



# Data exploring, github, correlation analysis, probability distribution

---

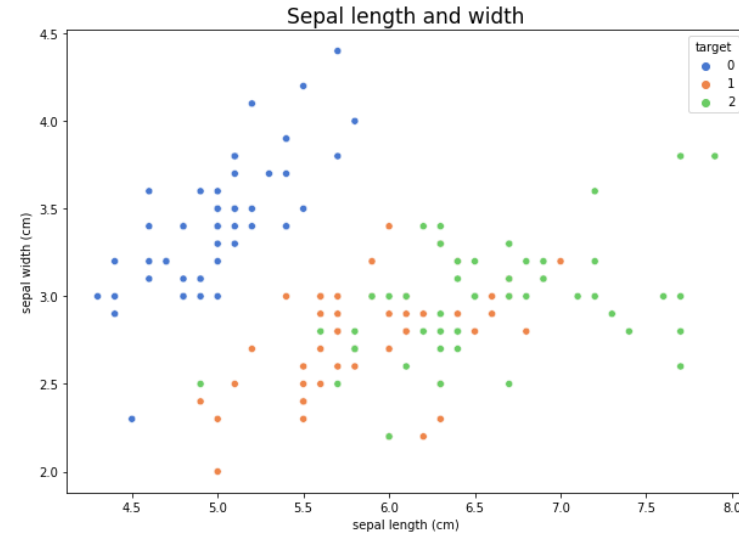
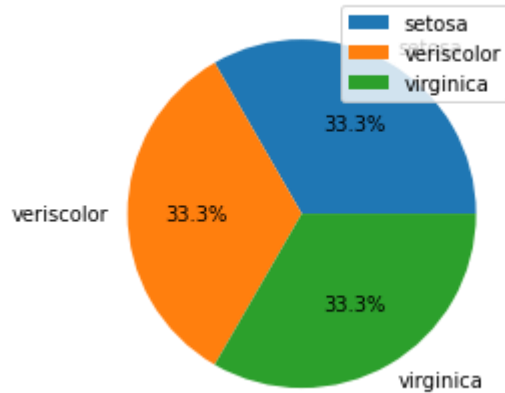
Advanced Python  
Stella Sangyoon Bae

2023. 06. 05



# 01 Data exploring (continued)

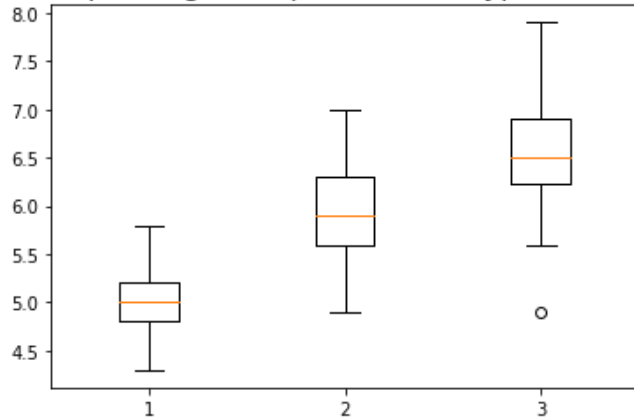
Pie chart among three types of iris



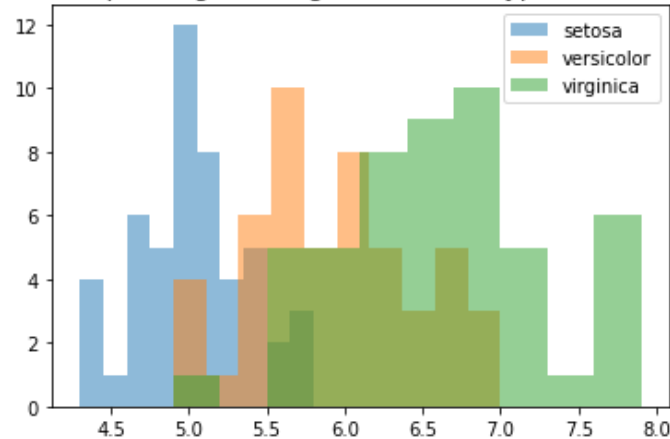
우리는 지금까지 한 종류의 데이터에 대해서만 plotting을 해 보았습니다.

그렇다면, 여러 종류의 데이터에 대해서 plotting 하려면 어떻게 해야 할까요?

