Artificial Intelligence 2: Revision for Probability

Yunwen Lei

School of Computer Science
University of Birmingham

Outline

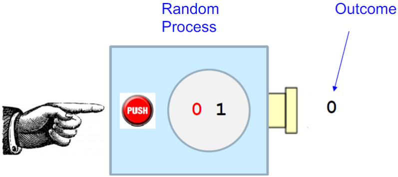
1. Sample Space and Events
2. Probability
3. Conditional Probability
4. Bayes' Formula and Independence
5. Discrete Random Variable
6. Continuous Random Variable
7. Joint Distributions
8. Descriptive Statistics

Sample Space and Events

Random Experiments

Random Experiments

A Random Experiment is a process that produces uncertain outcomes from a well-defined set of possible outcomes.



Sample Space

The set of all possible outcomes of an experiment is called the Sample Space and is denoted by Ω. Any individual outcome is called a Sample Point.

Event

An Event is any subset of the Sample Space. An event A is said to have occurred if the outcome of the random experiment is a member of A.

Examples

Rolling two dice

- Sample space: the set of all possible outcomes (36 sample points)

$$\Omega = \Big\{(1,1),(1,2),(1,3),(1,4),\dots,(6,3),(6,4),(6,5),(6,6)\Big\}$$

- An event of "double"

$$A = \big\{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\big\}, \quad |A| = 6$$

- Another event of sum being 4

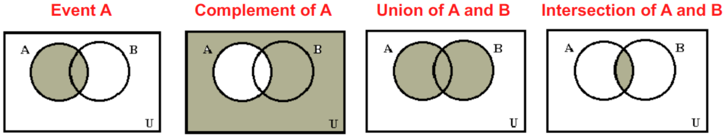
$$A = \{(1,3),(2,2),(3,1)\}, \quad |A| = 3$$

Lifetime of a device (measured in years)

- Sample space: $\Omega = [0,\infty)$
- An event $A = \big\{\text{device lasts for at least 5 years}\big\} = [5,\infty)$
- Another event $A = \big\{\text{device is dead by its } 6^{\text{th}} \text{ birthday}\big\} = [0,6)$

Operation on Events

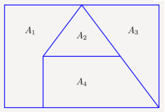
- The complement A^c of an event A is the event that A does not occur
- The union $A \cup B$ of two events A and B is the event that either A or B or both occurs
- The intersection $A \cap B$ of two events A and B is the event that both A and B occur



Definition: Partition

A collection of sets $\{A_1, \ldots, A_n\}$ is a partition to the universal set Ω if it satisfies the following conditions:

- (non-overlap) $\{A_1, \ldots, A_n\}$ is disjoint
- (decompose) $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$



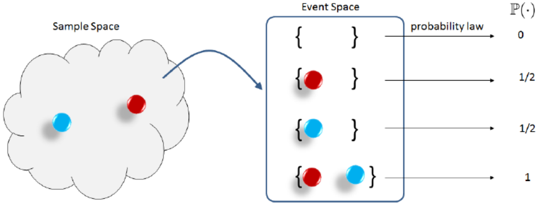
Probability

Probability Law

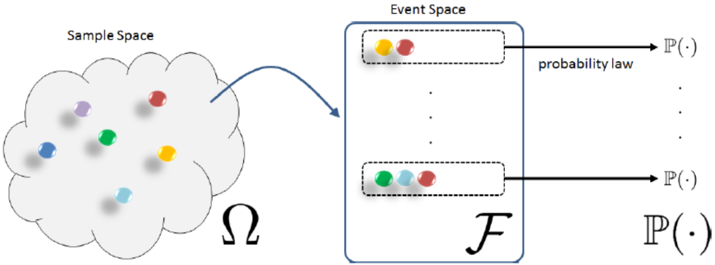
A Probability Law is a function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ that maps an event A to a real number in $[0, 1]$. It satisfies the following Kolmogorov axioms:

- Non-negativity: for any event $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$ (nonnegative area of event)
- Unit measure: $\mathbb{P}(\Omega) = 1$ (the area of the whole sample space is 1)
- Additivity of disjoint events: if $A_1, A_2, \ldots$ is a collection of disjoint events then (if two regions do not overlap, then the area of the combined region is the sum of the area of each region)

