g(Y_1), \dots, g(Y_{23})$?

Theorem 10. *If $Y_1, \dots, Y_n$ are exchangeable, then for any function g, $g(Y_1), \dots, g(Y_n)$ are exchangeable.*

Example 11 (Flipping 23 cards)**.** *From our previous reasoning, we get*

$$\mathbb{P}(X_{23} \textit{ is a spade}) = \mathbb{P}(X_1 \textit{ is a spade}) = 1/4$$

Example 12. *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

$$\mathbb{P}(X_3 = \text{red}, X_7 = \text{green}) = \mathbb{P}(X_1 = \text{red}, X_2 = \text{green}) = \frac{5}{8} \times \frac{3}{7} \approx 0.27$$

$\square$

1.4 Independent, Identically Distributed Random Variables

Lemma 13. *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 $\square$

Example 14 (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 $\mathbb{P}(X_1 \text{ is largest}) = 1/3$ □

Example 15. *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^{th} card.

$$\begin{aligned}
\mathbb{P}(X_6 = \text{ queen}|X_5 = \text{queen}, X_{10} = \text{queen}) &= \frac{\mathbb{P}(X_6 = \text{queen}, X_5 = \text{queen}, X_{10} = \text{queen})}{\mathbb{P}(X_5 = \text{queen}, X_{10} = \text{queen})} \\
&= \frac{\mathbb{P}(X_1 = \text{queen}, X_2 = \text{queen}, X_3 = \text{queen})}{\mathbb{P}(X_1 = \text{queen}, X_2 = \text{queen})} \\
&= \mathbb{P}(X_3 = \text{queen}|X_1 = \text{queen}, X_2 = \text{queen}) \\
&= \frac{2}{50} \approx 0.04
\end{aligned}$$

□

2 Empirical estimators of mean and variance

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 $X_1, ..., X_n$, define the **sample mean/empirical mean**

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

- What is the expectation of $\bar{X}_n$? (easy)
- What is the variance of $\bar{X}_n$? (needs more tools!)

2.1 Independence, expectation, variance: keys lemmas

Lemma 16 (Expectation of product of independent random variables)**.** *Let $X_1, \ldots, X_n$ be independent r.v.*

Let $g_1, \ldots, g_n$ be n functions $g_i : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[g_i(X_i)]$ is defined.

$$\mathbb{E}[g_1(X_1) \ldots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \ldots \mathbb{E}[g_n(X_n)]$$

Proof. (2 continuous r.v. case) Let X, Y be independent and continuous, $g, h : \mathbb{R} \to \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[g(X)h(Y)] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x)h(y)f_{X,Y}(x,y)dxdy \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\
&= \int_{-\infty}^{+\infty} g(x)f_X(x)dx \int_{-\infty}^{+\infty} h(y)f_Y(y)dy = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]
\end{aligned}$$

□

Important remark: The lemma above can be in fact seen as a characterization of independence, it is a crucial one to know. Remember the following carefully

- For **any** r.v. $X_1, \ldots, X_n$ the expectation of the sum $X_1 + \ldots + X_n$ is the sum of the expectations
$$\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots \mathbb{E}[X_n]$$
- For **independent** r.v. $X_1, \ldots, X_n$ the expectation of the product $X_1 \cdot \ldots \cdot X_n$ is the product of the expectations
$$\mathbb{E}[X_1 \ldots X_n] = \mathbb{E}[X_1] \ldots \mathbb{E}[X_n]$$

Most of the properties seen in the course follow easily from one or two of these facts. (In the lemma above we use $Y_1 = g_1(X_1), \ldots, Y_n = g_n(X_n)$ that are independent as functions of independent r.v.)

Lemma 17. *Let $X_1, \ldots, X_n$ be n **independent** random variables with finite variance*
$$\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$$

Proof. **Note:** A much simpler proof is provided using covariance later.
Denote $\mu_i = \mathbb{E}[X_i]$, we know that $\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n \mu_i$.

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^n X_i - \mu_i\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^n X_i - \mu_i\right)\left(\sum_{j=1}^n X_j - \mu_j\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n (X_i - \mu_i)^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^n (X_i - \mu_i)(X_j - \mu_j)\right] \\
\begin{array}{c}\text{(Linearity of Expectation} \\ \text{+ Expectation of product} \\ \text{of independent r.v.)}\end{array} \quad &= \sum_{i=1}^n \mathbb{E}\left[(X_i - \mu_i)^2\right] + \underbrace{\sum_{\substack{i,j=1 \\ i \neq j}}^n \underbrace{\mathbb{E}[(X_i - \mu_i)]}_{=0} \mathbb{E}[(X_j - \mu_j)]}_{=0} \qquad = \sum_{i=1}^n \mathrm{Var}(X_i)
\end{aligned}$$

□

Example 18 (Variance of a binomial random variable, easy computation)**.** *Let $X \sim \mathrm{Bin}(n, p)$, what is the variance of X?*

Solution. By definition, $X = B_1 + \ldots + B_n$ where $B_i \sim \mathrm{Ber}(p)$ are independent. We have $\mathrm{Var}(B_i) = p(1-p)$, so $\mathrm{Var}(X) = \mathrm{Var}(B_1) + \ldots + \mathrm{Var}(B_n) = np(1-p)$ □

Example 19 (Variance of negative binomial random variable)**.** *Let $X \sim \mathrm{NegBin}(n, p)$, i.e. $X = G_1 + \ldots + G_n$ with $G_i \sim \mathrm{Geom}(p)$ independent.*
What is the expectation and variance of X?

