Week 6: Statistics (Basic Math/Concepts for Stats.)

Instructor: Daejin Choi (djchoi@inu.ac.kr)

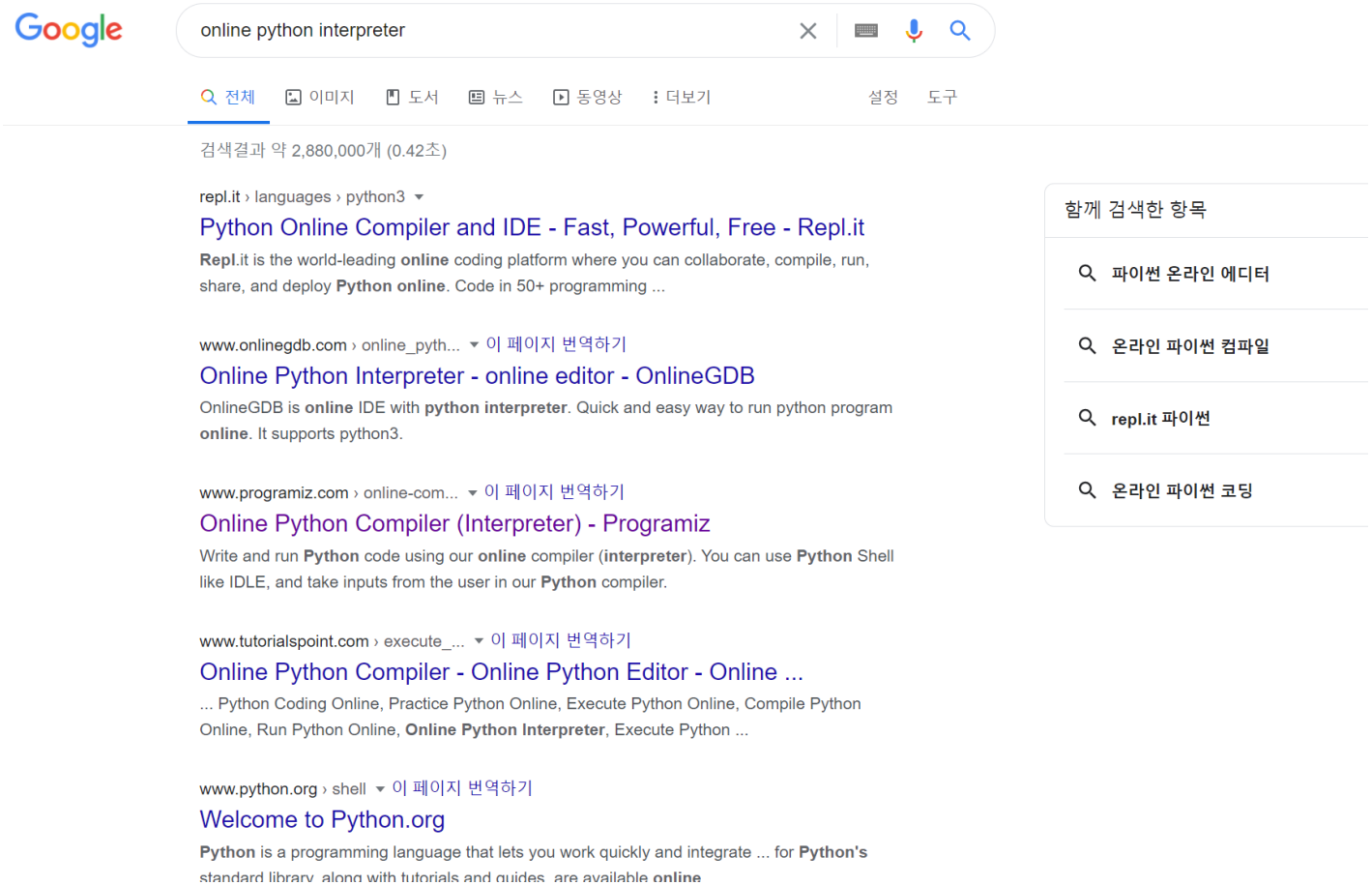
INCHEON NATIONAL UNIVERSITY

* Most slides are used from www.munmund.net & www.deeplearningbook.org/lecture_slides.html with small updates