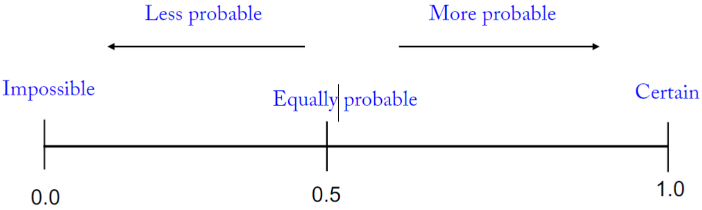
$$\mathbb{P}\big(\cup_{i=1}^{\infty} A_i \big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$



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



So we measure the probability of events on a real-number scale from 0 to 1:

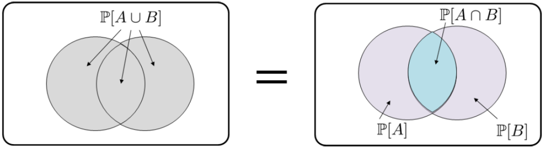


Properties of Probability Laws

Properties of Probability Laws

Consider a probability law, and let A and B be events

- If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$



- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Conditional Probability

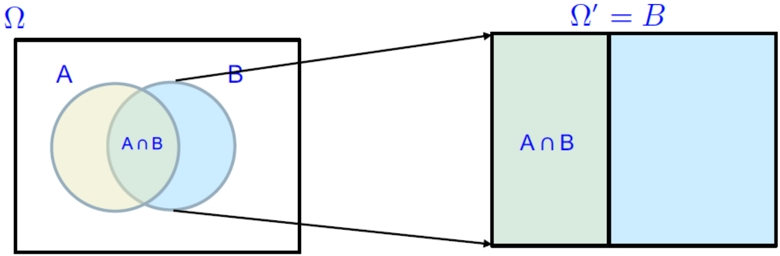
Conditional Probability

Conditioning the original sample space means changing the perspective: instead of finding the area of A inside Ω, we are finding the area of $A \cap B$ inside B

Conditional Probability

Let $\mathbb{P}(B) > 0$. The Conditional Probability of A, given B is defined as

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



Conditional Probability: $\mathbb{P}(B|A)$

Example. Roll two dice. $A =$"The total dots is more than 8" and $B =$"The first die shows 3 dots". What is $\mathbb{P}(B|A)$?

Original

Second Roll

First Roll	1	2	3	4	5	6	
1	2	3	4	5	6	7	
2	3	4	5	6	7	8	
3	4	5	6	7	8	9	B
4	5	6	7	8	9	10	
5	6	7	8	9	10	11	
6	7	8	9	10	11	12	A

Conditioned by knowing A happened:

Second Roll

First Roll	1	2	3	4	5	6
1						
2						
3						9
4					9	10
5				9	10	11
6			9	10	11	12

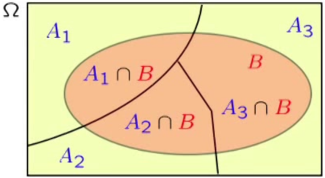
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{1/36}{10/36} = \frac{1}{10}$$

Law of Total Probability

Theorem: Law of Total Probability

Let $A_1, A_2, \ldots, A_n$ be a partition of sample space Ω. Let B be any event. Then

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(A_i \cap B) = \sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B|A_i).$$



Used to compute the probability of events when we have information on the conditional probability of that event!

Product Rule

We know

$$\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) \iff \mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B) \tag{1}$$

We can extend it to general case

Product Rule

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}) \tag{2}$$

Bayes' Formula and Independence

Bayes' Formula

- If $\mathbb{P}(A) \neq 0$, we know

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{3}$$

- We have law of total probability

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

Bayes' Formula

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)}$$

- We have prior information on how likely B (cause) would occur
- We also have information on conditional probability of A (effect) given B (cause)
- Up on observation of A, Bayes' rule gives update on probability of B

Can be extended to the case of more than 2 events!

Independence

In some special cases partial information on an experiment does not change the likelihood of an event

- Sex of first child has nothing to do with sex of second (independent)

Independence ($A \perp B$ or $A \perp\!\!\!\perp B$)

We say two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{4}$$

In this case (assuming $\mathbb{P}(B) \neq 0$)

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

Conditional independence, given C, is defined as independence under probability law $\mathbb{P}(\cdot|C)$

Conditional Independence

We say A and B are conditionally independent given C iff (we write $A \perp\!\!\!\perp B|C$)

Independence and Product Rule

- There are events that are independent but not conditionally independent
- There are events that are conditionally independent but not independent
- Conditional independence simplifies the product rule
  - Recall the product rule

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}) \tag{5}$$

  - If A_i is conditionally independent of $A_j, j < i-1$ given A_{i-1}, then

$$\mathbb{P}(A_i|A_1 \cap \cdots \cap A_{i-1}) = \mathbb{P}(A_i|A_{i-1})$$

    and therefore

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2)\cdots\mathbb{P}(A_n|A_{n-1})$$

Discrete Random Variable

Random Variable

Definition

A Random Variable X is a function $X : \Omega \mapsto \mathbb{R}$ that maps an outcome $\xi \in \Omega$ to a number $X(\xi) \in \mathbb{R}$.