Solution. We have $\mathbb{E}(G_i) = \frac{1}{p}$ and $\mathrm{Var}(G_i) = \frac{1-p}{p^2}$. So $\mathbb{E}(X) = \frac{n}{p}$, $\mathrm{Var}(X) = \frac{n(1-p)}{p^2}$. □

2.2 Empirical mean, estimators

2.2.1 Empirical mean

Definition 20. *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables drawn from the distribution of a random variable X with mean μ and variance σ^2. The **sample mean** or **empirical mean** of the first n observations is defined as*
$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$
It satisfies $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

Proof. Follows from the linearity of expectation for $\mathbb{E}(\bar{X}_n)$. For $\text{Var}(\bar{X}_n)$ it follows from above lemma using that the variables are i.i.d. □

Note: The variance of the empirical mean tends to 0 as $n \to +\infty$.
Gives the intuition that, as $n \to +\infty$, $\bar{X}_n$ converges to the mean of X, i.e. μ
(This will be shown properly with a proof of the law of large numbers)

2.2.2 Estimators

Definition 21 (Estimator)**.** *Let θ 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 θ from n observations is a function of the n i.i.d. r.v.*
2. *The bias of an estimator $\hat{\theta}$ of θ 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 22. *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 μ is the mean of the r.v. X*

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?

$\to$ **No!**

$$\begin{aligned}
\mathbb{E}[Y_n] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(\bar{X}_n - \mu + \mu - X_i)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(\bar{X}_n - \mu)^2 + (X_i - \mu)^2 - 2(\bar{X}_n - \mu)(X_i - \mu)\right] \\
&= \frac{1}{n}(n\frac{\sigma^2}{n}) + \frac{1}{n}n\sigma^2 - 2\,\mathbb{E}\left[\sum_{i=1}^n (\bar{X}_n - \mu)(X_i - \mu)\right] \\
&= \sigma^2\left(\frac{1}{n} + 1\right) - 2\,\mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2\left(1 - \frac{1}{n}\right) \neq \sigma^2
\end{aligned}$$

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 $\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2$

2.3 Decomposing r.v. to compute mean, variance

2.3.1 Decomposing with indicator random variables

Definition 23. *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the **indicator random variable** of an event $A \subset \Omega$ is defined as*

$$\mathrm{I}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Denote $p = \mathbb{P}(A)$, we have $\mathrm{I}_A \sim \text{Ber}(p)$ and $\mathbb{E}[\mathrm{I}_A] = \mathbb{P}(A)$

Example 24. *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 $X = \mathrm{I}_{A_1} + \mathrm{I}_{A_2} + \mathrm{I}_{A_3}$.

So $\mathbb{E}[X] = \mathbb{E}[\mathrm{I}_{A_1}] + \mathbb{E}[\mathrm{I}_{A_2}] + \mathbb{E}[\mathrm{I}_{A_3}] = 0.9$ □

2.3.2 Coupon collector problem

Example 25. *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 $\mathrm{Var}(T_n)$?

Approach

1. Could write the p.m.f. to compute $\mathbb{E}[T_n]$ and $\mathrm{Var}(T_n)$
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

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 $T_2 - T_1 \sim \mathrm{Geom}\left(\frac{n-1}{n}\right)$

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 $W_k \sim \mathrm{Geom}(\frac{n-k}{n})$

4. Finally

$$T_n = T_1 + T_2 - T_1 + \ldots + T_n - T_{n-1} = 1 + W_1 + \ldots + W_{n-1}$$

   So we can get $\mathbb{E}[T_n]$ without computing the p.m.f. of T_n!

5. $T_n = 1 + W_1 + \ldots + W_{n-1}$, $W_k \sim \mathrm{Geom}\left(\frac{n-k}{n}\right)$, how can we compute $\mathrm{Var}(T_n)$?

→ Needs W_k independent!

6. Intuitively yes, why the waiting time for the k^{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 $\mathbb{P}(W_k = a_k | W_j = a_l) = \mathbb{P}(W_k = a_k)$

8. Variance can then be computed as before

Final results We have $\mathbb{E}[W_k] = \frac{1}{p_k} = \frac{n}{n-k}, \mathrm{Var}(W_k) = \frac{1-p_k}{p_k^2} = \frac{k/n}{(n-k)^2/n^2} = \frac{kn}{(n-k)^2}$ so

$$\mathbb{E}[T_n] = 1 + \sum_{k=1}^{n-1} \frac{n}{n-k} = n \cdot \frac{1}{n} + n \sum_{k=1}^{n-1} \frac{1}{n-k} = n \sum_{j=1}^{n} \frac{1}{j}$$

$$\mathrm{Var}(T_n) = \sum_{k=1}^{n-1} \frac{kn}{(n-k)^2} = \sum_{j=1}^{n-1} \frac{n(n-j)}{j^2} = n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} - n \sum_{j=1}^{n-1} \frac{1}{j}$$

□

Example 26. *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. Denote $Y = 5T_5$ the time of the walk, $\mathbb{E}[Y] = 5\,\mathbb{E}[T_5] = 57$, $\mathrm{Var}(Y) = 25\,\mathrm{Var}(T_5) = 629$ $\square$

MATH/STAT395: Probability II	Spring 2020

Covariance, Correlation, Multivariate normal

Sections 8.4 8.5 8.6 of ASV

Instructor: Vincent Roulet — *Teaching Assistant: Zhenman Yuen*

1 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. Here 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

1.1 Definition, interpretation

Definition 1 (Covariance)**.** *Let X, Y be two random variables defined on the same probability space with expectations μ_X, μ_Y. The **covariance** of X and Y is defined by*

$$\mathrm{Cov}(X,Y) \triangleq \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

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 $\mathrm{Cov}(X,Y) > 0$

Example 2. *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 $\mathrm{Cov}(X_4, X_6)$, we should get that $\mathrm{Cov}(X_4, X_6) < 0$, i.e.,*
as X_4 tends to be higher than its mean, X_6 tends to be lower than its mean

Lemma 3. *The **covariance** of X and Y can be formulated as*

$$\mathrm{Cov}(X,Y) \triangleq \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$$

Proof.

$$\begin{aligned}\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] &= \mathbb{E}[XY - \mu_X Y - \mu_Y Y + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_X\,\mathbb{E}[Y] - \mu_Y\,\mathbb{E}[X] + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mu_X\mu_Y\end{aligned}$$

□

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,
   $$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y)dxdy$$
   If X, Y are discrete (and integer valued),
   $$\mathrm{Cov}(X,Y) = \sum_{k=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty}(k-\mu_X)(j-\mu_Y)\,\mathbb{P}(X=k, Y=j)$$

Terminology We say that two r.v. X, Y are

1. positively correlated if $\mathrm{Cov}(X,Y) > 0$
2. negatively correlated if $\mathrm{Cov}(X,Y) < 0$
3. uncorrelated if $\mathrm{Cov}(X,Y) = 0$

Example 4. *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 in the following

Example 5. *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* $\mathrm{Cov}(X,Y)$.