Before Starting

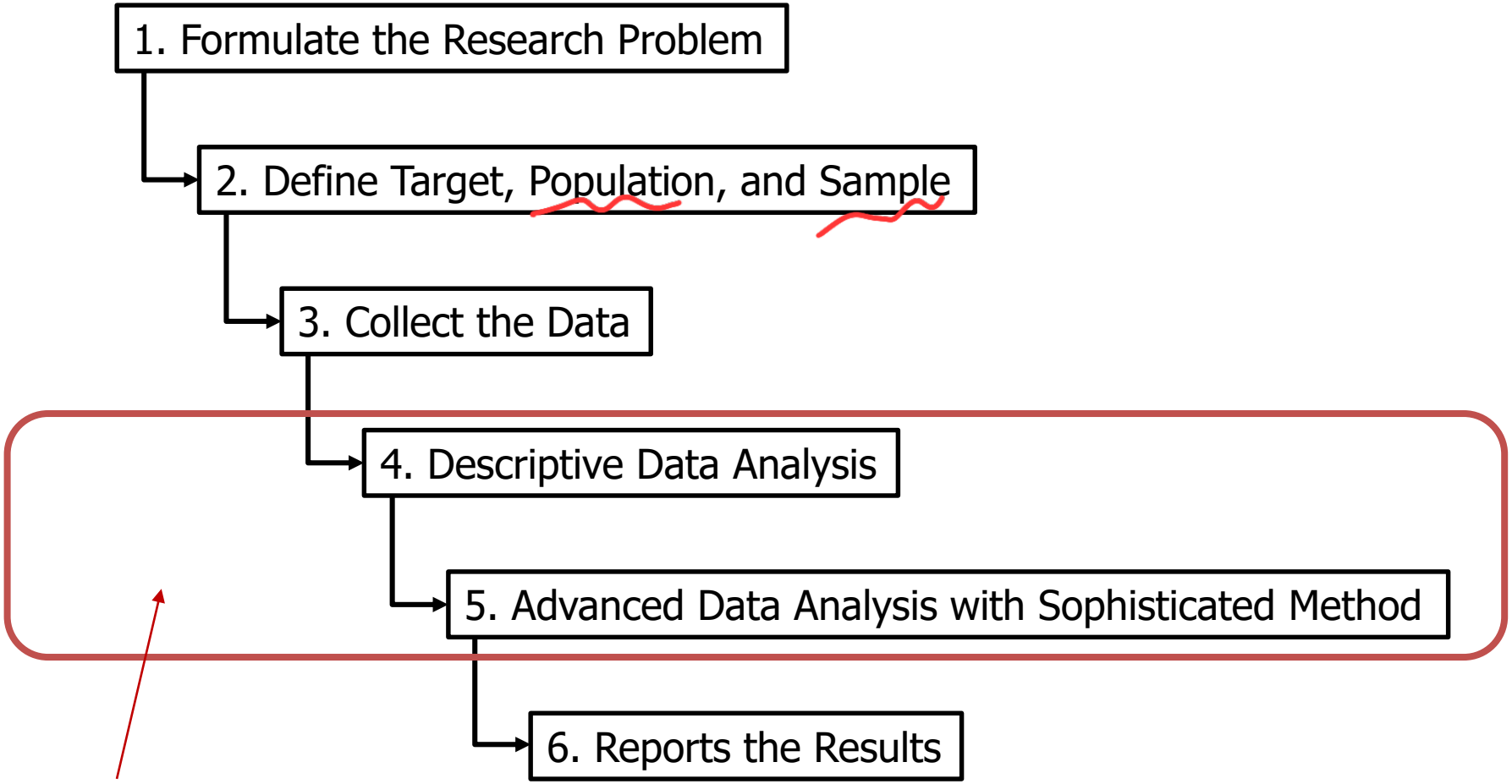
- In the lecture today, python-based codes are used for better explanation

- You can try if you have already installed, or you can use online python interpreter



Why Statistics?

- In general, analysis process is …



You can easily start with statistical methods!
But, a little knowledge is required. → Today we will cover!

Contents for Today

- Goal: Introducing the brief concepts / maths for Statistics.

- Contents
  - Terminologies
  - Central Tendency
  - Probability
  - Distribution & Information Theory

Terminologies

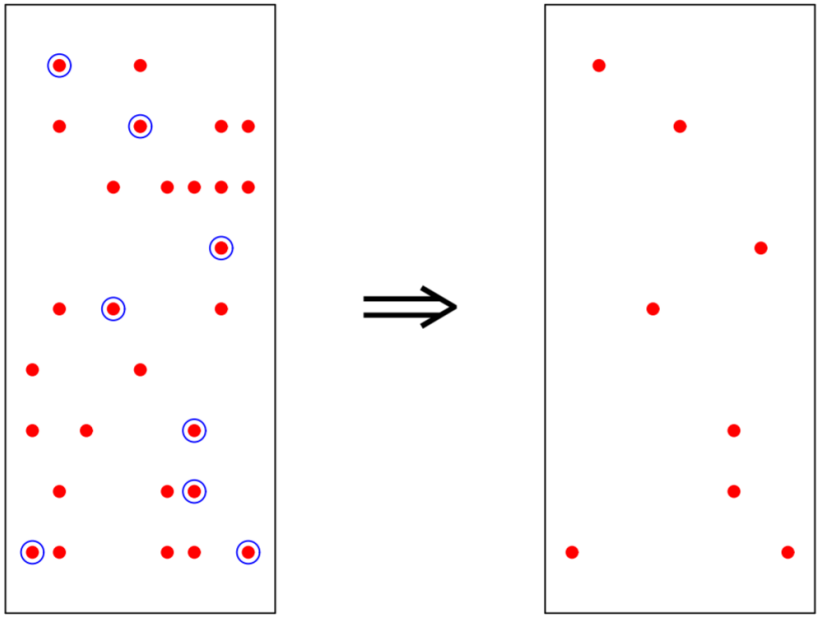
Terminology: Population vs. Sample

- Population: A (statistical) population is the set of measurements (or record of some qualitative trait) corresponding to the **entire collection of units** for which inferences are to be made

Target 집단 수집한 데이터셋

- Sample: A sample from statistical population is the set of measurements that are actually collected in the course of an investigation

Population vs. Sample



Terminology: Descriptive vs. Inferential

- Descriptive Statistics: consists of methods for **organizing and summarizing information** (Weiss, 1999) ⇒ 요약한 정보
- Inferential Statistics: consists of methods for **drawing and measuring the reliability of conclusions** about 신뢰성 population based on information obtained from a **sample** of the population (Weiss, 1999) 유의미한결과 나타내기 위한 통계학적 test
- NOTE: Descriptive and inferential statistics are interrelated
- Assume that you have just started data analysis. What you will do?
  - Preliminary descriptive analysis
  - (Choice of appropriate inferential method)
  - Inferential analysis

Terminology: Parameters vs. Statistics

- Parameters: an **unknown** numerical summary of the population (Agresti & Finalay, 1997) α, β

- Statistics: a **known** numerical summary of the sample which can be used to make inference about parameters (Agresti & Finalay, 1997) 알려진것, 평균 이런거
  ↳ 방법+방법으로 나타난 결과도