We use $X(\Omega)$ or R_X to denote the range of X, i.e., $X(\Omega) = \{X(\xi) : \xi \in \Omega\}$.

- X is a discrete random variable if $X(\Omega)$ is countable
- X is a continuous random variable if $X(\Omega)$ is uncountable

Example: Toss 3 Coins, $X(\Omega) = \{0, 1, 2, 3\}$

	Sample Space Ω								
	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT	
$P(\xi)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	
$\mathbf{X}(\xi)$	3	2	2	1	2	1	1	0	$\leftarrow$ number of heads
$\mathbf{Y}(\xi)$	1	0	0	0	0	0	0	1	$\leftarrow$ matching tosses
$\mathbf{Z}(\xi)$	8	2	2	$\frac{1}{2}$	2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$	$\leftarrow$ H: double your money T: halve your money

Probability Mass Function (PMF)

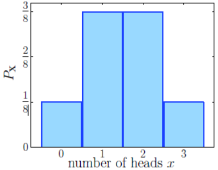
$$\underbrace{\{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}}_{\Omega} \xrightarrow{X} \underbrace{\{0, 1, 2, 3\}}_{X(\Omega)}$$

Each possible value x of the random variable X corresponds to an event

x	0	1	2	3
Event	{TTT}	{HTT, THT, TTH}	{HHT, HTH, THH}	{HHH}

For each $x \in X(\Omega)$, compute $\mathbb{P}(X = x)$ by adding the outcome-probabilities

	possible values $x \in \mathbf{X}(\Omega)$			
x	0	1	2	3
$P_{\mathrm{X}}(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



Probability Mass Function (PMF)

The Probability Mass Function $P_X(a)$ is the probability for the random variable X to take value a

$$P_X(a) = \mathbb{P}(X = a).$$

Cumulative Distribution Function

Cumulative Distribution Function (CDF)

The Cumulative Distribution Function $F_X(x)$ is the probability for the random variable X to be at most x

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

Properties

- $F_X(x)$ is a non-decreasing function of x.
- $F_X(-\infty) = 0$ and $F_X(\infty) = 1$
- $\mathbb{P}(a < X \leq b) = F_X(b) - F_x(a)$.

Special Distribution

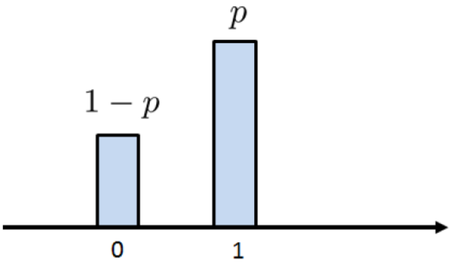
Definition: Bernoulli Distribution

Suppose you have a coin where the probability of a heads is p and we define the random variable

$$X = \text{“the number of heads showing on one tossed coin”}$$

Then we say that X is distributed according to the Bernoulli Distribution with parameter p, and write this as

$$X \sim \text{Bernoulli}(p).$$



Special Distribution

Binomial Random Variable: Sum of Bernoullis

X is the number of successes in n independent trials with success probability p on each trial: $X = X_1 + \cdots + X_n$, where $X_i \sim$ Bernoulli(p)

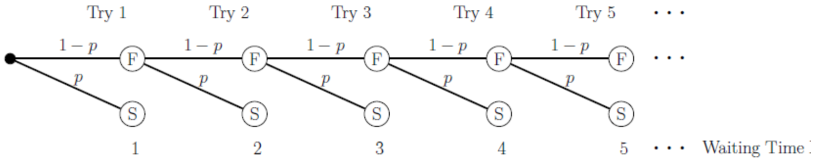
$$P_X(k) = B(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{6}$$

Definition: Geometric Distribution

We say $X \sim$ Geometric(p) with the range $X(\Omega) = \{1, 2, 3, \ldots\}$ iff

$$P_x(k) = (1-p)^{k-1} p \quad \text{for } k = 1, 2, 3, \ldots.$$

Let p be the probability to succeed on a random trial. Let X be the number of trials that appear until the first success.