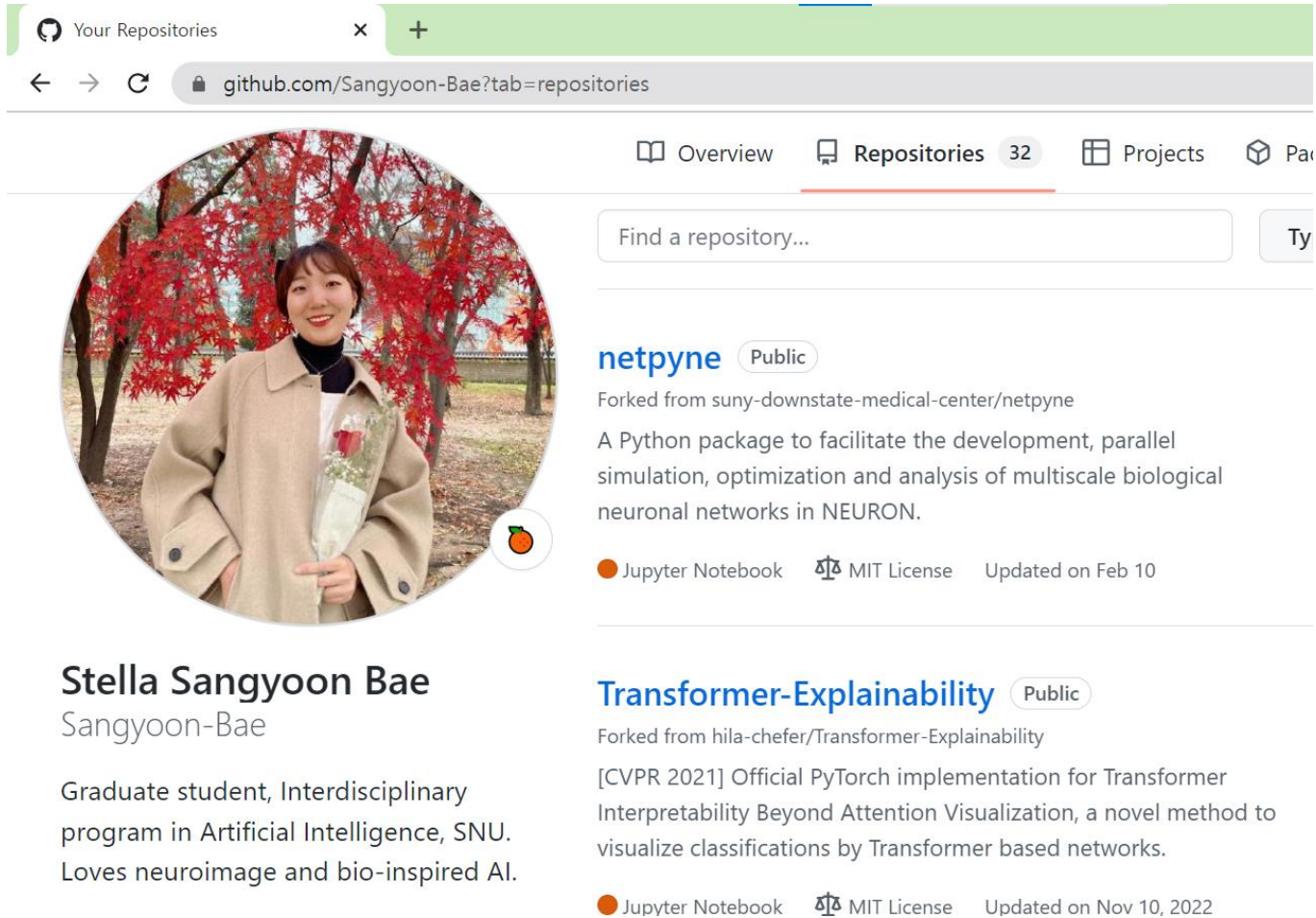
Sepal length boxplot of three types of iris



Sepal length histogram of three types of iris



## 02 Github



Stella Sangyoon Bae  
Sangyoon-Bae

Graduate student, Interdisciplinary program in Artificial Intelligence, SNU. Loves neuroimage and bio-inspired AI.

**netpyne** Public  
Forked from suny-downstate-medical-center/netpyne  
A Python package to facilitate the development, parallel simulation, optimization and analysis of multiscale biological neuronal networks in NEURON.  
Jupyter Notebook MIT License Updated on Feb 10

**Transformer-Explainability** Public  
Forked from hila-chefer/Transformer-Explainability  
[CVPR 2021] Official PyTorch implementation for Transformer Interpretability Beyond Attention Visualization, a novel method to visualize classifications by Transformer based networks.  
Jupyter Notebook MIT License Updated on Nov 10, 2022

일종의 원격 코드 저장소로, 많은 개발자들이 포트폴리오로 사용합니다.

우선 회원가입부터 해볼까요?



## 02 Github

---

깃헙에 원격으로 접속할 키를 만들어봅시다! 주피터 터미널을 열어서 아래와 같은 커맨드를 입력해봅시다.

```
ssh-keygen -t ed25519 -C "{name of key} {your email using in github}"
```


엔터를 쑥 치면 어디에 키가 저장되어 있는 지 알려줄거예요. 복사해서 그 파일로 가봅시다.

```
vi {path to the key}
```



# 02 Github

깃헙에 그 키를 등록해봅시다.

1.  **Stella Sangyoon Bae**  
Your personal account [Switch to another account](#)

Public profile

Account

Appearance

Accessibility

Notifications

Access

Billing and plans

Emails

Password and authentication

Sessions

SSH and GPG keys

Public p

Name

Stella Sangy

Your name may  
remove it at any

Public email

Select a ver

You have set yo  
uncheck "Keep

Bio

Graduate st  
Intelligence

## 2. SSH keys

This is a list of SSH keys associated with your account. Remove any keys that you do not recognize.

New SSH key

## 3. SSH keys / Add new

Title

Key type

Authentication Key


Key

Begins with 'ssh-rsa', 'ecdsa-sha2-nistp256', 'ecdsa-sha2-nistp384', 'ecdsa-sha2-nistp521', 'ssh-ed25519', 'sk-ecdsa-sha2-nistp256@openssh.com', or 'sk-ssh-ed25519@openssh.com'






# 02 Github

이제 깃헙을 사용할 준비가 되었어요! Repository를 만들어볼까요?



[Pull requests](#) [Issues](#) [Codespaces](#) [Marketplace](#) [Explore](#)


  


[Overview](#) [Repositories 32](#) [Projects](#) [Packages](#) [Stars 6](#)

Type ▾


Language ▾

Sort ▾

 New




Stella Sangyoon Bae





[netpyne](#) [Public](#)

Forked from suny-downstate-medical-center/netpyne

A Python package to facilitate the development, parallel simulation, optimization and analysis of multiscale biological neuronal networks in NEURON.

 [Star](#) ▾

 Jupyter Notebook

 MIT License

Updated on Feb 10

[Transformer-Explainability](#) [Public](#)



# 02 Github

이제 깃헙을 사용할 준비가 되었어요! Repository를 만들어볼까요?

Owner \*

Sangyoon-Bae ▾

/

Repository name \*

name ✓

Great repository names are short and memorable. Need inspiration? How about [probable-octo-couscous](#)?

Description (optional)

☒ Public

Anyone on the internet can see this repository. You choose who can commit.

☐ Private

You choose who can see and commit to this repository.

Initialize this repository with:

Skip this step if you're importing an existing repository.

☒ Add a README file

This is where you can write a long description for your project. [Learn more.](#)

Add .gitignore


Choose which files not to track from a list of templates. [Learn more.](#)

.gitignore template: None ▾

Choose a license

A license tells others what they can and can't do with your code. [Learn more.](#)

License: MIT License ▾

This will set  main as the default branch. Change the default name in your [settings](#).