- Example
  - The proportion p of 18-30 year-olds who use Twitter at least once a day → Parameter (parameter)
  - The proportion p' of 18-30 year-olds using Twitter at least once a day, calculated from a sample of 18-30 year-olds

sample에서 계산
적어도 한 번 사용 → 확률(p')

Terminology: Discrete vs. Continuous

- Discrete Variables: indicates a finite numbers of values or as many values as there are integers. E.g.) tweet counts per day

- Continuous Variables: indicates any real values e.g., time difference between two consecutive tweet postings

중심적 경향

Central Tendency

Measure of Central Tendency: Mean

- The sum of observed values in a data divided by the number of observations

- p.s. Often affected by extreme values in the data (i.e., outliers)

```
>>> import numpy
>>> numpy.mean(numpy.array([1,2,3,1,3,3,54,6]))
9.125
```

Measure of Central Tendency: Median

- The value of the variable in a data set that divides the set of the observed values in half, so that the observed values in one half are less than or equal to the median value and the observed values in the other half are greater or equal to the median value

- To compute median, the observed values of variable in a data should be sorted
  - If the # observation is odd, then the sample median is the observed value exactly in the middle of the ordered list
  - If the # observation is even, then the sample median is the number halfway between two middle observed values in the ordered list

```
>>> import numpy
>>> numpy.median(numpy.array([1,2,3,1,3,3,54,6]))
3
```

Measure of Central Tendency: Mode

- A discrete variable (or its value) is that value of the variable which occurs with the greatest frequency in a data set

```
>>> from statistics import mode
>>> mode([1, 1, 2, 3, 3, 3, 3, 4])
3
```

- p.s. if the greatest frequency is 1 (i.e., no value occurs more than once), then the variable has no mode

Measure of Variation: Range

- The difference between its maximum and minimum values in a data set
- Formally, range = max. – min.
- p.s. range is affected by only max. and min. values, which means that other values may be ignored.
  - E.g., X = [1, 10, 10, 11, 11, 10, 11, 10, 11, 10, 11, 11, 10, 11, 100]
  - Can we say the range of X is [1, 100]?

Measure of Variation: Percentile & Quartiles

- A method dividing a given dataset with the equal portions
  - Quartile (4 parts)

- Let n denote the number of observations in a data set. Arrange the observed values of variable in a data in increasing order
  - The 1st quartile Q1 is at position (n+1)/4
  - The 2nd quartile Q2 (the median) is at position (n+1)/2
  - The 3rd quartile Q3 is at position 3(n+1)/4

  In the ordered list

- Interquartile range: the difference between 1st and 3rd quartiles of the variable (i.e., IQR = Q3-Q1)
  - Roughly, the IQR gives the range of the middle 50% of the observed values

Measure of Variation: Interquartile Range = IQR

- The sample of interquartile range represents the length of the interval covered by the center half of the observed values of the variable

- This measure of variation is not disturbed if a small fraction the observed values are very large or very small outlier 제거 효과

- Five-number summary: consists of min., max., and quartiles written in increasing order
  - Min, Q1, Q2, Q3, Max

Measure of Variation: Standard Deviation

- A measure of a distribution's deviation from its mean. (square root of variance)

- For a variable x, the sample standard deviation is

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2\right]}, \quad \text{where} \quad \mu = \frac{1}{N}(x_1+\cdots+x_N),$$

- Preferred when the mean is used as the measure of center (e.g., symmetric distribution)

Measure of Variation: Standard Deviation

- Trivially, more variation → larger standard deviation

- However, it can be strongly affected by a few extreme observations

```
>>> import numpy
>>> numpy.std([0.1,2.8,3.7,2.6,5,3.4])
1.4851
```

Probability

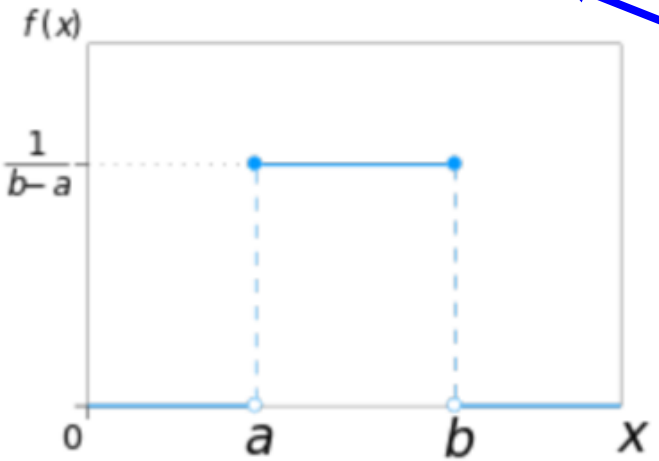
Probability Mass Function (PMF)

- For discrete random variables
- The domain of P must be the set of all possible states of x
- $\forall x \in$ x, $0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring
- $\sum_{x \in \mathrm{X}} P(x) = 1$. We refer to this property as being normalized. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring
- Uniform distribution among k states: $P(\mathrm{x} = x_i) = \frac{1}{k}$

이산적 확률

Probability Density Function (PDF)

- For continuous random variables
- The domain of P must be the set of all possible states of x
- $\forall x \in$ x, $p\ (x) \geq 0$. Note that we do not require $p(x) \leq 1$
- $\int p(x)dx = 1$
- Uniform distribution $u(x; a, b) = \frac{1}{b-a}$.