Continuous Random Variable

Continuous Random Variable

Continuous Random Variables

A random variable having a continuous CDF is said to be a continuous random variable.

Definition: PDF

Let X be a continuous random variable. The probability density function of X is a function $f_X : \mathbb{R} \mapsto \mathbb{R}_+$, when integrated over an interval $[a, b]$, yields the probability of obtaining $a \leq X \leq b$:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

Example . Let

$$f_X(x) = \begin{cases} 3x^2, & \text{if } x \in [0, 1] \\ 0, & \text{otherwise.} \end{cases}$$

Let $A = [0, 0.5]$. Then the probability $\mathbb{P}(\{X \in A\})$ is

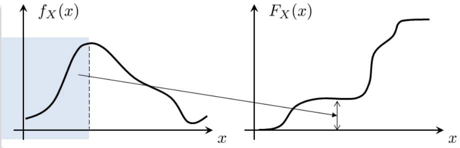
$$\mathbb{P}(0 \leq X \leq 0.5) = \int_0^{0.5} 3x^2 dx = \int_0^{0.5} dx^3 = 1/8.$$

Cumulative Distribution Function

Definition: CDF

Let X be a continuous random variable. The Cumulative Distribution Function of X is

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



Connecting PDF and CDF

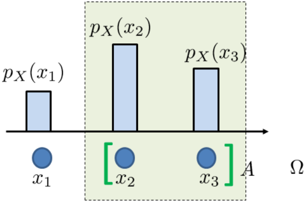
- If X is a continuous random variable and $a \leq b$, then (integration)

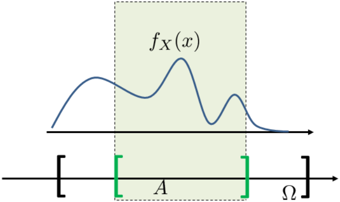
$$\int_a^b f_X(x)dx = \mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) \tag{7}$$

- If F_X is differentiable at x, then (differentiation)

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f_X(y)dy. \tag{8}$$

Property





Intuition

- Probability is a measure of the size of set
- Use length/area/volume to measure the size of a continuous set
- $f_X(x)$ is the weight when calculating the size
  - $f_X(x) \geq 0$
  - $\int_{\Omega(X)} f_X(x)dx = 1$

Definition and properties

- Probability per unit length
- $f_X(x) \geq 1$ is okay

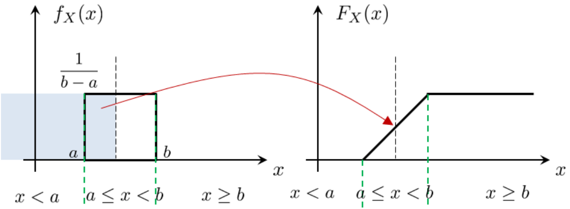
Uniform Random Variable

Definition: Uniform Random Variable

We say X is a continuous uniform random variable on $[a, b]$ if the PDF is

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{Uniform}(a, b)$.



The CDF of a uniform random variable is

$$F_X(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{otherwise.} \end{cases}$$

Gaussian Random Variable

Definition: Gaussian Random Variable

We say X is a Gaussian random variable if the PDF is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{9}$$

where (μ, σ^2) are parameters of the distribution. We write

$$X \sim \text{Gaussian}(\mu, \sigma^2) \quad \text{or} \quad X \sim \mathcal{N}(\mu, \sigma^2).$$

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

- it is symmetric around μ
- σ^2 determines how sharply the variable is around its center

When we sum many independent random variables, the resulting random variable is a Gaussian

$$\sum_{i=1}^{n} X_i \rightarrow \text{a Gaussian random variable if } X_i \text{ are independent}$$

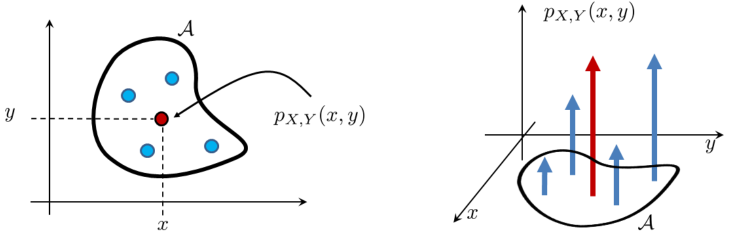
Joint Distributions

Joint PMF

Definition: Joint PMF

Let X and Y be two discrete random variables. The joint PMF of X and Y is defined as

$$P_{X,Y}(x,y) = \mathbb{P}(X = x \text{ and } Y = y). \tag{10}$$



A joint PMF for a pair of discrete random variables consists of an array of impulses. To measure the size of the event $\mathcal{A}$, we sum all the impulses inside $\mathcal{A}$.

