데이터 통계 기초

빅데이터분석
컴퓨터AI공학부
천세진

Outline

- Probability
- Discrete random variables
  - Random Variables
  - Cumulative Distribution Function(CDF)
  - Expectation
- Continuous random variables
  - PDF
  - Gaussian random variables

Probability

Probability is a measure of the size of a set.

Probability

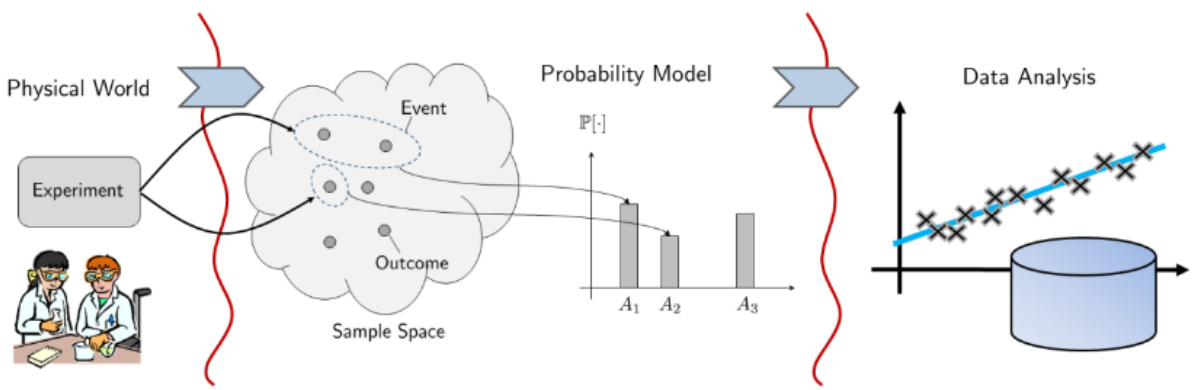


Figure 2.12: Given an experiment, we define the collection of all outcomes as the sample space. A subset in the sample space is called an event. The probability law is a mapping that maps an event to a number that denotes the size of the event.

Sample Space, Event Space, and Probability Law

- Set 이론 위에서 진행

- **Sample Space** Ω: The set of all possible outcomes from an experiment.
- **Event Space** $\mathcal{F}$: The collection of all possible events. An event E is a subset in Ω that defines an outcome or a combination of outcomes.
- **Probability Law** $\mathbb{P}$: A mapping from an event E to a number $\mathbb{P}[E]$ which, ideally, measures the size of the event.

Thinking) 주사위, 고객방문, 구매확률

Sample space

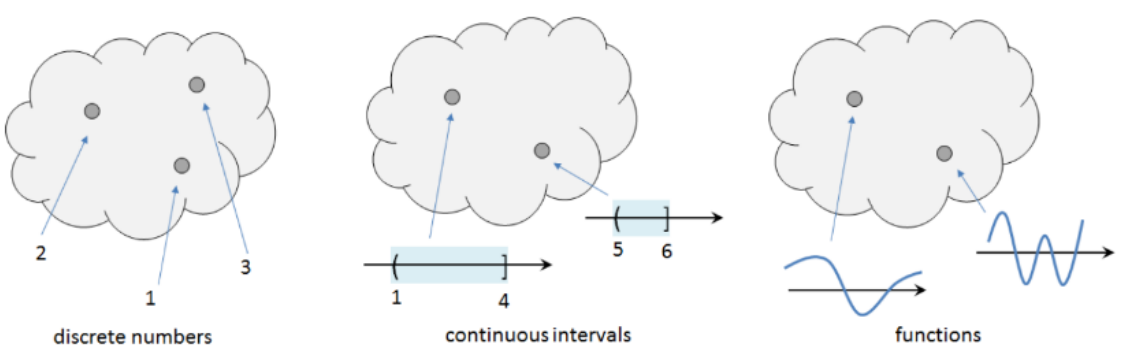


Figure 2.13: The sample space can take various forms: it can contain discrete numbers, or continuous intervals, or even functions.