Parameterized by

Rules associated with Probability

- Computing marginal probability with the Sum Rule
  - $\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y).$
  - $p(x) = \int p(x, y) dy.$
- Conditional Probability

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}.$$

- Bayes' Rule: $P(x \mid y) = \frac{P(x)\, P(y \mid x)}{P(y)}$
- Chain Rule of Probability

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$$

E.g., $P(a,b,c) = P(a|b,c)\, P(b|c)\, P(c)$

Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y).$$

Notation: $x \perp y$ $P(A)P(B) = 0$

- Conditional Independence
  - $\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}$
  - $p(\mathrm{x} = x, \mathrm{y} = y|\mathrm{z} = z) = p(\mathrm{x} = x|\mathrm{z} = z)p(\mathrm{y} = y|\mathrm{z} = z)$ = conditional Independence
  - Equivalently, $p(x|y, z) = p(x|z)$
    - Proof?
  - Notation: $x \perp y \mid z$ → 독립

Expectation

- Discrete variable: $\mathbb{E}_{\mathbf{x}\sim P}[f(x)] = \sum_x P(x)f(x),$
- Continuous variable: $\mathbb{E}_{\mathbf{x}\sim p}[f(x)] = \int p(x)f(x)dx.$
- Linearity of Expectations

$$\mathbb{E}_{\mathbf{x}}[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}_{\mathbf{x}}[f(x)] + \beta\mathbb{E}_{\mathbf{x}}[g(x)],$$

  - NOTE: this always holds, even when f(x) and g(x) are dependent

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$$

Covariance

Cov (X, Y) = E[(X − E(X))(Y − E(Y))]

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right].$$

Note: Covariance can be negative 음수가 될 수 있음

- Covariance Matrix: $\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j)$.
  - Diagonal elements are variance of i-th element

Randomized Algorithms (=Probabilistic Algo.)

	전두경사 **Las Vegas** ⟷	**Monte Carlo**
Type of Answer	Exact	Random amount of error
Runtime	Random (until answer found)	Chosen by user (longer r untime gives less error)

Estimating sums / integrals with samples

$$s = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) f(\boldsymbol{x}) = E_p[f(\mathbf{x})]$$

$$s = \int p(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = E_p[f(\mathbf{x})]$$

$$\hat{s}_n = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}^{(i)}).$$

Justification

- Unbiased:
  - The expected value for finite *n* is equal to the correct value
  - The value for any specific n samples will have random error, but the errors for different sample sets cancel out

- Low variance:
  - Variance is O(1/n)
  - For very large n, the error converges "almost surely" to 0