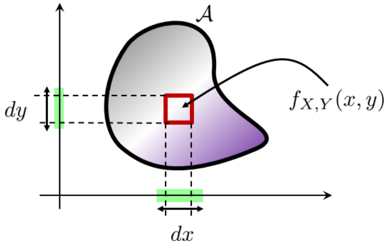
Joint PDF

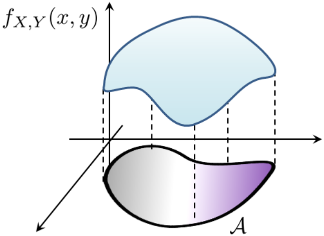
Definition: Joint PDF

Let X and Y be two continuous random variables. The joint PDF of X and Y is a function $f_{X,Y}(x,y)$ that can be integrated to yield a probability:

$$\mathbb{P}(\mathcal{A}) = \int_{\mathcal{A}} f_{X,Y}(x,y)dxdy \tag{11}$$

for any event $\mathcal{A} \subseteq X(\Omega) \times Y(\Omega)$.





Marginal PMF and Marginal PDF

Definition: Marginal PMF and Marginal PDF

The marginal PMF is defined as

$$P_X(x) = \sum_{y \in Y(\Omega)} P_{X,Y}(x,y) \quad \text{and} \quad P_Y(y) = \sum_{x \in X(\Omega)} P_{X,Y}(x,y).$$

The marginal PDF is defined as

$$f_X(x) = \int_{Y(\Omega)} f_{X,Y}(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{X(\Omega)} f_{X,Y}(x,y)dx.$$

Joint CDF

Definition: Joint CDF

Let X and Y be two random variables. The joint CDF of X and Y is the function $F_{X,Y}(x,y)$ such that

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

Theorem

If X and Y are discrete, then

$$F_{X,Y}(x,y) = \sum_{y' \leq y} \sum_{x' \leq x} P_{X,Y}(x',y'). \tag{12}$$

If X and Y are continuous, then

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(x',y')dx'dy'. \tag{13}$$

Conditional Probability of Random Variables

Conditional PMF under an Event

For a discrete random variable X and event A, the conditional PMF of X given A is defined as

$$P_{X|A}(x_i) = \mathbb{P}(X = x_i|A) \qquad = \frac{\mathbb{P}(X = x_i \text{ and } A)}{\mathbb{P}(A)}, \quad \text{for any } x_i \in X(\Omega). \tag{14}$$

Conditional PMF

Let X and Y be two discrete random variables. The conditional PMF of X given Y is

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}.$$

Definition: Conditional PDF

Let X and Y be two continuous random variables. The conditional PDF of X given Y is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Independence

Definition: Independence for Two Variables

We say two random variables X and Y are independent iff

$$P_{X,Y}(x,y) = P_X(x)P_Y(y) \quad \text{or} \quad f_{X,Y}(x,y) = f_X(x)f_Y(y). \tag{15}$$

Definition: Independence for Multiple Variables

We say a sequence of random variables $X_1, X_2, \ldots, X_N$ are independent iff the joint PDF (or joint PMF) can be factorized

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = \prod_{n=1}^{N} f_{X_n}(x_n). \tag{16}$$

Descriptive Statistics

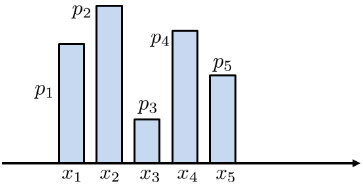
Expectation

Expectation: The Expectation of a random variable X is

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x P_X(x). \tag{17}$$

Intuition: it gives expected value before experiment

$$\mathbb{E}[X] = \underbrace{\sum_{x \in X(\Omega)}}_{\text{sum over all states}} \underbrace{x}_{\text{a state } X \text{ takes}} \underbrace{P_X(x)}_{\text{the percentage}}$$



$\mathbb{E}[X] = p_1 x_1 + \ldots + p_5 x_5$

Linearity of Expectation

Let $X_1, X_2, \ldots, X_k$ be random variables. Let $a_1, \ldots, a_k$ be constants. Then

$$\mathbb{E}[\sum_{i=1}^{k} a_i X_i] = \sum_{i=1}^{k} a_i \mathbb{E}[X_i].$$

Variance: Size of Deviations From the Mean

Let X=sum of 2 dice.

$$\mathbb{E}[X] = \frac{1}{36}\cdot 2 + \frac{2}{36}\cdot 3 + \frac{3}{36}\cdot 4 + \frac{4}{36}\cdot 5 + \cdots + \frac{1}{36}\cdot 12 = 7 \leftarrow \mu$$

Let $\Delta = X - \mu$, which measures the deviation from the mean.