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)
   $\mathbb{E}[X] = \iint_T x f_{X,Y}(x,y)dxdy = \int_0^1\int_0^{1-y} 2x dxdy = \frac{1}{3}$
   By symmetry, $\mathbb{E}[Y] = \frac{1}{3}$ and
   $\mathbb{E}[XY] = \iint_T xy f_{X,Y}(x,y)dxdy = \int_0^1\int_0^{1-y} 2xy dxdy = \frac{1}{12}$
   So $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \frac{1}{12} - \frac{1}{3}\cdot\frac{1}{3} = -\frac{1}{36} < 0$

□

1.1.1 Covariance of indicator random variables

Lemma 6. *Let A, B be two events on a proba. space $\Omega, \mathcal{F}, \mathbb{P}$.*

$$\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) = \mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B)$$

If $\mathbb{P}(B) > 0$, $\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$

Proof.

$$\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) = \mathbb{E}[\mathrm{I}_A\,\mathrm{I}_B] - \mathbb{E}[\mathrm{I}_A]\,\mathbb{E}[\mathrm{I}_B]$$

$$(\mathrm{I}_A\,\mathrm{I}_B)(\omega) = \mathrm{I}_A(\omega)\,\mathrm{I}_B(\omega) = \begin{cases} 1 & \text{if } \omega \in A \text{ and } \omega \in B \\ 0 & \text{otherwise} \end{cases}. \quad \text{Thus } \mathrm{I}_A\,\mathrm{I}_B = \mathrm{I}_{A\cap B}$$

$$\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) = \mathbb{E}[\mathrm{I}_{A\cap B}] - \mathbb{E}[\mathrm{I}_A]\,\mathbb{E}[\mathrm{I}_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(B)(\mathbb{P}(A|B) - \mathbb{P}(A))$$

provided that $\mathbb{P}(B) > 0$ (for the last equality). □

Interpretation of covariance for indicator random variables

1. $\mathrm{I}_A, \mathrm{I}_B$ are **positively** correlated $(\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) > 0) \Leftrightarrow \mathbb{P}(A|B) - \mathbb{P}(A) > 0$

$\rightarrow$ *the occurrence of B **increases** the chances of A.*

2. $\mathrm{I}_A, \mathrm{I}_B$ are **negatively** correlated $(\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) < 0) \Leftrightarrow \mathbb{P}(A|B) - \mathbb{P}(A) < 0$

$\rightarrow$ *the occurrence of B **decreases**the chances of A*

3. $\mathrm{I}_A, \mathrm{I}_B$ are uncorrelated $(\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) = 0) \Leftrightarrow A$, B are independent.

The covariance of **indicator random variables** is a measure of the independence of the corresponding events.

Example 7. *Let S be the sum of two fair dice X_1 and X_2.*
Are $\mathrm{I}_{\{S>10\}}, \mathrm{I}_{\{X_2=6\}}$ positively, negatively correlated or uncorrelated?

Solution.

$$\mathrm{Cov}(\mathrm{I}_{S>10}, \mathrm{I}_{\{X_2=6\}}) = \mathbb{P}(S > 10, X_2 = 6) - \mathbb{P}(S > 10)\,\mathbb{P}(X_2 = 6) = \frac{2}{36} - \frac{3}{36}\frac{1}{6} > 0$$

So $\mathrm{I}_{\{S>10\}}, \mathrm{I}_{\{X_2=6\}}$ are positively correlated. □

1.1.2 Covariance and independence

Theorem 8. *Let X, Y be two random variables,*

$$X, Y \textit{ are independent} \Rightarrow \mathrm{Cov}(X, Y) = 0$$

*but the the converse **does not hold in general***

Proof. 1. Let X, Y be two independent r.v., then

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

2. Counter example: take $X \sim \mathrm{Unif}(\{-1, 0, 1\})$ and $Y = X^2$, then

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\,\mathbb{E}[X^2]$$

We have $X^3 = X$ and $\mathbb{E}[X] = 0$ therefore $\mathrm{Cov}(X, Y) = 0$
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.

□

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 below)

1.1.3 Properties of Covariance

Example 9. *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* $\mathrm{Cov}(X_4, X_6)$ *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*

Lemma 10 (Properties of covariance 1)**.** *Provided that the covariances defined below are well defined,*

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
2. $\mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X, Y)$ *for any* $a, b \in \mathbb{R}$

Proof. 1. clear from definition

2.

$$\begin{aligned}\mathrm{Cov}(aX+b, Y) &= \mathbb{E}[(aX+b)Y] - \mathbb{E}[aX+b]\,\mathbb{E}[Y] \\ &= a\,\mathbb{E}[XY] + b\,\mathbb{E}[Y] - a\,\mathbb{E}[X]\,\mathbb{E}[Y] - b\,\mathbb{E}[Y] \\ &= a\,(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]) = a\,\mathrm{Cov}(X, Y)\end{aligned}$$

□

Lemma 11 (Bilinearity of covariance)**.** *Provided that the covariances defined below are well defined, For* $X_1, \ldots, X_m$, $Y_1, \ldots Y_n$ *r.v. and* $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$,

$$\mathrm{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \,\mathrm{Cov}(X_i, Y_j)$$