Distribution & Information Theory

Bernoulli Distribution

베르누이

- PDF

$$P(\mathrm{x}=1)=\phi$$

$$P(\mathrm{x}=0)=1-\phi$$

$$P(\mathrm{x}=x)=\phi^{x}(1-\phi)^{1-x}$$

$p^x(1-p)^{1-x}$

- Expectation: $\mathbb{E}_{\mathrm{x}}[\mathrm{x}]=\phi$

$E(x)=p$

- Variance: $\mathrm{Var}_{\mathrm{x}}(\mathrm{x})=\phi(1-\phi)$

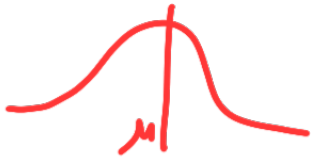
$V(x)=pq$

Gaussian Distribution

- Parameterized by variance:
  - E[x] = μ, Var[x] = σ^2

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

정규분포

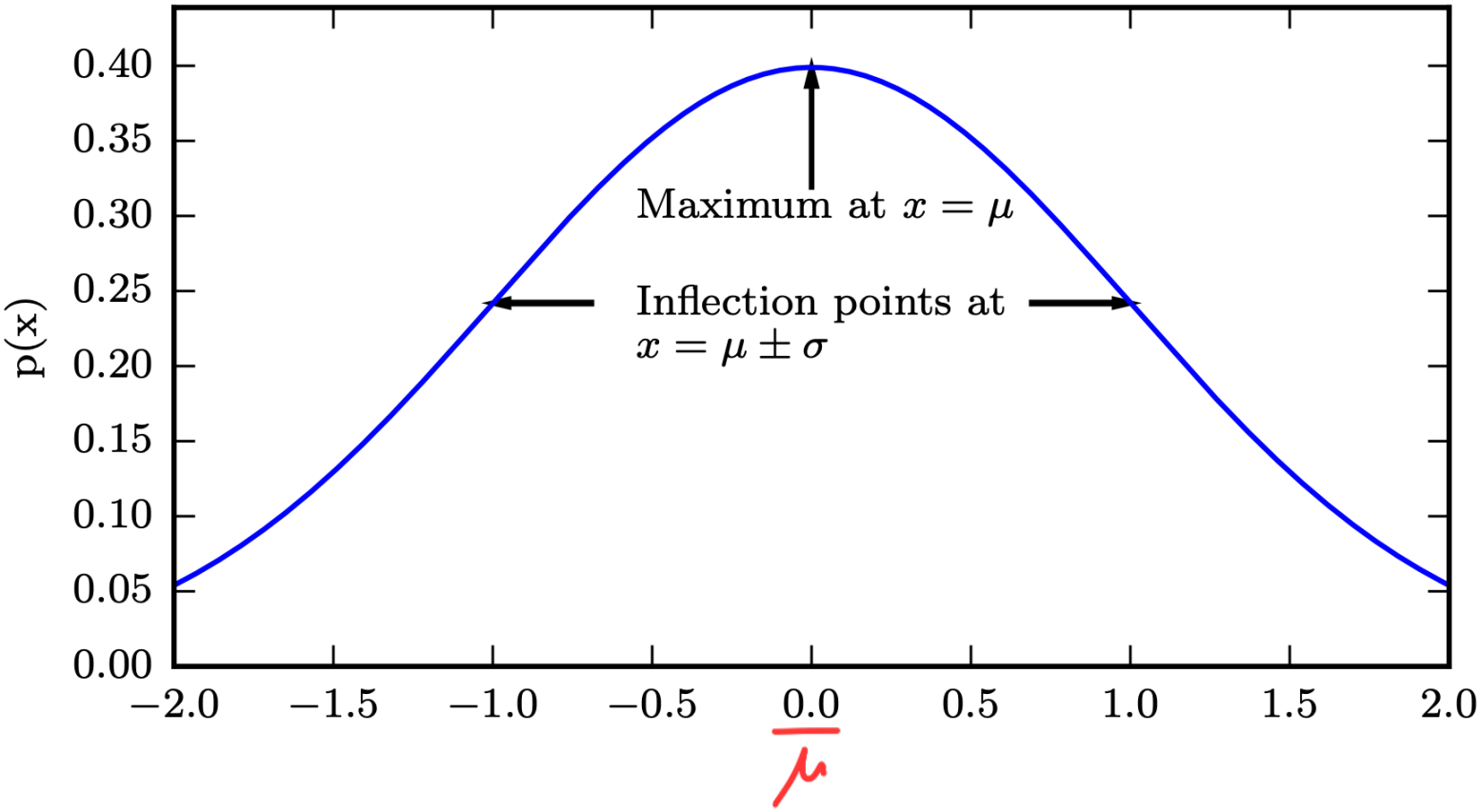


- Parameterized by precision:

$$\mathcal{N}(x;\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{1}{2}\beta(x-\mu)^2\right).$$

$\sigma^2 = \frac{1}{\beta}$

Gaussian Distribution



Theorems related to Gaussian Distribution

- Central limit theorem: the sum of many independent random variables is approximately normally distributed

Normalize

$$\frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \rightarrow N(0,1) \quad \text{as} \quad n \rightarrow \infty$$

대수 법칙

- Law of large numbers: the sample average converges to the expectation as the sample size goes to infinity

$$\overline{X_n} \rightarrow \mu \quad \text{as} \quad n \rightarrow \infty, \text{ where } \overline{X_n} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

$E(X) = \mu$

기대 값이 평균에 수렴한다

Multivariate Gaussian

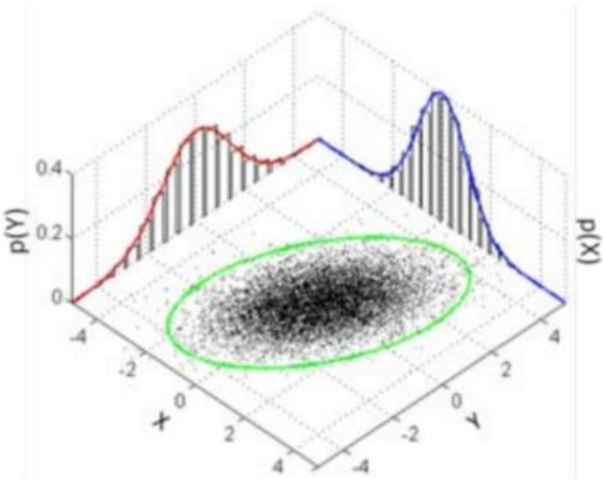
- Parameterized by covariance matrix

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

vector

od-

- μ is a vector
- Σ is a covariance matrix



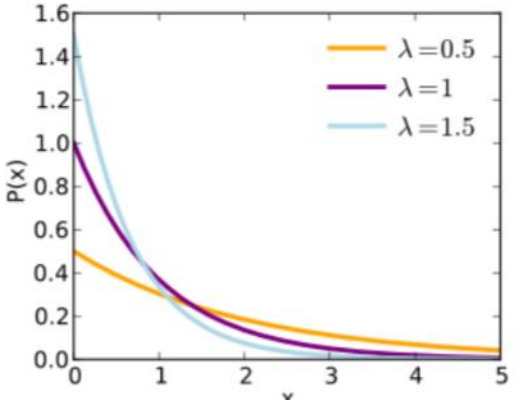
- Parameterized by precision matrix

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

More Distributions

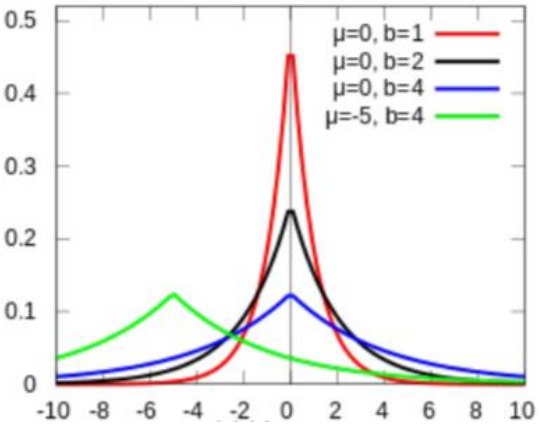
- Exponential:

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x).$$



- Laplace:

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right).$$



- Dirac Delta: It is zero-valued everywhere except at μ, yet integrates to 1

$$p(x) = \delta(x - \mu).$$

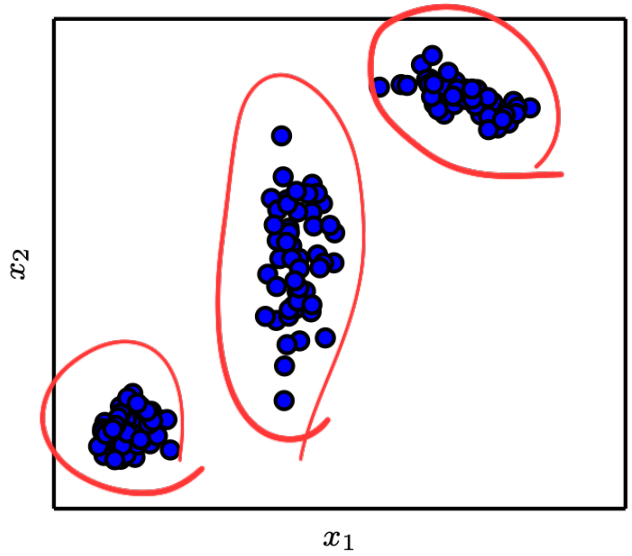
- Empirical Distribution:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