$\mathbf{X}$	2	3	4	5	6	7	8	9	10	11	12	
$\boldsymbol{\Delta}$	-5	-4	-3	-2	-1	0	1	2	3	4	5	$\leftarrow \mathbf{X} - \mu$
$P_{\mathbf{X}}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

Variance and Standard Deviation

Variance, $\text{Var}(X)$, is the expected value of the squared deviations

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

The Standard Deviation, σ, is the square-root of the variance: $\sigma = \sqrt{\mathbb{E}[\Delta^2]}$.

$$\begin{aligned}\text{Var}(X) &= \frac{1}{36}\cdot(-5)^2 + \frac{2}{36}(-4)^2 + \frac{3}{36}(-3)^2 + \frac{4}{36}(-2)^2 + \frac{5}{36}(-1)^2 + \frac{6}{36}0^2 \\ &\quad + \frac{1}{36}\cdot(5)^2 + \frac{2}{36}(4)^2 + \frac{3}{36}(3)^2 + \frac{4}{36}(2)^2 + \frac{5}{36}(1)^2 = \frac{35}{6}.\end{aligned}$$

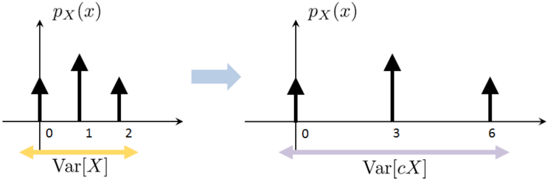
Properties of Variance: let X be a random variable

- Variance is the expectation of square minus the square of expectation

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \tag{18}$$

- Scale. For any constant c

$$\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$$



- Shift. For any constant c

$$\mathrm{Var}(X + c) = \mathrm{Var}(X)$$

- If X and Y are independent then

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

A problem

Throw a dice twice. Let X be the first number, and Y be the second number. Define

$$Z = \max\{X, Y\}.$$

1. Find the PMF of Z
2. Find $\mathbb{P}(Z \leq 5 | X \geq 4)$
3. Find the expectation of Z, i.e., $\mathbb{E}[Z]$

Solution

We have the table (different rows correspond to the outcome of the first die, different columns correspond to the outcome of the second die)

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

- We have

$$P_Z(1) = \frac{1}{36},\ P_Z(2) = \frac{3}{36},\ P_Z(3) = \frac{5}{36}$$
$$P_Z(4) = \frac{7}{36},\ P_Z(5) = \frac{9}{36},\ P_Z(6) = \frac{11}{36}$$

Solution

According to the table, we know

$$\mathbb{P}(X \geq 4) = \frac{18}{36}$$

and (the 10 outcomes are
$(4,1),(4,2),(4,3),(4,4),(4,5),(5,1),(5,2),(5,3),(5,4),(5,5)$)

$$\mathbb{P}(Z \leq 5, X \geq 4) = \frac{10}{36}$$

It then follows that

$$\mathbb{P}(Z \leq 5 | X \geq 4) = \frac{\mathbb{P}(Z \leq 5, X \geq 4)}{\mathbb{P}(X \geq 4)} = \frac{10}{18} = \frac{5}{9}.$$

The expectation of Z is

$$\begin{aligned}
\mathbb{E}[Z] &= 1 * P_Z(1) + 2 * P_Z(2) + 3 * P_Z(3) + 4 * P_Z(4) + 5 * P_Z(5) + 6 * P_Z(6) \\
&= 1 * \frac{1}{36} + 2 * \frac{3}{36} + 3 * \frac{5}{36} + 4 * \frac{7}{36} + 5 * \frac{9}{36} + 6 * \frac{11}{36} \\
&= \frac{1 + 2 * 3 + 3 * 5 + 4 * 7 + 5 * 9 + 6 * 11}{36} = \frac{161}{36}
\end{aligned}$$