DISCRETE RANDOM VARIABLES

Mapping Probability to Data, and vice versa

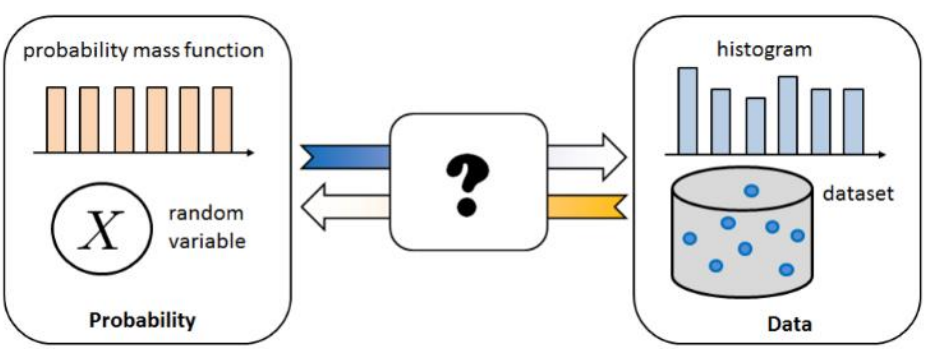


Figure 3.1: The landscape of probability and data. Often we view probability and data analysis as two different entities. However, probability and data analysis are inseparable. The goal of this chapter is to link the two.

Key Concepts

Key Concept 1: What are random variables?

Random variables are mappings from events to numbers.

Key Concept 2: What are probability mass functions (PMFs)?

Probability mass functions are the ideal histograms of random variables.

Key Concept 3: What is expectation?

Expectation = Mean = Average computed from a PMF.

Random Variables

- Symbol ♣, ♢, ♡, ♠
- Encode each symbol with a number

  $\clubsuit \leftarrow 1, \diamondsuit \leftarrow 2, \heartsuit \leftarrow 3, \spadesuit \leftarrow 4,$
- 각 outcome에 을 얻을 확률

$$\mathbb{P}[\{\clubsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{\diamondsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\spadesuit\}] = \frac{1}{6}.$$

- Function X와 encode된 symbol을 함께

$$X(\clubsuit) = 1, \quad X(\diamondsuit) = 2, \quad X(\heartsuit) = 3, \quad X(\spadesuit) = 4.$$

Random Variables

$$\mathbb{P}[X=1]=\frac{1}{6},\quad \mathbb{P}[X=2]=\frac{2}{6},\quad \mathbb{P}[X=3]=\frac{2}{6},\quad \mathbb{P}[X=4]=\frac{1}{6}.$$

PMF

Definition 3.2. *The* **probability mass function (PMF)** *of a random variable X is a function which specifies the probability of obtaining a number $X(\xi)=x$. We denote a PMF as*

$$p_X(x)=\mathbb{P}[X=x]. \tag{3.1}$$

The set of all possible states of X is denoted as $X(\Omega)$.

PMF

Example 3.5. Flip a coin twice. The sample space is $\Omega = \{\text{HH, HT, TH, TT}\}$. We can assign a random variable X = number of heads. Therefore,

$$X(\text{“HH”}) = 2, X(\text{“TH”}) = 1, X(\text{“HT”}) = 1, X(\text{“TT”}) = 0.$$

So the random variable X takes three states: 0, 1, 2. The PMF is therefore

$$p_X(0) = \mathbb{P}[X = 0] = \mathbb{P}[\{\text{“TT”}\}] = \frac{1}{4},$$
$$p_X(1) = \mathbb{P}[X = 1] = \mathbb{P}[\{\text{“TH”}, \text{“HT”}\}] = \frac{1}{2},$$
$$p_X(2) = \mathbb{P}[X = 2] = \mathbb{P}[\{\text{“HH”}\}] = \frac{1}{4}.$$