ⓘ You are creating a public repository in your personal account.

Create repository



## 02 Github

---

깃헙 설정을 해볼까요? 먼저 초기화입니다.

```
git init
```

```
PS C:\Users\stell\GIRE basic python> git init  
Initialized empty Git repository in C:/Users/stell/GIRE basic python/.git/
```

이런 화면이 뜨면 성공한 거예요. 이제 사용자 등록을 해봅시다.

```
git config --global user.name {your user name}  
git config --global user.email {your email}
```





## 02 Github

원격 저장소와 우리 컴퓨터의 저장소를 연동해봅시다.

```
git clone {remote repository address}
```

Sangyoon-Bae / **GIRE\_basic\_python** Public

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags

Sangyoon-Bae sample code

- Iris\_sample.ipynb sample code
- README.md Initial commit

README.md

### GIRE\_basic\_python

Basic python course in GIRE

Local Codespaces New

Clone

HTTPS SSH GitHub CLI

git@github.com:Sangyoon-Bae/GIRE\_basic\_python

Use a password-protected SSH key.

Open with GitHub Desktop

Open with Visual Studio

Remote address는 여기서 얻을 수 있어요!



## 02 Github

---

이제 저 폴더 안에서 작업을 해봐요.

`cd {folder name}` #cd란? 현재 위치를 변경해주는 명령어

`git status`

```
On branch main
Your branch is up to date with 'origin/main'.
Untracked files:
  (use "git add <file>..." to include in what will be committed)
    .ipynb_checkpoints/
    Iris_sample.ipynb

nothing added to commit but untracked files present (use "git add" to track)
```

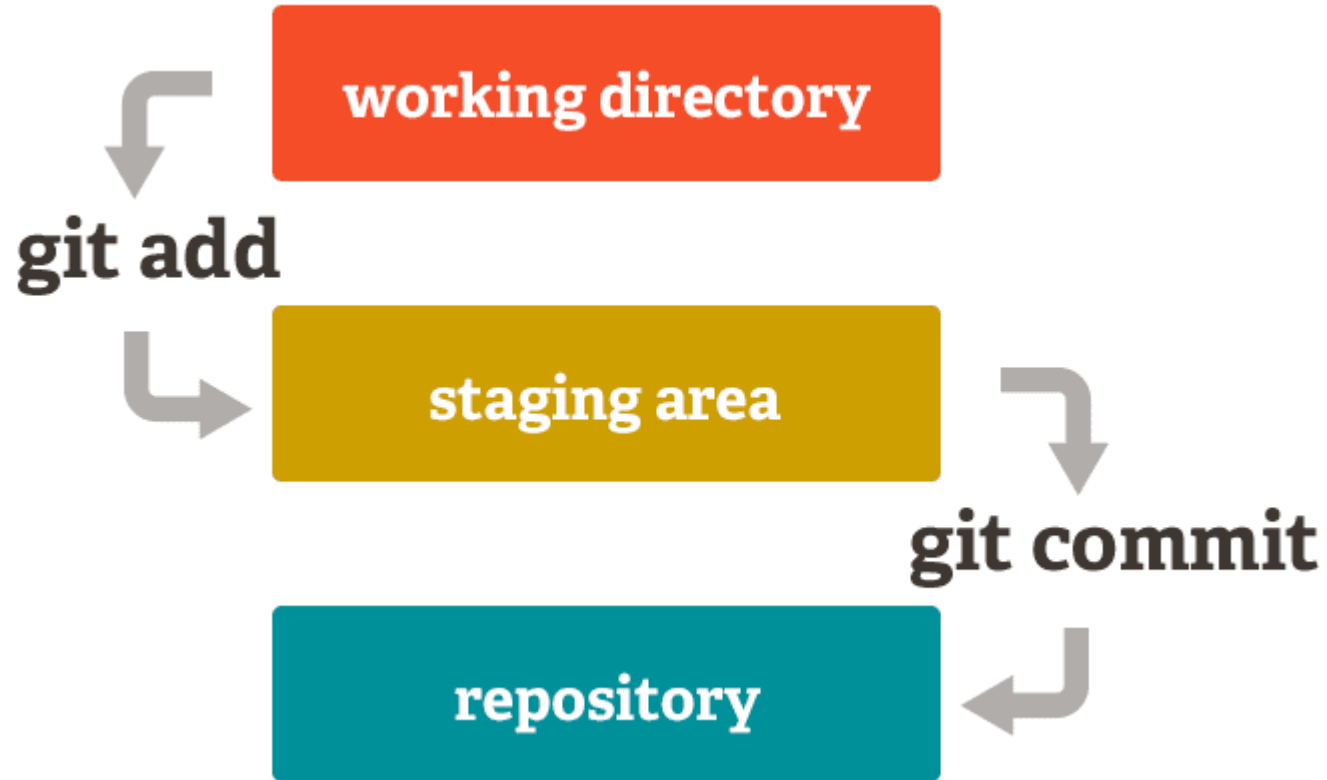
아직 tracking이 안 되었네요.



## 02 Github

---

Working directory (현재 있는 디렉토리) 에서 git add로 staging area에 올리고,  
git commit으로 repository에 올려야 해요.



## 02 Github

---

일단 working directory -> staging area

```
git add {file you want to upload on the github repository}
```

꼭 지금 상태가 어떤 지 확인해봐야 합니다.

```
git status
```

```
On branch main
Your branch is up to date with 'origin/main'.

Changes to be committed:
  (use "git restore --staged <file>..." to unstage)
    new file:   Iris_sample.ipynb

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    .ipynb_checkpoints/
```

잘 tracking 하네요.