Proof. $\mu_{X_i} = \mathbb{E}[X_i]$, $\mu_{Y_i} = \mathbb{E}[Y_i]$ so $\mathbb{E}\left[\sum_{i=1}^m X_i\right] = \sum_{i=1}^m \mu_{X_i}$, $\mathbb{E}\left[\sum_{j=1}^n Y_j\right] = \sum_{j=1}^n \mu_{Y_j}$

$$\begin{aligned}\mathrm{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) &= \mathbb{E}\left[\left(\sum_{i=1}^m a_i X_i - \sum_{i=1}^m a_i \mu_{X_i}\right)\left(\sum_{j=1}^n b_j Y_j - \sum_{j=1}^n b_j \mu_{Y_j}\right)\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^m a_i (X_i - \mu_{X_i})\right)\left(\sum_{j=1}^n b_j (Y_j - \mu_{Y_j})\right)\right] \\ &= \mathbb{E}\left[\sum_{i=1}^m \sum_{j=1}^n a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\right] \\ &= \sum_{i=1}^m \sum_{j=1}^n a_i b_j \,\mathbb{E}\left[(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\right] = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \,\mathrm{Cov}(X_i, Y_j)\end{aligned}$$

□

Example 12. *Let* $(X_1,\dots,X_r) \sim \mathrm{Multinom}(n,r,p_1,\dots,p_r)$ *with* $p_1+\ldots+p_r=1$*, i.e.*

1. *An experiment has* r *outcomes*
2. *The* i^{th} *outcome has proba* p_i
3. X_i *is the number of times outcome* i *occurs when performing* n *independent trials*

Find $\mathrm{Cov}(X_i,X_j)$ *for* $i,j \in \{1,\dots n\}$

Solution. 1. **Idea:** Decompose X_i and X_j as sum of simple r.v., i.e. $X_i = \sum_{k=1}^n \mathrm{I}_{k,i}$ where

$$\mathrm{I}_{k,i} = \begin{cases} 1 & \text{if trial } k \text{ gives outcome i} \\ 0 & \text{if trial } k \text{ gives an outcome other than i} \end{cases}$$

2. $\mathrm{I}_{k,i} \sim \mathrm{Ber}(p_i)$ so $X_i \sim \mathrm{Bin}(n,p_i)$ and $\mathrm{Cov}(X_i,X_i) = \mathrm{Var}(X_i) = np_i(1-p_i)$
3. For $i \neq j$, by bilinearity of the covariance,

$$\mathrm{Cov}(X_i,X_j) = \mathrm{Cov}\left(\sum_{k=1}^n \mathrm{I}_{k,i}, \sum_{\ell=1}^n \mathrm{I}_{\ell,j}\right) = \sum_{k=1}^n\sum_{\ell=1}^n \mathrm{Cov}(\mathrm{I}_{k,i},\mathrm{I}_{\ell,j}) = \sum_{k=1}^n \mathrm{Cov}(\mathrm{I}_{k,i},\mathrm{I}_{k,j})$$

   using that if $k \neq l$, $\mathrm{I}_{k,i}, \mathrm{I}_{\ell,j}$ are independent by definition.
   Since $i \neq j$, $\mathrm{I}_{k,i}\,\mathrm{I}_{k,j} = 0$, because on trial k both outcomes cannot occur

$$\mathrm{Cov}(\mathrm{I}_{k,i},\mathrm{I}_{k,j}) = \mathbb{E}[\mathrm{I}_{k,i}\,\mathrm{I}_{k,j}] - \mathbb{E}[\mathrm{I}_{k,i}]\,\mathbb{E}[\mathrm{I}_{k,j}] = 0 - p_ip_j$$

   Therefore $\mathrm{Cov}(X_i,X_j) = -np_ip_j < 0$,
   $\rightarrow$ the more often i occurs, the fewer opportunities for outcome j

$\square$

1.2 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 ?

$\rightarrow$ Needs to take into account the interactions btw the elements of the sum, i.e. their covariance!

Corollary 13. *Let* $X_1,\dots,X_n$ *be* n *r.v. with finite variance and covariances (between each pair)*

$$\mathrm{Var}(X_1+\ldots+X_n) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{1\le i<j\le n} \mathrm{Cov}(X_i,X_j)$$

Proof. $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) = \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i,X_j)$
Then identify $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i,X_i)$ in the sum and simplify the rest of the sum. Namely

$$\begin{aligned}\sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i,X_j) &= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{\substack{i,j=1\\ i\neq j}}^n \mathrm{Cov}(X_i,X_j) \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{1\le i<j\le n} \mathrm{Cov}(X_i,X_j)\end{aligned}$$

where we used in the last equality that $\mathrm{Cov}(X_i,X_j) = \mathrm{Cov}(X_j,X_i)$. $\square$

Corollary 14. *Let* $X_1,\dots X_n$ *be* n *uncorrelated r.v.* ($\mathrm{Cov}(X_i,X_j)$=0 *for* $i,j\in\{1,\dots,n\}, i\neq j$)

$$\mathrm{Var}(X_1+\ldots+X_n) = \sum_{i=1}^n \mathrm{Var}(X_i)$$

Remark: We retrieve a previous result 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

Example 15. *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 $X = \mathrm{I}_A + \mathrm{I}_B$ 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 $\mathrm{Var}(X)$.

Solution. $\mathrm{I}_A \sim \mathrm{Ber}(p_A)$ with $p_A = 0.5$, $\mathrm{I}_B \sim \mathrm{Ber}(p_B)$ with $p_B = 0.2$

Using

$$\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) = \mathbb{E}[\mathrm{I}_A \, \mathrm{I}_B] - \mathbb{E}[\mathrm{I}_A] \, \mathbb{E}[\mathrm{I}_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A) \, \mathbb{P}(B)$$

We get

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(\mathrm{I}_A) + \mathrm{Var}(\mathrm{I}_B) + 2\,\mathrm{Cov}(\mathrm{I}_A, \mathrm{I}_B) \\ &= p_A(1 - p_A) + p_B(1 - p_B) + 2(\mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B)) \\ &= 0.25 + 0.16 + 2(0.1 - 0.1) = 0.41 \end{aligned}$$

□

1.3 Correlation

Motivation

- We said that $\mathrm{Cov}(X, Y)$ could be a good proxy of dependence
- Yet, by bilinearity, $\mathrm{Cov}(10X, 7Y) = 70\,\mathrm{Cov}(X, Y)$
- 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 a scaling invariant measure: correlation!