Mixture Distributions

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

- Gaussian mixture: $P(x|c = i)$ is Gaussian

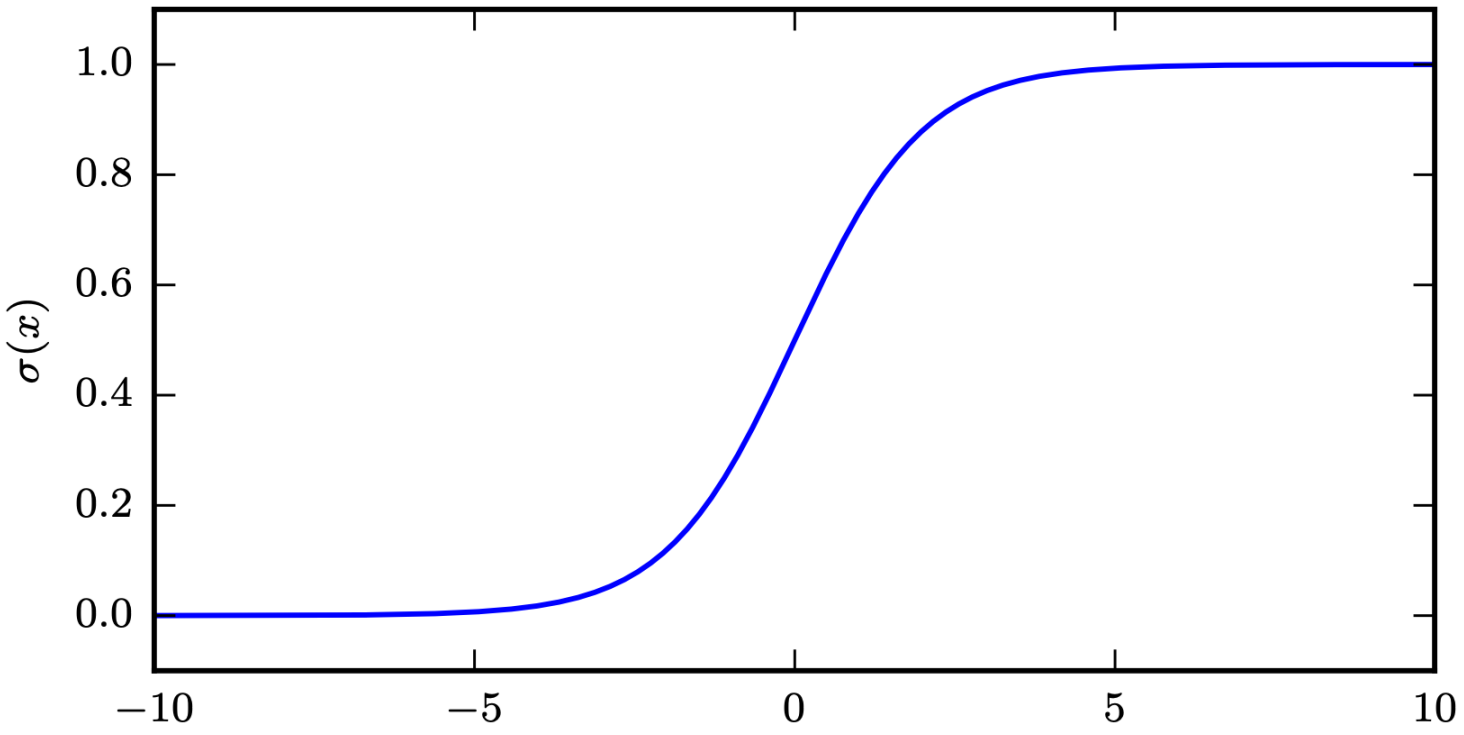


Gaussian mixture with three components

복잡 복잡이
가우시안 분포 더하기

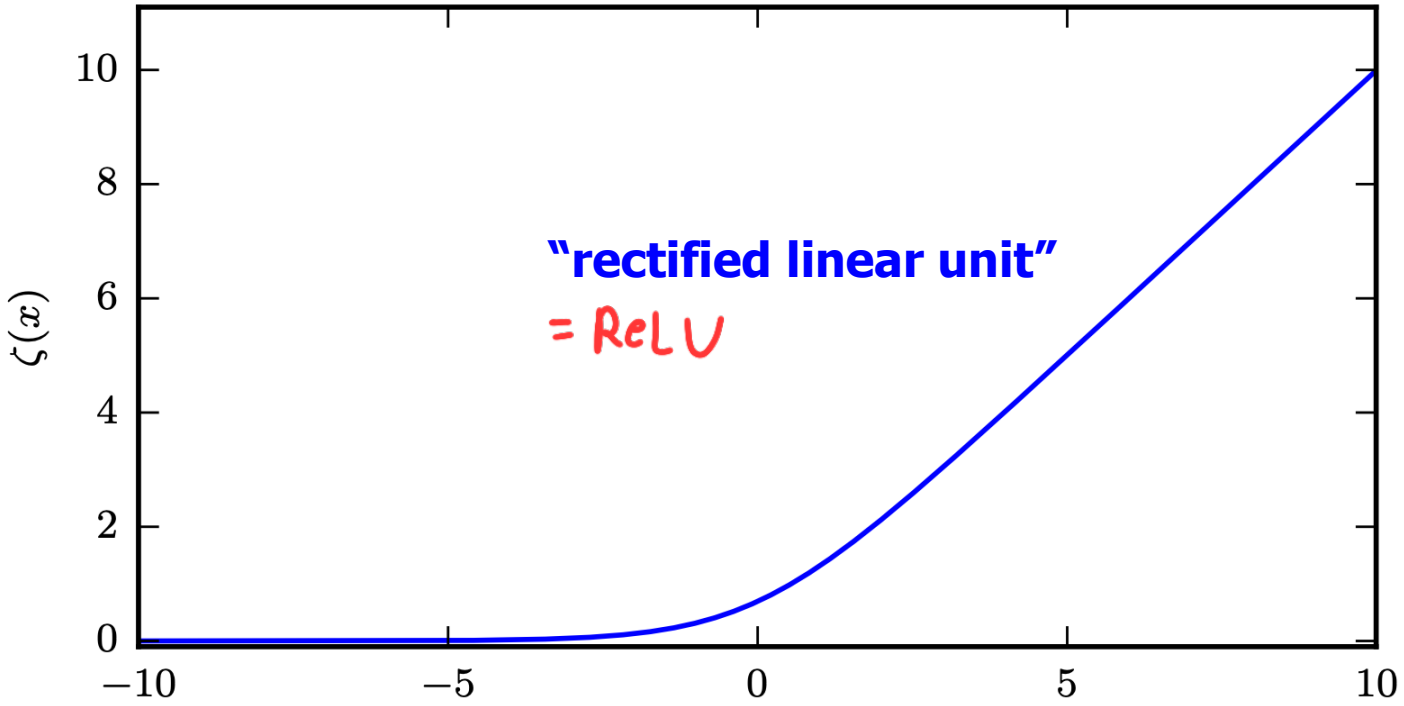
Logistic Sigmoid

- $\sigma(x) = \frac{1}{1+\exp(-x)}$



Softplus Function

- $\zeta(x) = \log(1 + \exp(x))$



- "softened" version of $x^+ = \max(0, x)$

Information Theory

- Information theory: quantifying how much information is present in a signal

정보의 양을 측정