## 02 Github

tracked file을 staging area에서 repository로 commit 해봅시다.

```
git commit -m "message you want to write"
```

Commit한 파일을 최종 업로드 해볼게요.

```
git push origin main
```

```
Enumerating objects: 4, done.
Counting objects: 100% (4/4), done.
Delta compression using up to 8 threads
Compressing objects: 100% (3/3), done.
Writing objects: 100% (3/3), 137.28 KiB | 3.71 MiB/s, done.
Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/Sangyoon-Bae/GIRE_basic_python.git
b1676e1..f4d5475  main -> main
```

Sangyoon-Bae / GIRE\_basic\_python Public

<> Code Issues Pull requests Actions Projects Wiki

main 1 branch 0 tags



Sangyoon-Bae sample code



Iris\_sample.ipynb

sample code



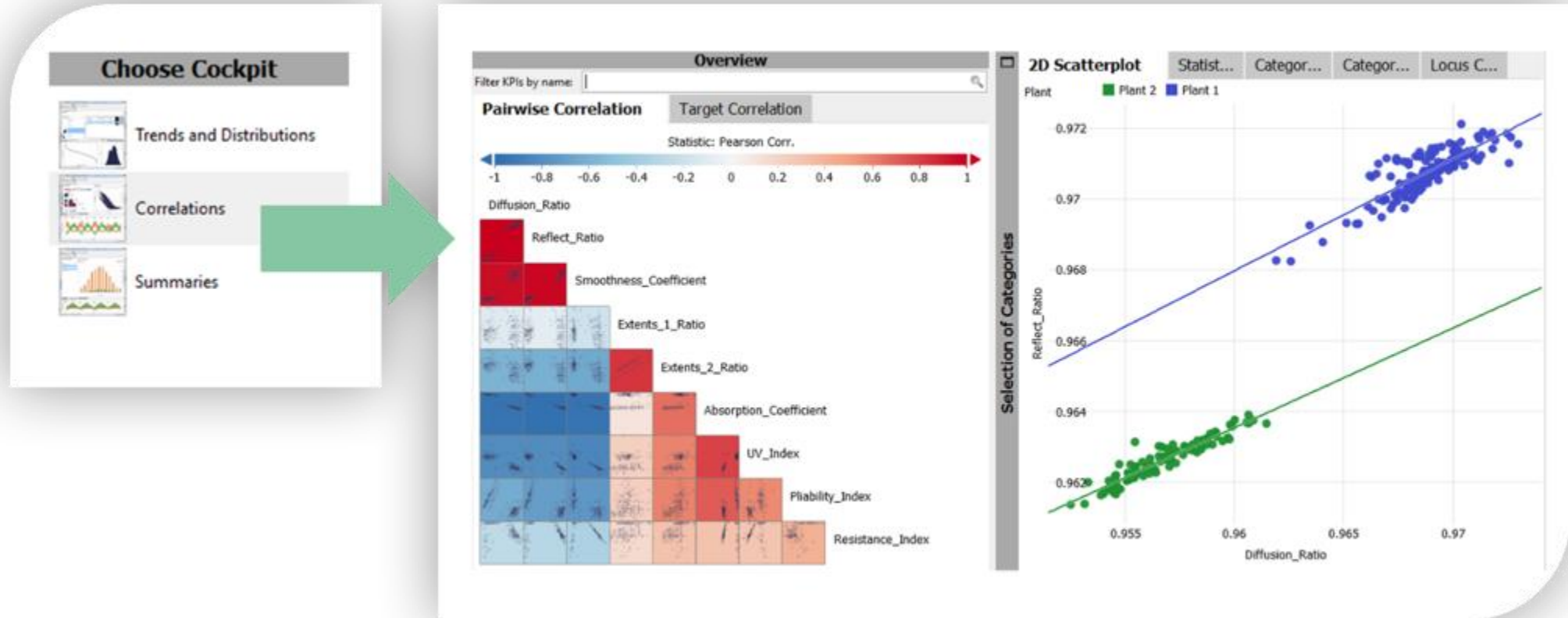
README.md

Initial commit

깃헙 레포지토리에서도 잘 올라가 있는 걸 확인할 수 있습니다.