Generative perspective

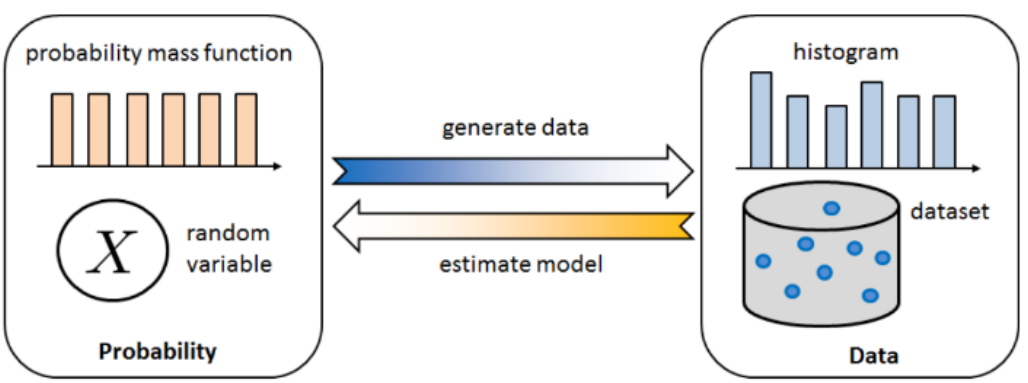
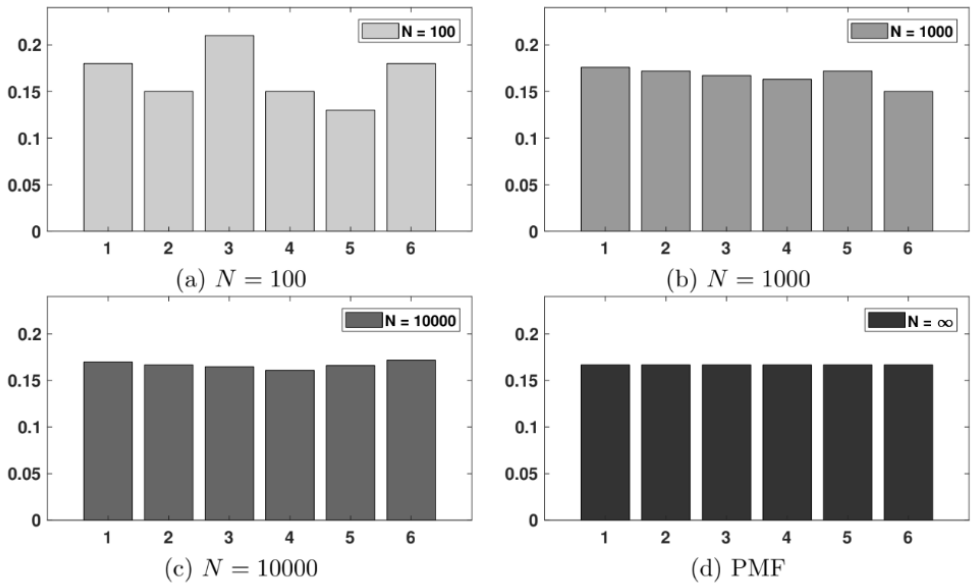


Figure 3.9: When analyzing a dataset, one can treat the data points are samples drawn according to a latent random variable with certain a PMF. The dataset we observe is often finite, and so the histogram we obtain is empirical. A major task in data analysis is statistical inference, which tries to retrieve the model information from the available measurements.

Gene



(a) $N = 100$ (b) $N = 1000$

(c) $N = 10000$ (d) PMF

Figure 3.8: Histogram and PMF, when throwing a fair die N times. As N increases, the histograms are becoming more similar to the PMF.

CDF: Cumulative distribution function

Definition 3.3. *Let X be a discrete random variable with $\Omega = \{x_1, x_2, \ldots\}$. The* **cumulative distribution function** *(CDF) of X is*

$$F_X(x_k) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x_k] = \sum_{\ell=1}^{k} p_X(x_\ell). \tag{3.6}$$

If $\Omega = \{\ldots, -1, 0, 1, 2, \ldots\}$, then the CDF of X is

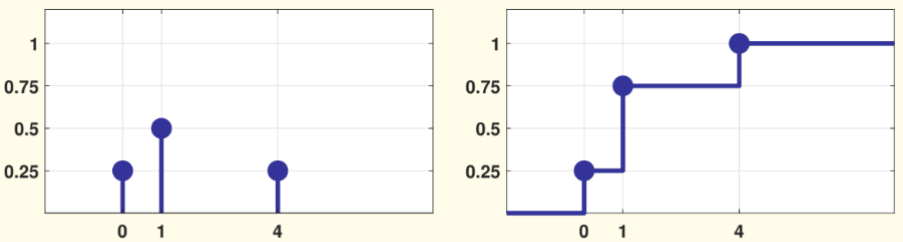
$$F_X(k) \stackrel{\text{def}}{=} \mathbb{P}[X \leq k] = \sum_{\ell=-\infty}^{k} p_X(\ell). \tag{3.7}$$