- Learning that **an unlikely event has occurred** is more informative than learning that **a likely event has occurred**

확률적으로 적은 것들을 갖고오는게 더 의미

- Self-Information of x
  - $I(x) = -\log P(x)$

P(x) ⇑ → I(x) ⇓, P(x) ⇓ → I(x) ⇑

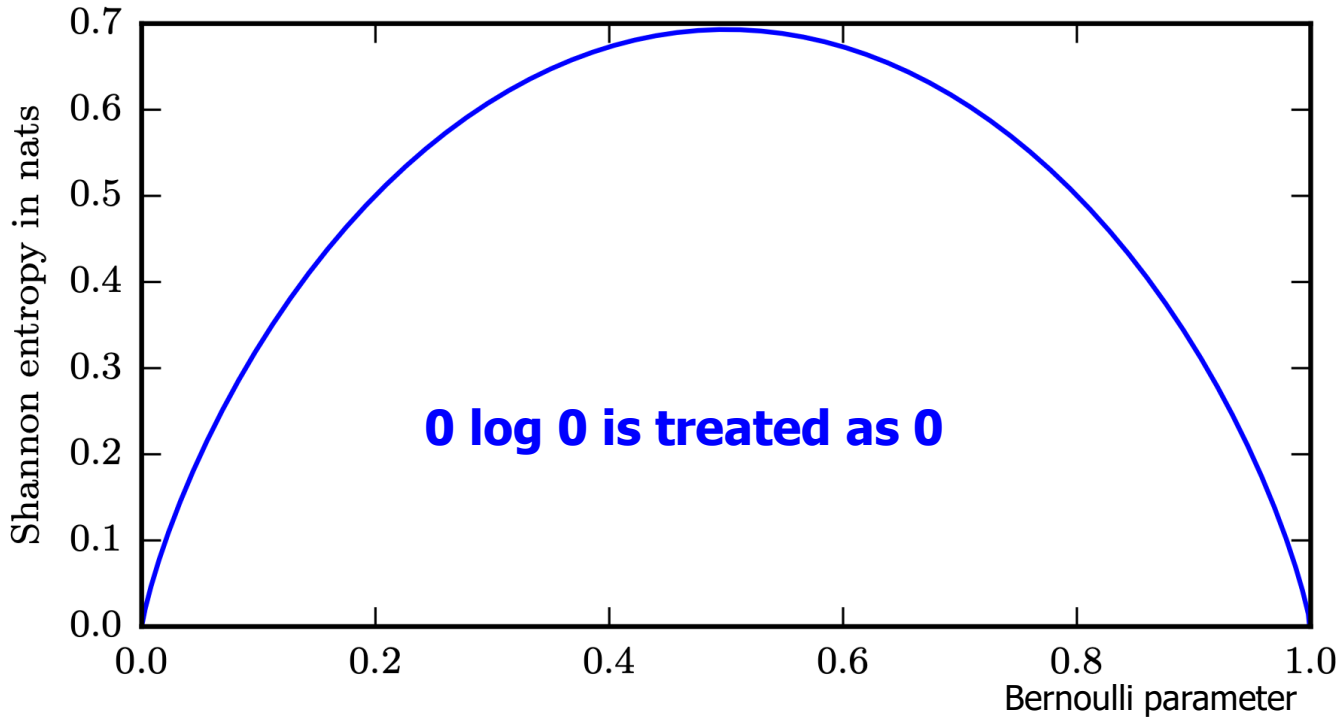
  - Intuition: minimum # of bits to express (encode) an event with probability P(x)

의미 → x값을 나타내기 위한 최소 bit

  - Rare event has a large information content

Information Theory: Entropy

- Entropy: expectation of self-information
  - Minimum expected # of bits to express a distribution
  - For Bernoulli variable,
    - $H(x) = -p \log p - (1-p)\log(1-p)$