# 03 correlation analysis

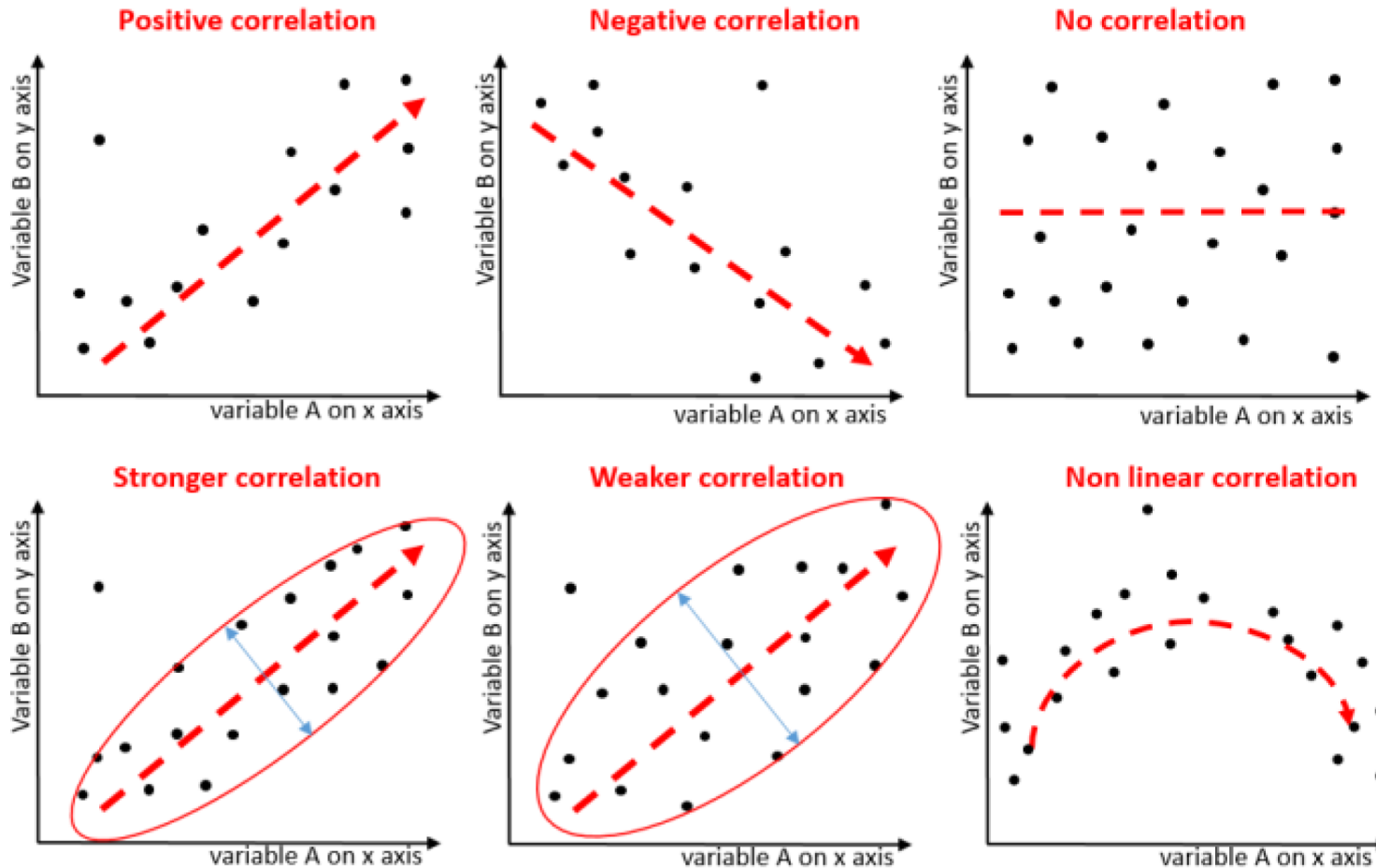


무엇을 분석할 지 알기 위해서 하는 분석입니다.

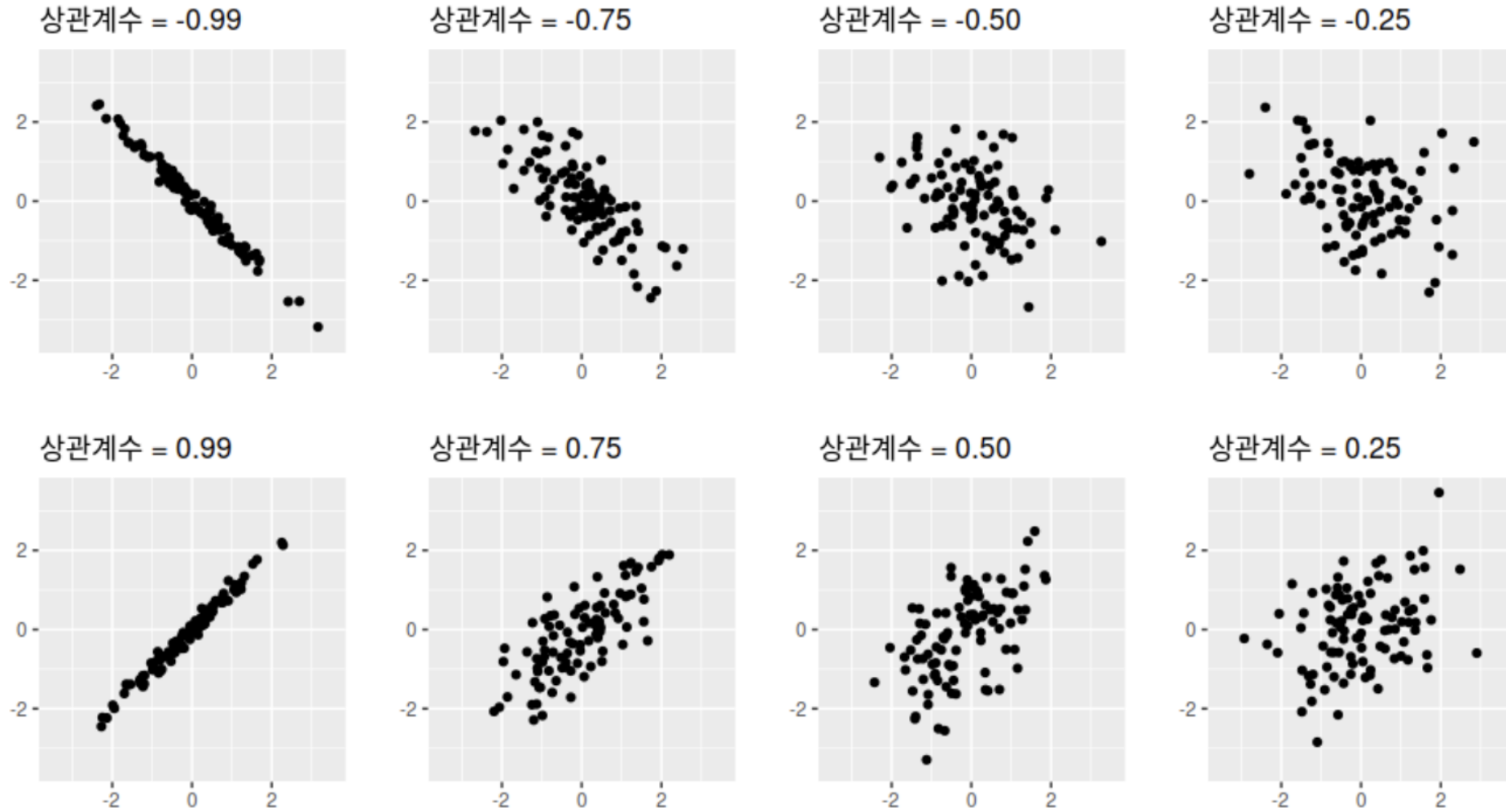
서로 관련이 없는 feature-target 사이 모델을 만들면 아무리 모델을 잘 만들어 봤자 소용이 없겠죠!