CDF:

Example 3.6. Consider a random variable X with PMF $p_X(0) = \frac{1}{4}$, $p_X(1) = \frac{1}{2}$ and $p_X(4) = \frac{1}{4}$. The CDF of X can be computed as

$$F_X(0) = \mathbb{P}[X \le 0] = p_X(0) = \frac{1}{4},$$
$$F_X(1) = \mathbb{P}[X \le 1] = p_X(0) + p_X(1) = \frac{3}{4},$$
$$F_X(4) = \mathbb{P}[X \le 4] = p_X(0) + p_X(1) + p_X(4) = 1.$$

As shown in **Figure 3.13**, the CDF of a discrete random variable is a staircase function.



(a) PMF $p_X(k)$ (b) CDF $F_X(k)$

Figure 3.13: Illustration of a PMF and a CDF.

Expectation

Definition 3.4. *The* **expectation** *of a random variable X is*

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\, p_X(x). \tag{3.10}$$

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

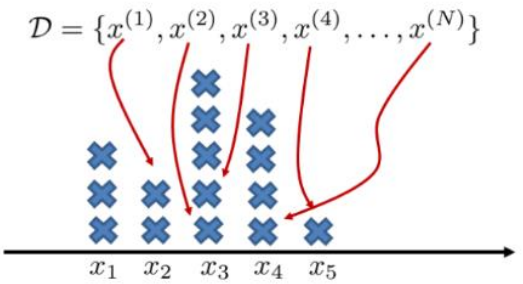


Figure 3.16: If we have a dataset $\mathcal{D}$ containing N samples, and if there are only K distinct values, we can effectively put these N samples into K bins. Thus, the "average" (which is the sum divided by the number N) is exactly the same as our definition of expectation.

$$\text{average} = \underbrace{\sum_{k=1}^{K}}_{\text{sum of all states}} \underbrace{\text{value } x_k}_{\text{a state } X \text{ takes}} \times \underbrace{\frac{\text{number of samples with value } x_k}{N}}_{\text{the percentage}},$$

COMMON DISCRETE RANDOM VARIABLES

Power of Probability

- Ability to summarize microstates using macro descriptions
- Moment-generating functions
  - N 개의 Random variables을 합하는 편리한 방법

Distribution	PMF / PDF	$\mathbb{E}[X]$	$\mathrm{Var}[X]$	$M_X(s)$
Bernoulli	$p_X(1) = p$ and $p_X(0) = 1 - p$	p	$p(1-p)$	$1 - p + pe^s$
Binomial	$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$	$(1 - p + pe^s)^n$
Geometric	$p_X(k) = p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^s}{1-(1-p)e^s}$
Poisson	$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$e^{\lambda(e^s-1)}$
Gaussian	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$\exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}$
Exponential	$f_X(x) = \lambda \exp\{-\lambda x\}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - s}$
Uniform	$f_X(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb} - e^{sa}}{s(b-a)}$

Table 6.1: Moment-generating functions of common random variables.

BERNOULLI

- 결과가 두 가지 중 하나로만 나오는 실험이나 시행(trials)

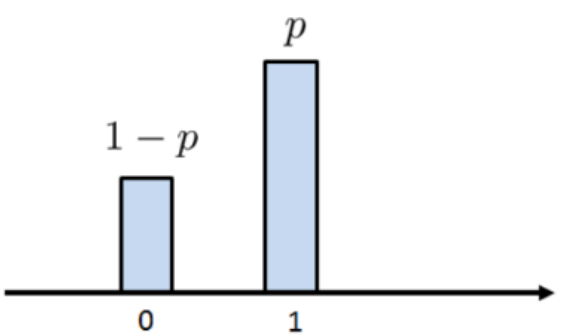


Figure 3.22: A Bernoulli random variable has two states with probability p and $1 - p$.