KL Divergence

- Measure the difference of two distributions P(x) and Q(x)

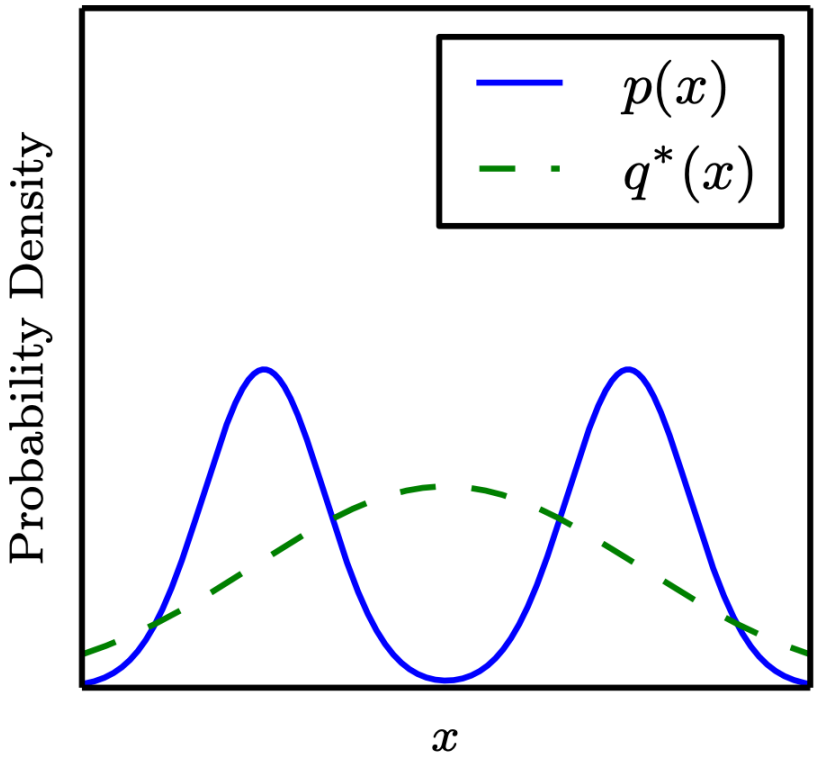
$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right].$$

- Properties
  - Always nonnegative: 0 if and only if P and Q are the same
    - Intuition: If x ~ P, the best (minimal) encoding is given by assigning log P(x) bits for each x
  - Not symmetric:

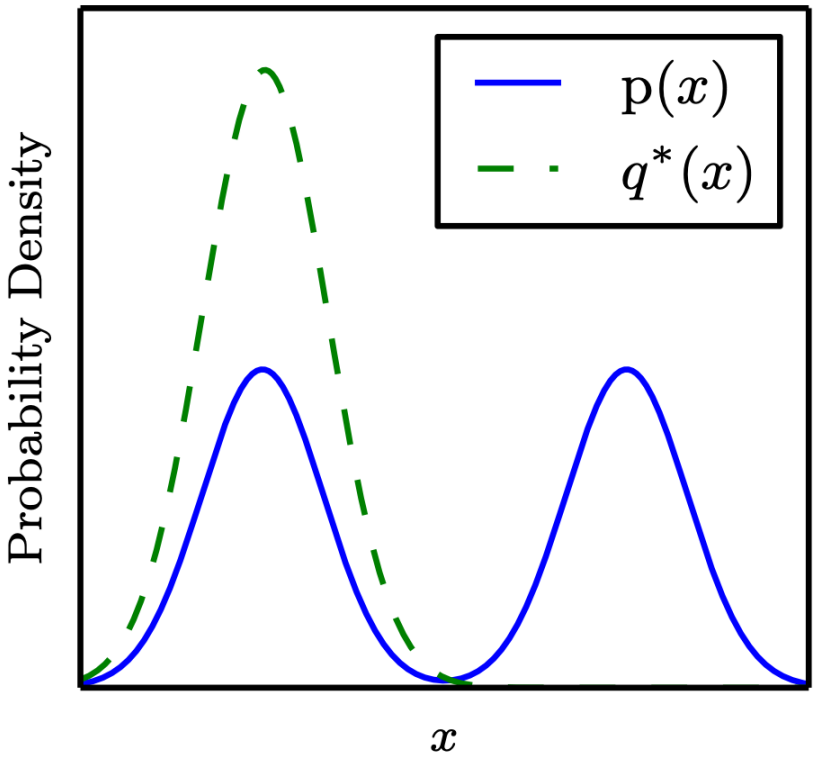
$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

KL is Asymmetric

$q^* = \text{argmin}_q D_{\text{KL}}(p\|q)$



$q^* = \text{argmin}_q D_{\text{KL}}(q\|p)$



Cross-Entropy

- Average # of bits needed to identify an event from the true distribution P, if we use a coding scheme optimized for unnatural distribution Q

- $H(P,Q) = H(P) + D_{KL}(P||Q) = -E_{x\sim P} \log Q(x)$

Entropy + KLD

P의 분포를 따를때 log Q의 평균

- Minimizing the cross-entropy w.r.t. Q is equivalent to minimizing the KL divergence

KLD↑ → CE↑

linear 함

Summary & Next

- "Remember" the key terminology / concepts that popularly used in Stats.
  - Population, Sample, Descriptive, Inferential, Parameters, Statistics, …
  - Mean, Median, Mode, Range/Quartile, Standard Dev.
  - Probability, Expectation, Covariance
  - Distributions
  - Self-Information, Entropy, KL-Divergence, Cross-Entropy

- Next?
  - Statistical methods (Pearson, …)

Thank you!

Instructor: Daejin Choi (djchoi@inu.ac.kr)