Definition 16 (Correlation)**.** *The **correlation** (or **correlation coefficient**) of two r.v. X, Y with positive finite variances is*

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

It is sometimes denoted $\rho(X, Y)$ or $\rho_{X,Y}$.

Lemma 17 (Scaling invariance)**.** *Let X, Y be two r.v. with positive finite variances and $a, b \in \mathbb{R}$, $a \neq 0$*

$$\mathrm{Corr}(aX + b, Y) = \frac{a}{|a|}\,\mathrm{Corr}(X, Y)$$

Proof. $\mathrm{Corr}(aX + b, Y) = \frac{\mathrm{Cov}(aX+b,Y)}{\sqrt{\mathrm{Var}(aX+b)}\sqrt{\mathrm{Var}(Y)}} = \frac{a\,\mathrm{Cov}(X,Y)}{\sqrt{a^2\,\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{a}{|a|}\,\mathrm{Corr}(X, Y)$ □

Lemma 18 (Properties of correlation 1)**.** *Let X, Y be two r.v. with positive finite variances. Then* $-1 \leq \mathrm{Corr}(X, Y) \leq 1$

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 $\mu_X = \mathbb{E}[X], \sigma_X^2 = \mathrm{Var}(X), \mu_Y = \mathbb{E}[Y], \sigma_Y^2 = \mathrm{Var}(Y)$, s.t.

$$\mathbb{E}[\tilde{X}] = 0, \qquad \mathrm{Var}(\tilde{X}) = \mathbb{E}[\tilde{X}^2] = \mathbb{E}\left[\frac{(X - \mu_X)^2}{\sigma_X^2}\right] = 1$$

Same for $\tilde{Y}$ and finally

$$\mathbb{E}[\tilde{X}\tilde{Y}] = \mathbb{E}\left[\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right] = \frac{\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y} = \mathrm{Corr}(X,Y)$$

Proof. of the lemma

$$0 \leq \mathbb{E}[(\tilde{X}-\tilde{Y})^2] = \mathbb{E}[\tilde{X}^2] + \mathbb{E}[\tilde{Y}^2] - 2\,\mathbb{E}[\tilde{X}\tilde{Y}] = 2(1-\mathrm{Corr}(X,Y))$$

Therefore $1-\mathrm{Corr}(X,Y) \geq 0$, i.e. $\mathrm{Corr}(X,Y) \leq 1$.
Similarly $0 \leq \mathbb{E}[(\tilde{X}+\tilde{Y})^2] = 2(1+\mathrm{Corr}(X,Y))$ so $\mathrm{Corr}(X,Y) \geq -1$ □

Lemma 19 (Properties of correlation 2)**.** *Let X, Y be two r.v. with positive finite variances.*

1. $\mathrm{Corr}(X,Y) = 1 \iff \exists a > 0, b \in \mathbb{R},$ *s.t.* $Y = aX + b$
2. $\mathrm{Corr}(X,Y) = -1 \iff \exists a < 0, b \in \mathbb{R},$ *s.t.* $Y = aX + b$

and naturally $\mathrm{Corr}(X,X) = 1$

Proof. of 1. (proof of 2. is analogous)
If $Y = aX + b$ then by scaling invariance $\mathrm{Corr}(X,Y) = \frac{a}{|a|}\mathrm{Corr}(X,X) = 1$
Assume $\mathrm{Corr}(X,Y) = 1$, denote $\tilde{X} = \frac{X-\mu_X}{\sigma_X}, \tilde{Y} = \frac{Y-\mu_Y}{\sigma_Y}$ and $Z = \tilde{X} - \tilde{Y}$

$$\mathbb{E}[Z] = 0, \qquad \mathrm{Var}(Z) = \mathbb{E}[(\tilde{X}-\tilde{Y})^2] = 2(1-\mathrm{Corr}(X,Y)) = 0$$

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 $a = \frac{\sigma_Y}{\sigma_X} > 0$. □

Remark 20. *Let X, Y be two r.v. such that* $\mathrm{Corr}(X,Y) = 1$.
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

$$\mathbb{P}(Y = aX + b) = \int_{-\infty}^{+\infty}\int_{ax+b}^{ax+b} f_{X,Y}(x,y)dydx = 0$$

In this case we would say that the random vector (X,Y) is **degenerated**
(similarly as when $\mathrm{Var}(X) = 0$ for a single r.v.) □

Example 21. 1. *Roll a die 10 times, denote X_1, X_2 the number of* 1 *and* 2 *that you get.*

   (a) *Compute* $\mathrm{Corr}(X_1, X_2)$

2. *Flip a coin 10 times, denote X_1, X_2 the number of tails and heads respectively.*

   (a) *Compute* $\mathrm{Corr}(X_1, X_2)$
   (b) *How could you have found it ?*

Solution. 1. We give directly the correlation of $(X_1,\ldots,X_n) \sim \mathrm{Multinom}(n, r, p_1, \ldots, p_r)$ (here $n = 10$, $r = 6$, $p_i = 1/6$)
We saw that

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} np_i(1-p_i) & \text{if } i = j \\ -np_ip_j & \text{if } i \neq j \end{cases}$$

So for $i \neq j$

$$\mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(X_j)}\mathrm{Var}(X_j)} = \frac{-np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_j(1-p_j)}} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}$$

So here $\mathrm{Corr}(X_i, X_j) = -\sqrt{1/25} = -1/5 = -0.2$

2. In the case $r = 2$ s.t. $p_1 = 1 - p_2$,

$$\mathrm{Corr}(X_1, X_2) = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}} = -1$$

That reflects that $X_2 = n - X_1$ for binomial.

□

2 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!