BINOMIAL(이항 분포) n>2

- 연속된 *n*번의 독립적 시행에서 각 시행이 확률 p를 가질 때

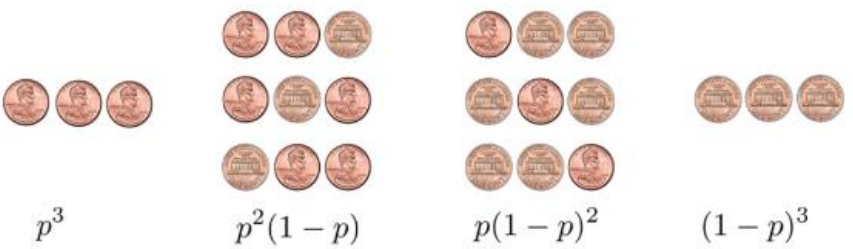
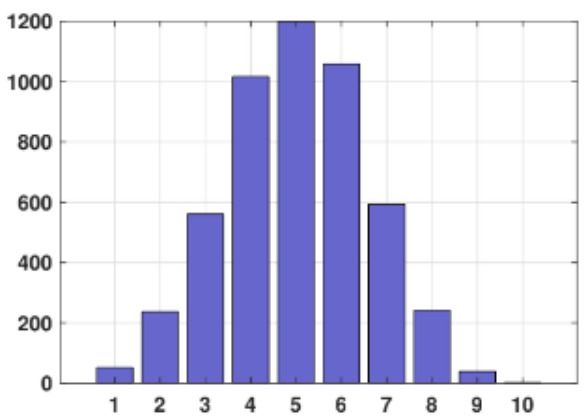
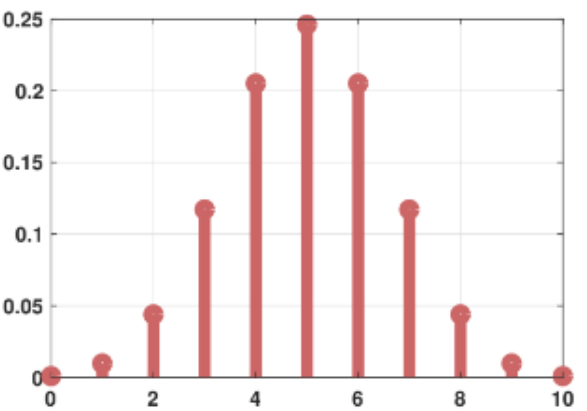


Figure 3.27: The probability of getting k heads out of $n = 3$ coins.