# 03 correlation analysis

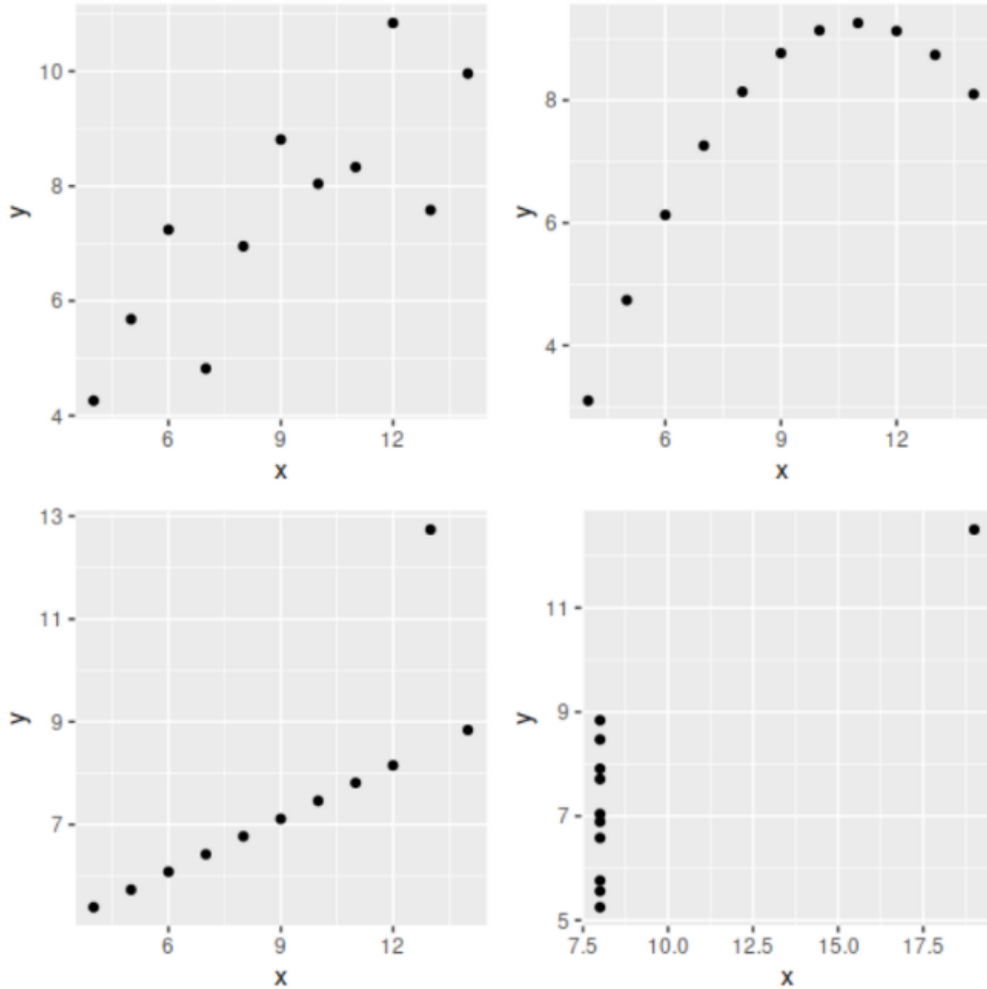


# 03 correlation analysis





## 03 correlation analysis



### WARNING!

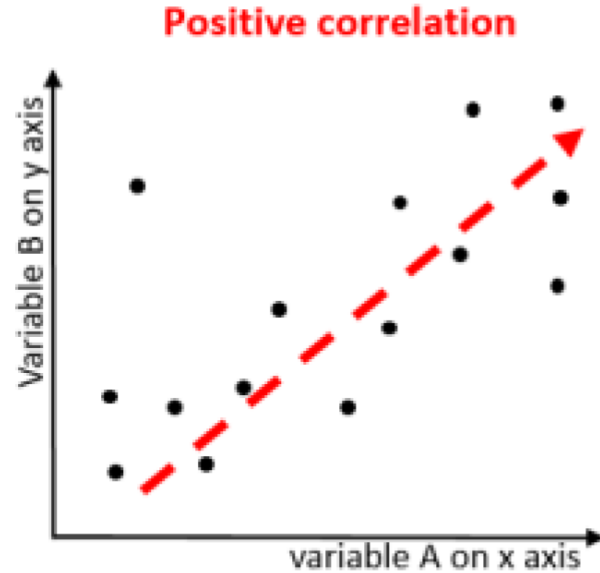
4개의 그래프에서 correlation coefficient는 모두 0.82이지만, 데이터의 분포 형태는 각기 다름을 알 수 있습니다.

상관계수에 너무 의존하지 않고 그래프를 그려보는 습관을 들이도록 합시다.

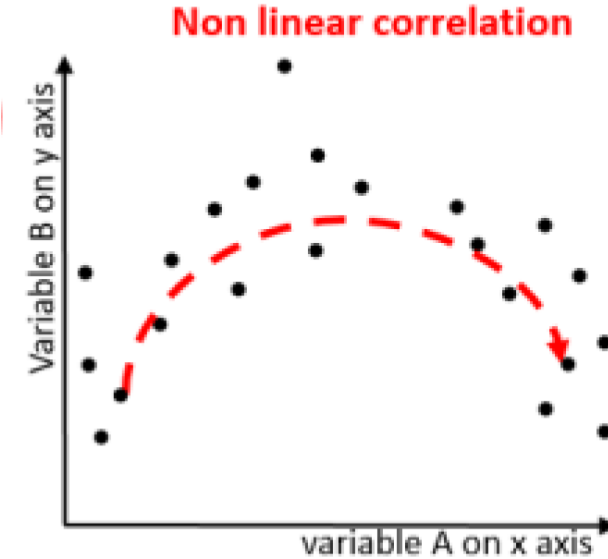


Anscombe, F. J. (1973). Graphs in statistical analysis. *The american statistician*, 27(1), 17-21.

## 03 correlation analysis



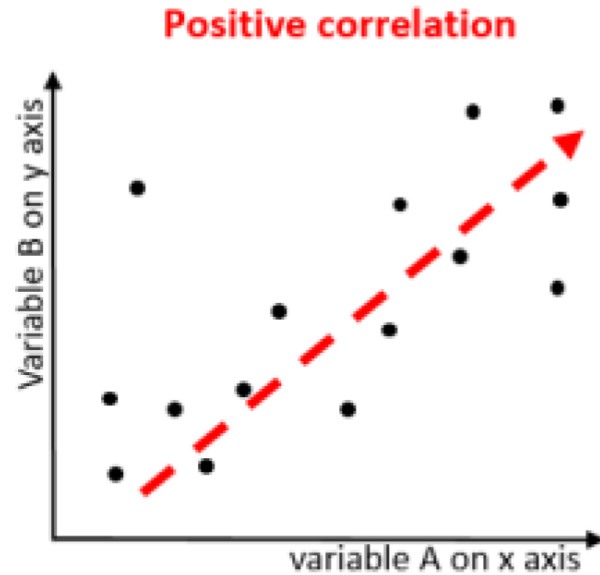
그래프가 선형이라서 '선형 상관관계' 라고 부릅니다.  
Pearson 상관계수를 이용해서 측정합니다.