2.1 Mean vector, covariance matrix

Definition 22 (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*

2.1.1 Mean vector

Definition 23 (Mean vector). *Let $\mathbf{X} = (X_1, ..., X_n)^\top$ be a random vector, its mean vector is defined as*

$$\boldsymbol{\mu}_{\boldsymbol{X}} \triangleq \mathbb{E}[\boldsymbol{X}] \triangleq \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

Example 24. *Let $\boldsymbol{X} = (X_1, \dots, X_n) \sim \mathrm{Multinom}(n, r, p_1, \dots, p_r)$, what is $\boldsymbol{\mu}_{\boldsymbol{X}}$?*

Solution. $X_i \sim \mathrm{Bin}(n, p_i)$ (use decomposition seen previously)

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \begin{pmatrix} np_1 \\ \vdots \\ np_n \end{pmatrix}$$

□

Lemma 25. *Let $\mathbf{X} = (X_1, ..., X_n)^\top$, $A = (A_{ij})_{\substack{i=1,\dots,p \\ j=1,\dots,n}} \in \mathbb{R}^{p\times n}$ and $b = (b_i)_{i=1}^p \in \mathbb{R}^p$, then*

$$\mathbb{E}[A\boldsymbol{X} + b] = A\,\mathbb{E}[\boldsymbol{X}] + b$$

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 $\mathbb{E}[A\mathbf{X} + b] = A\,\mathbb{E}[\mathbf{X}] + b$ □

2.1.2 Covariance Matrix

Definition 26. *Let $\mathbf{X} = (X_1, ..., X_n)^\top$ be a random vector its* ***covariance matrix*** *is defined as*

$$S_{\mathbf{X}} = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \dots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \dots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} = (\mathrm{Cov}(X_i, X_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

1. *$S_{\mathbf{X}}$ is symmetric, i.e. $(S_{\mathbf{X}})_{ij} = (S_{\mathbf{X}})_{ji} = \mathrm{Cov}(X_i, X_j)$*
2. *The diagonal of $S_{\mathbf{X}}$ represents the variances $(S_{\mathbf{X}})_{ii} = \mathrm{Var}(X_i)$*

Example 27. *Let X, Y be two random variables, their covariance matrix is*

$$S = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathrm{Var}(Y) \end{pmatrix}$$

Lemma 28. *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 material at the end of the lecture notes) □

2.2 Multivariate Normal Random variables

Definition 29 (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?

Property 30. *A standard normal random vector $\mathbf{X} = (X_1, \dots, X_n)^\top$ satisfies*

$$\boldsymbol{\mu}_{\mathbf{X}} = 0_n \triangleq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \qquad S_{\mathbf{X}} = \mathrm{I}_n \triangleq \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

It is denoted $\mathbf{X} \sim \mathcal{N}(0_n, \mathrm{I}_n)$

Definition 31 (Multivariate Normal Distribution)**.** *A random vector $\mathbf{X} = (X_1, \dots, X_n)^\top$ is a* ***normal random vector***
if there exist $\boldsymbol{\mu} \in \mathbb{R}^n, A \in \mathbb{R}^{p \times n}, \mathbf{Z} \sim \mathcal{N}(0_n, \mathrm{I}_n)$ s.t.

$$\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$$

Question What are the mean and covariance matrix of a standard normal random vector?

Property 32. *A normal random vector $\mathbf{X} = (X_1, \dots, X_n)^\top$ satisfies as defined above*

$$\boldsymbol{\mu}_{\mathbf{X}} = \boldsymbol{\mu} \qquad S_{\mathbf{X}} = AA^\top$$

It is denoted $\mathbf{X} \sim \mathcal{N}(\mu, S)$ with $S = S_{\mathbf{X}}$.

Definition 33. *A normal random vector* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathcal{N}(\mu, S)$ *with **invertible covariance matrix** has a joint p.d.f.*

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Proof. (See additional material for a sketch of proof) $\square$

Lemma 34. *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathcal{N}(\mu, S)$ *with finite positive marginal variances* ($0 < \mathrm{Var}(X_i) < +\infty$),

$$\mathrm{Cov}(X_i, X_j) = 0 \textit{ for all } i \neq j \iff X_1, \ldots, X_n \textit{ are independent}$$

Proof. If $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, then

$$S = \begin{pmatrix} \sigma^2_{X_1} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \sigma^2_{X_n} \end{pmatrix} \qquad S^{-1} = \begin{pmatrix} \sigma^{-2}_{X_1} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \sigma^{-2}_{X_n} \end{pmatrix}$$

So

$$\begin{aligned} f_{\mathbf{X}}(x_1, \ldots, x_n) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\ &= \frac{1}{(2\pi)^{n/2}\sigma_{X_1}\ldots\sigma_{X_n}} e^{-\left(\frac{(x_1-\mu_1)^2}{2\sigma^2_{X_1}} + \ldots + \frac{(x_n-\mu_n)^2}{2\sigma^2_{X_n}}\right)} = f_1(x_1)\ldots f_n(x_n) \end{aligned}$$

The joint p.d.f. factorizes in functions of each r.v. (that can be shown to be the marginals) so $(X_1, \ldots X_n)$ are independent. $\square$

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

3 Additional material*

3.1 Characterization of independence*

The key lemma to show that Independence $\Rightarrow \mathrm{Cov}(X, Y) = 0$ is

Lemma 35. *If X, Y are two independent r.v. then for any $g, h : \mathbb{R} \to \mathbb{R}$ 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 36. *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. □

Covariance only checks for one particular choice of h and g.
It is not sufficient.

3.2 Multivariate Normal Distribution*

3.2.1 Covariance Matrix*

Definition 37. *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 38. *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 □

Lemma 39. *Let $\mathbf{X} = (X_1, ..., X_n)^\top$ be a random vector its **covariance matrix** reads*

$$S_{\mathbf{X}} = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top\right]$$

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

$$(\mathbb{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top])_{ij} = \mathrm{Cov}(X_i, X_j)$$

which gives the result. □

Lemma 40. *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., $(A^\top)_{ij} = A_{ji}$

Proof.

$$\begin{aligned}
S_{\mathbf{Y}} &= \mathbb{E}[(A\mathbf{X} + b - (A\,\mathbb{E}[\mathbf{X}] + b))(A\mathbf{X} + b - (A\,\mathbb{E}[\mathbf{X}] + b))^\top] \\
&= \mathbb{E}[(A(\mathbf{X} - \mathbb{E}[\mathbf{X}]))(A(\mathbf{X} - \mathbb{E}[\mathbf{X}]))^\top] \\
&= A\,\mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\right]A^\top \\
&= AS_{\mathbf{X}}A^\top
\end{aligned}$$