BINOMIAL

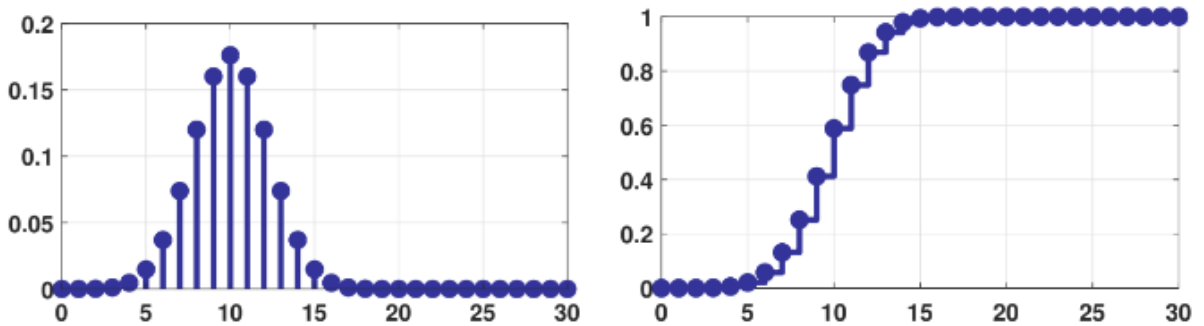


(a) Histogram based on 5000 samples



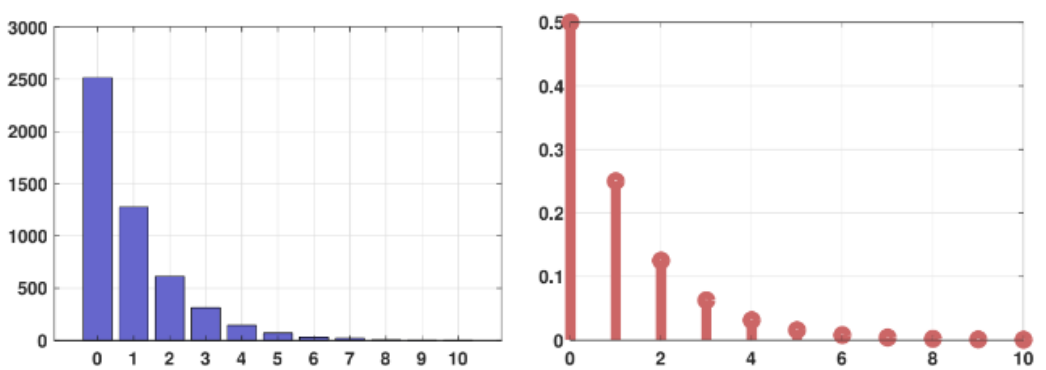
(b) PMF

BINOMIAL



Geometric (기하 분포)

- 동일한 베르누이 분포를 따르는 시행의 독립적인 반복에서 처음으로 성공하기까지의 시도횟수



(a) Histogram based on 5000 samples (b) PMF

Figure 3.31: An example of a geometric distribution with $p = 0.5$.

Poisson*

- 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현
- Arrivals of events
  - Telephone call, website updates, electron emission, the number of conversations per user, the number of transactions per paying

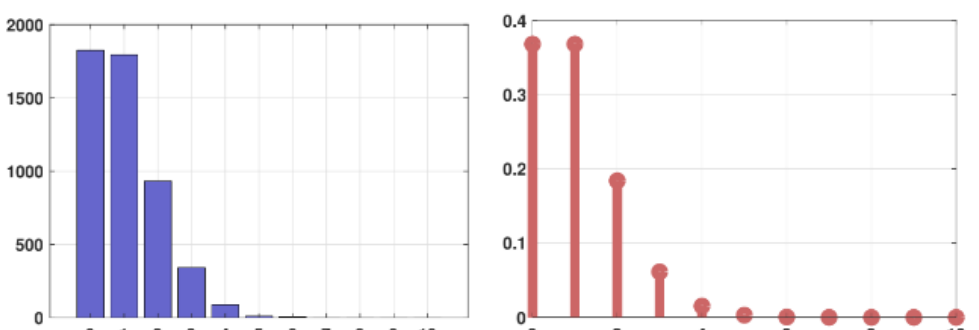
Definition 3.11. *Let X be a* **Poisson random variable**. *Then, the PMF of X is*

$$p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$ is the Poisson rate. We write

$$X \sim \text{Poisson}(\lambda)$$

to say that X is drawn from a Poisson distribution with a parameter λ.



(a) Histogram based on 5000 samples (b) PMF

Figure 3.33: An example of a Poisson distribution with $\lambda = 1$.

- Lamba와 함께 변화의 모양을 구성할 수 있음

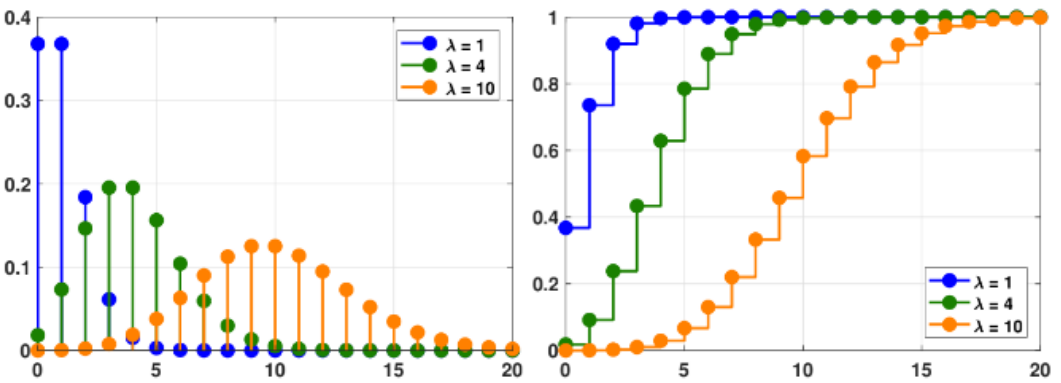
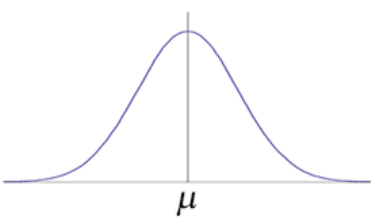