그래프가 선형이 아니라서 '비선형 상관관계' 라고 부릅니다.  
Spearman 상관계수를 이용해서 측정합니다.  
순서값의 Pearson 상관계수라고 생각하면 됩니다.  
그래서 한 변수가 증가할 때 다른 변수가  
감소하는 지 증가하는 지 정도만 봅니다.



## 03 correlation analysis



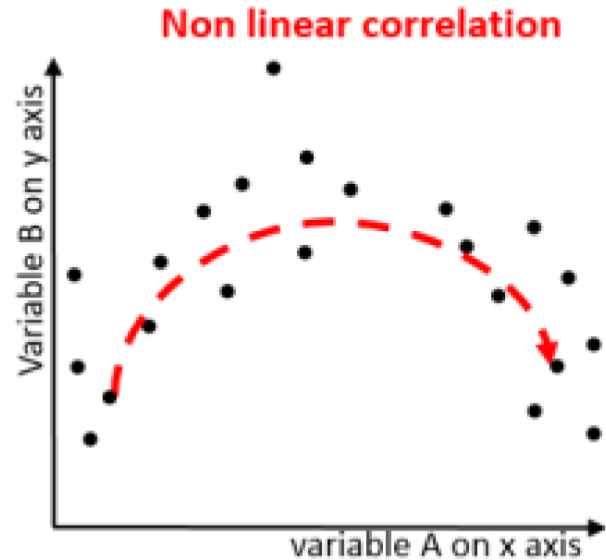
Strength of Association	Coefficient, r	
	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

그래프가 선형이라서 '선형 상관관계' 라고 부릅니다.

Pearson 상관계수를 이용해서 측정합니다.



## 03 correlation analysis



### Grading Standards

$\rho = 0$
$0 <  \rho  \leq 0.19$
$0.20 \leq  \rho  \leq 0.39$
$0.40 \leq  \rho  \leq 0.59$
$0.60 \leq  \rho  \leq 0.79$
$0.80 \leq  \rho  \leq 1.00$
1.00

### Correlation Degree

no correlation
very weak
weak
moderate
strong
very strong
monotonic correlation

그래프가 선형이 아니어서 '비선형 상관관계' 라고 부릅니다.

Spearman 상관계수를 이용해서 측정합니다.

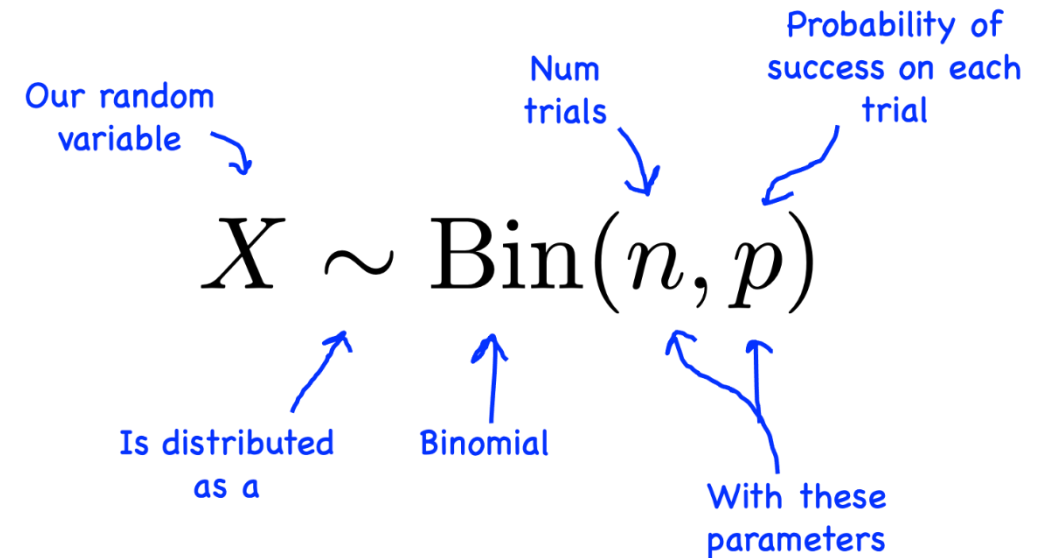
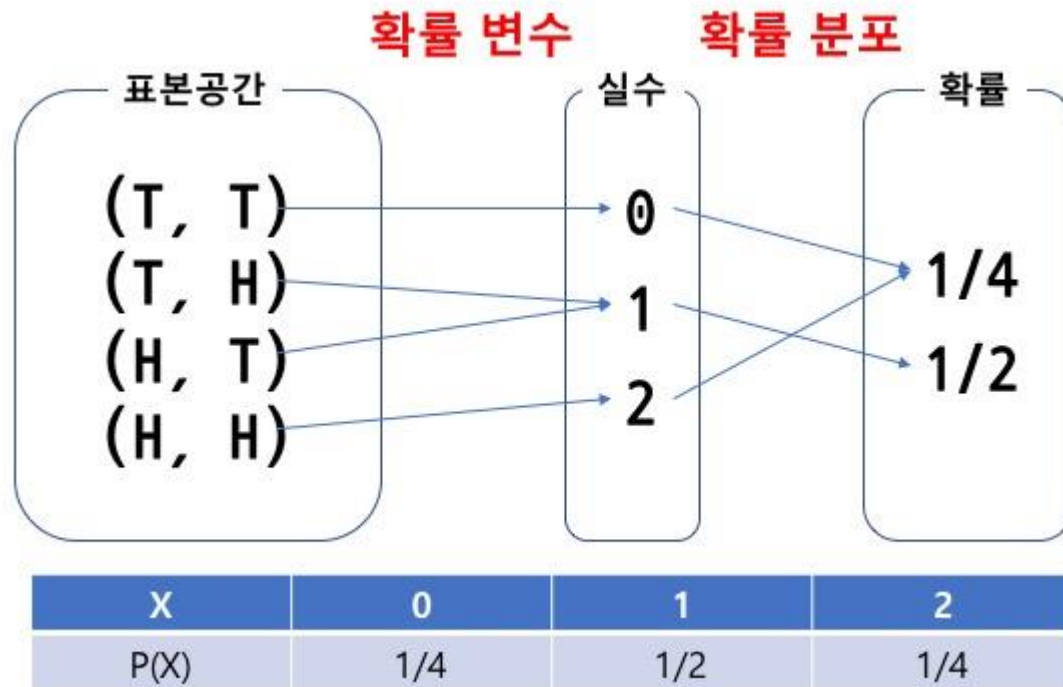
순서값의 Pearson 상관계수라고 생각하면 됩니다.