□

3.2.2 Multivariate Normal distribution p.d.f.*

Definition 41. *A normal random vector* $\mathbf{X} = (X_1, \ldots, X_n)^\top \sim \mathcal{N}(\mu, S)$ *with* ***invertible covariance matrix*** *has a joint p.d.f.*

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(S)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

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 $S = AA^\top$

□

MATH/STAT395: Probability II Spring 2020

Moment Generating Functions, Concentration inequalities

Sections 5.1 8.3 9.1 of ASV

Instructor: Vincent Roulet *Teaching Assistant: Zhenman Yuen*

1 Moment Generating 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 to describe a r.v.

   (rather than using p.m.f/p.d.f or c.d.f.)

1.1 Definition

Definition 1. *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

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx = \mathcal{L}(f)(-t) \qquad \phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x)dx = \mathcal{F}(f)(-t)$$

where $\mathcal{L}(f), \mathcal{F}(f)$ are the *Laplace* and *Fourier* transforms of f

$\rightarrow$ 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)

Example 2. *Let $X \sim \mathrm{Poisson}(\lambda)$, for $\lambda > 0$. Compute $M_X(t)$.*

Solution.

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{+\infty} e^{tk}\,\mathbb{P}(X = k) = \sum_{k=0}^{+\infty} e^{tk}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{+\infty}\frac{(e^t\lambda)^k}{k!} = e^{-\lambda}e^{\lambda e^t}$$

$$M_X(t) = e^{\lambda(e^t-1)}$$

□

Example 3. *1. Let $X \sim \mathcal{N}(0,1)$. Compute $M_X(t)$*

2. Let $X \sim \mathcal{N}(\mu,\sigma^2)$. Compute $M_X(t)$

Solution. 1.

$$\mathbb{E}[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx} e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{t^2/2} e^{-(x-t)^2/2}$$
$$= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-t)^2/2} = e^{t^2/2}$$

2. $X = \sigma Z + \mu$ for $Z \sim \mathcal{N}(0,1)$

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(\sigma Z+\mu)}] = e^{t\mu}\, \mathbb{E}[e^{t\sigma Z}] = e^{\mu t + \sigma^2 t^2/2}$$

□

Example 4. *Let $X \sim \mathbb{E}(\lambda)$, $\lambda > 0$. Compute $M_X(t)$.*

Solution.

$$\mathbb{E}[e^{tX}] = \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} dx = \lim_{b \to +\infty} \lambda \int_0^b e^{(t-\lambda)x} dx$$

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 $t = \lambda$, $\mathbb{E}[e^{tX}] = \lambda \lim\limits_{b\to+\infty} \int_0^b dx = \lambda \lim\limits_{b\to+\infty} b = +\infty$
2. if $t \neq \lambda$,

$$\mathbb{E}[e^{tX}] = \lambda \lim_{b\to+\infty} \frac{e^{(t-\lambda)b} - 1}{t - \lambda} = \begin{cases} +\infty & \text{if } t > \lambda \\ \frac{\lambda}{\lambda - t} & \text{otherwise} \end{cases}$$

So

$$M_X(t) == \begin{cases} +\infty & \text{if } t \geq \lambda \\ \frac{\lambda}{\lambda - t} & \text{otherwise} \end{cases}$$

□

1.2 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)

$$M_X(t) = \sum_{k\in X(\Omega)} e^{tk}\, \mathbb{P}(X = k), \qquad M'_X(t) = \sum_{k\in X(\Omega)} k e^{tk}\, \mathbb{P}(X = k)$$
$$M'_X(0) = \sum_{k\in X(\Omega)} k\, \mathbb{P}(X = k) = \mathbb{E}[X]$$

More generally, $M'_X(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}[\frac{d}{dt}e^{tX}] = \mathbb{E}[Xe^{tX}]$, s.t. $M'_X(0) = \mathbb{E}[X]$.

Lemma 5. *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^{th} derivative of the moment generating function on 0

Example 6. *Let $X \sim \mathrm{Ber}(p)$ 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 $\mathbb{E}[X^n] = \mathbb{E}[X] = p$

$\square$

Example 7. *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, λ) for any $a < 0$, i.e. an open interval around 0. We can compute for $t < \lambda$

$$M_X'(t) = \lambda(\lambda - t)^{-2}, M_X''(t) = 2\lambda(\lambda - t)^3, \ldots, M_X^{(n)}(t) = n!\lambda(\lambda - t)^{-n-1}$$

So $\mathbb{E}[X^n] = M_X^{(n)}(0) = n!\lambda^{-n}$

$\square$

1.3 Characterization of distributions by moment generating function

Definition 8 (Equality in distribution (Reminder))**.** *Two r.v. X, Y are **equal in distribution**, denoted $X \overset{d}{=} Y$ if*

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

Theorem 9. *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 $M_X(t) = M_X(t)$ then $X \overset{d}{=} Y$,*

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 a 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 $X \overset{d}{=} Y$

Example 10. *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 $X \overset{d}{=} Y$, i.e. X has the same distribution as Y.

$\square$

1.4 Moment generating function 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 11. *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.

$$M_{X_1+\dots+X_n}(t) = \mathbb{E}[e^{t(X_1+\dots+X_n)}] = \mathbb{E}[e^{tX_1} \dots e^{tX_n}] = \mathbb{E}[e^{tX_1}] \dots \mathbb{E}[e^{tX_n}] = M_{X_1}(t) \dots M_{X_n}(t)$$

□

Example 12. *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 $X + Y \overset{d}{=} Z \sim \text{Poisson}(\lambda+\mu)$

□

Example 13. *Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ independent (recall that $M_X(t) = e^{\mu_1 t + \frac{\sigma_1 t^2}{2}}$)*

1. *Compute $M_{X+Y}(t)$*
2. *What can you conclude about the distribution of $X+Y$?*

Solution.