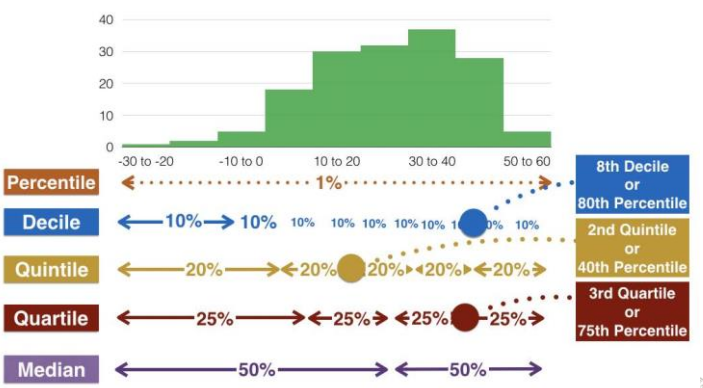
Figure 3.34: A Poisson random variable using different λ's. [Left] Probability mass function $p_X(k)$. [Right] Cumulative distribution function $F_X(k)$.

Gaussian(=Normal) 분포

- 평균과 표준편차를 통해 모양을 결정하는 분포
  - 사용자 패턴 등 많은 부분에서 사용됨



$$X \sim N(\mu, \sigma^2)$$



TRAFFIC MODEL WITH POISSON

Traffic model

- Lamba와 함께 변화의 모양을 구성할 수 있음

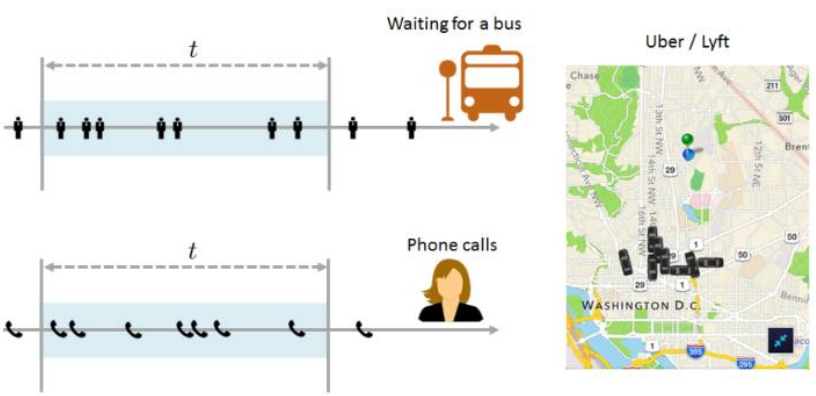


Figure 3.36: The Poisson random variable can be used to model passenger arrivals and the number of phone calls, and can be used by Uber or Lyft to provide shared rides.

Joint Distributions

Joint Distribution

Joint distributions are **high-dimensional** PDFs (or PMFs or CDFs).

$$\underbrace{f_X(x)}_{\text{one variable}} \implies \underbrace{f_{X_1,X_2}(x_1,x_2)}_{\text{two variables}} \implies \cdots \implies \underbrace{f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)}_{N \text{ variables}}.$$

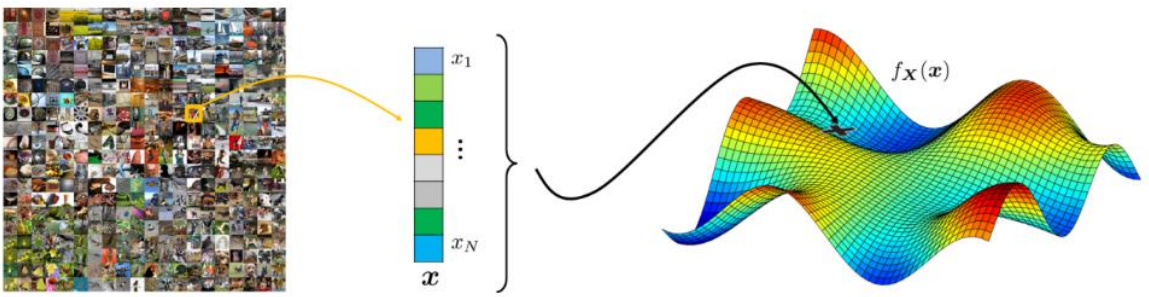


Figure 5.1: Joint distributions are ubiquitous in modern data analysis. For example, an image from a dataset can be represented by a high-dimensional vector $\boldsymbol{x}$. Each vector has a certain probability of being present. This probability is described by the high-dimensional joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$. The goal of this chapter is to understand the properties of this $f_{\boldsymbol{X}}$.

2-dimensional PDF

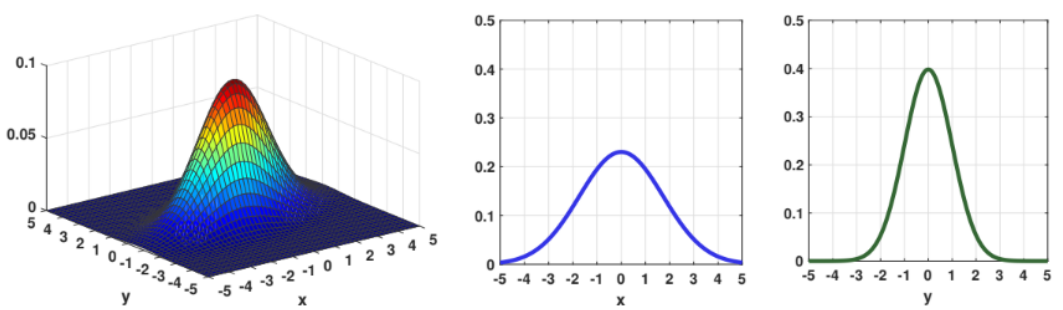


Figure 5.2: A 2-dimensional PDF $f_{X,Y}(x,y)$ of a pair of random variables (X,Y) and their respective 1D PDFs $f_X(x)$ and $f_Y(y)$.

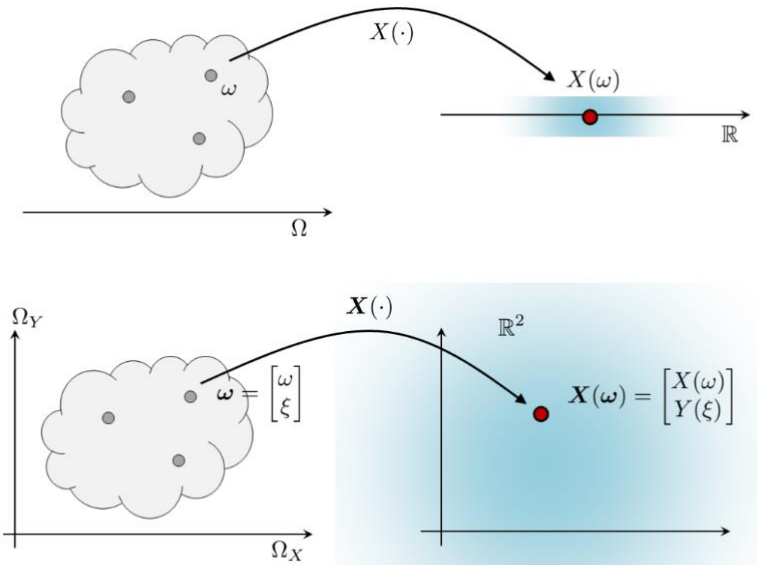


Figure 5.3: When there is a pair of random variables, we can regard the sample space as a set of coordinates. The random variables are 2D mappings from a coordinate $\boldsymbol{\omega}$ in $\Omega_X \times \Omega_Y$ to another coordinate $\boldsymbol{X}(\boldsymbol{\omega})$ in $\mathbb{R}^2$.

Credits

- https://probabilitycdatascience.com/