1. $M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\mu_1 t + \frac{\sigma_1 t^2}{2}} e^{\mu_2 t + \frac{\sigma_2 t^2}{2}} = e^{(\mu_1+\mu_2)t + \frac{(\sigma_1^2+\sigma_2^2)t}{2}}$
2. $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

□

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?

2.1 Monotonicity of Expectation

Theorem 14 (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 $\mathbb{P}(X \le Y) = 1$ then $\mathbb{E}[X] \le \mathbb{E}[Y]$.*

Proof. Denote $Z = Y - X$ s.t. $\mathbb{P}(Z \ge 0) = 1$

1. (Discrete case) If Z is discrete, for any $k < 0$, $0 \le \mathbb{P}(Z = k) \le \mathbb{P}(Z < 0) = 0$ so

$$\mathbb{E}[Z] = \sum_{k \in Z(\Omega)} k \, \mathbb{P}(Z = k) \ge 0$$

2. (Continuous case) If Z is continuous, then (as in exercise 4.2 of homework 1)

$$\int_{-\infty}^{0} z f_Z(z)dz = -\int_{-\infty}^{0}\int_{z}^{0} f_Z(z)dtdz = -\iint_{z\le t\le 0, z\le 0} f_Z(z)dtdz = -\int_{-\infty}^{0}\int_{-\infty}^{t} f_Z(z)dzdt$$

So $\int_{-\infty}^{0} z f_Z(z)dz {=} -\int_{-\infty}^{0}\mathbb{P}(Z\le t){=}0$ since $0{\le}\mathbb{P}(Z\le t){\le}\mathbb{P}(Z\le 0){=}0$ for all $t{\le}0$.

Therefore $\mathbb{E}[Z] = \int_{-\infty}^{0} z f_Z(z)dz + \int_{0}^{+\infty} z f_Z(z)dz \ge 0$

3. So in both cases $\mathbb{E}[Z] = \mathbb{E}[Y - X] \ge 0$, i.e. $\mathbb{E}[X] \le \mathbb{E}[Y]$

□

2.2 Markov's Inequality

Question: What can be said about the proba. of X if we know $\mathbb{E}[X]$?

Theorem 15 (Markov inequality)**.** *Let X be a non-negative r.v. with finite mean then for any $c > 0$,*

$$\mathbb{P}(X \ge c) \le \frac{\mathbb{E}[X]}{c}$$

Proof. Define the indicator random variable $\mathrm{I}_{X\ge c}$. We have

$$X \ge c\,\mathrm{I}_{X\ge c}$$

1. when $X \ge c$ the inequality reads $X \ge c$,
2. when $X \le c$ the inequality reads $X \ge 0$, true by assumption

Now applying previous theorem,

$$\mathbb{E}[X] \ge c\,\mathbb{E}[\mathrm{I}_{X\ge c}] = c\,\mathbb{P}(X \ge c)$$

□

Example 16. *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.

$$\mathbb{P}(X \ge 1400) \le \frac{\mathbb{E}[X]}{1400} = \frac{1000}{1400} = 5/7 \approx 0.71 < 0.8 \qquad \to \text{so the answer is no}$$

□

Example 17. *Let $X \sim \mathrm{Ber}(p)$, $p \in (0,1)$*

1. *What is $\mathbb{P}(X \ge 0.01)$?*
2. *What gives Markov inequality?*

Solution.

1. Clearly $\mathbb{P}(X \ge 0.01) = \mathbb{P}(X = 1) = p$
2. Markov's inequality gives

$$\mathbb{P}(X \ge 0.01) \le \frac{\mathbb{E}[X]}{0.01} = 100p$$

Here Markov's inequality is useless (we may even have $100p \ge 1$ s.t. it is even less informative than knowing that $\mathbb{P}(X \ge 0.01) \le 1$)

□

2.3 Chebyshev's inequality

Question What can be said about the proba. of X if we know $\mathbb{E}[X]$ and $\mathrm{Var}(X)$?

Theorem 18 (Chebyshev's Inequality)**.** *Let X be a r.v. with finite mean μ and finite variance σ^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 $\mathbb{P}(X \geq \mu + c)$, $\mathbb{P}(X \leq \mu - c)$

Example 19. *A donut vendor sells on average* 1000 *donuts per day with a variance of* 200*. Provide a bound on*

1. *the proba. that there will be between* 950 *and* 1050 *customers tomorrow*
2. *the proba. that there will be at least* 1400 *customers tomorrow*

Solution.

1. $\mathbb{P}(950 < X < 1050) = \mathbb{P}(|X-1000| < 50) = 1 - \mathbb{P}(|X-1000| \geq 50)$ By Chebyshev's inequality,

$$\mathbb{P}(|X-1000| \geq 50) = \mathbb{P}(|X-\mathbb{E}[X]| \geq 50) \leq \frac{\mathrm{Var}(X)}{50^2} = \frac{200}{50^2} = \frac{2}{25} = 0.08$$

So $\mathbb{P}(950 < X < 1050) \geq 1 - 0.08 = 0.92$

2. $\mathbb{P}(X \geq 1400) = \mathbb{P}(X - 1000 \geq 400) \leq \frac{200}{400^2} = \frac{1}{800} = 0.00125$

□

3 Additional material

3.1 Generalization of Markov's inequality

Question: What if we know more moments? How can proba of X be bounded?

Lemma 20. *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. $\square$

Corollary 21 (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*

$$\mathbb{P}(X \geq c) \leq e^{-tc}\, \mathbb{E}[e^{tX}] \qquad \textit{for all } t \in (0, \theta]$$

Proof. Apply above lemma for $f(x) = e^{tx}$ $\square$

Example 22. *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

$$\mathbb{P}(X \geq c) \leq e^{-c^2/2}$$

$\square$

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 23 (Sub-Gaussian distribution)**.** *The proba. distribution of a r.v. X is called **sub-Gaussian** if there are positive constants C, ν, 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 $x \to +\infty$
that $\int_c^{+\infty} f_X(x)d \leq Ce^{-\nu c^2/2}$

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, ...