그래서 한 변수가 증가할 때 다른 변수가  
감소하는 지 증가하는 지 정도만 봅니다.



# 04 Probability distribution

확률 변수 (random variable)?

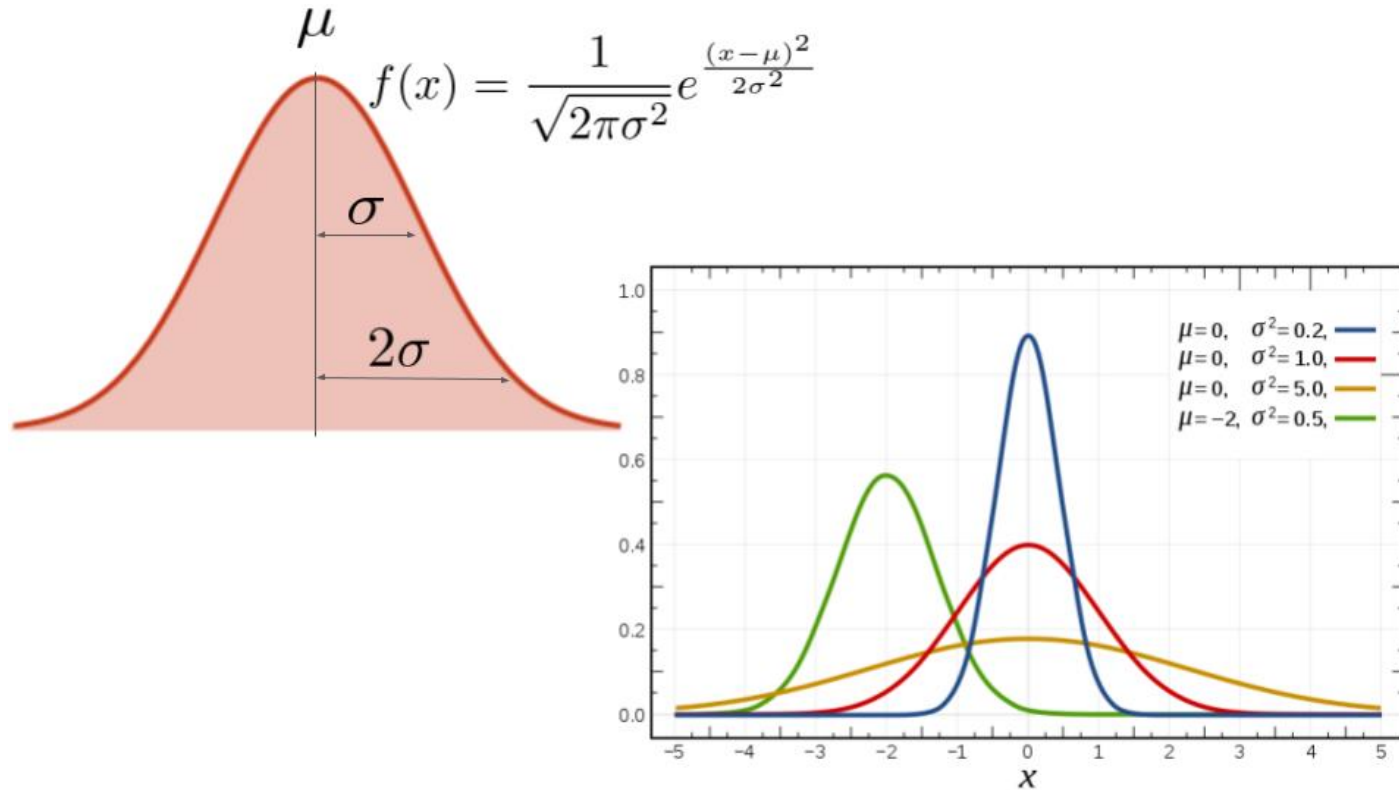


Random도 아니고 variable도 아님에 유의!



## 04 Probability distribution

확률 변수가 특정한 값을 가질 확률을 나타내는 함수

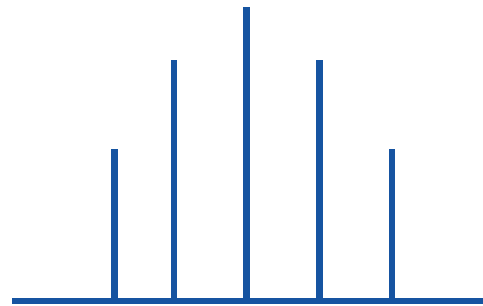


데이터의 형태를 알 수 있습니다.



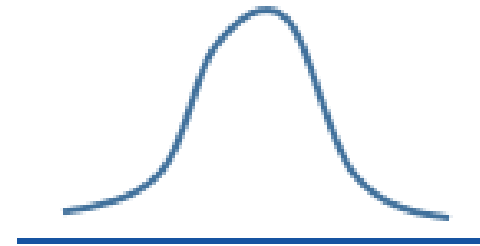
## 04 Probability distribution

---



이산확률분포

Discrete

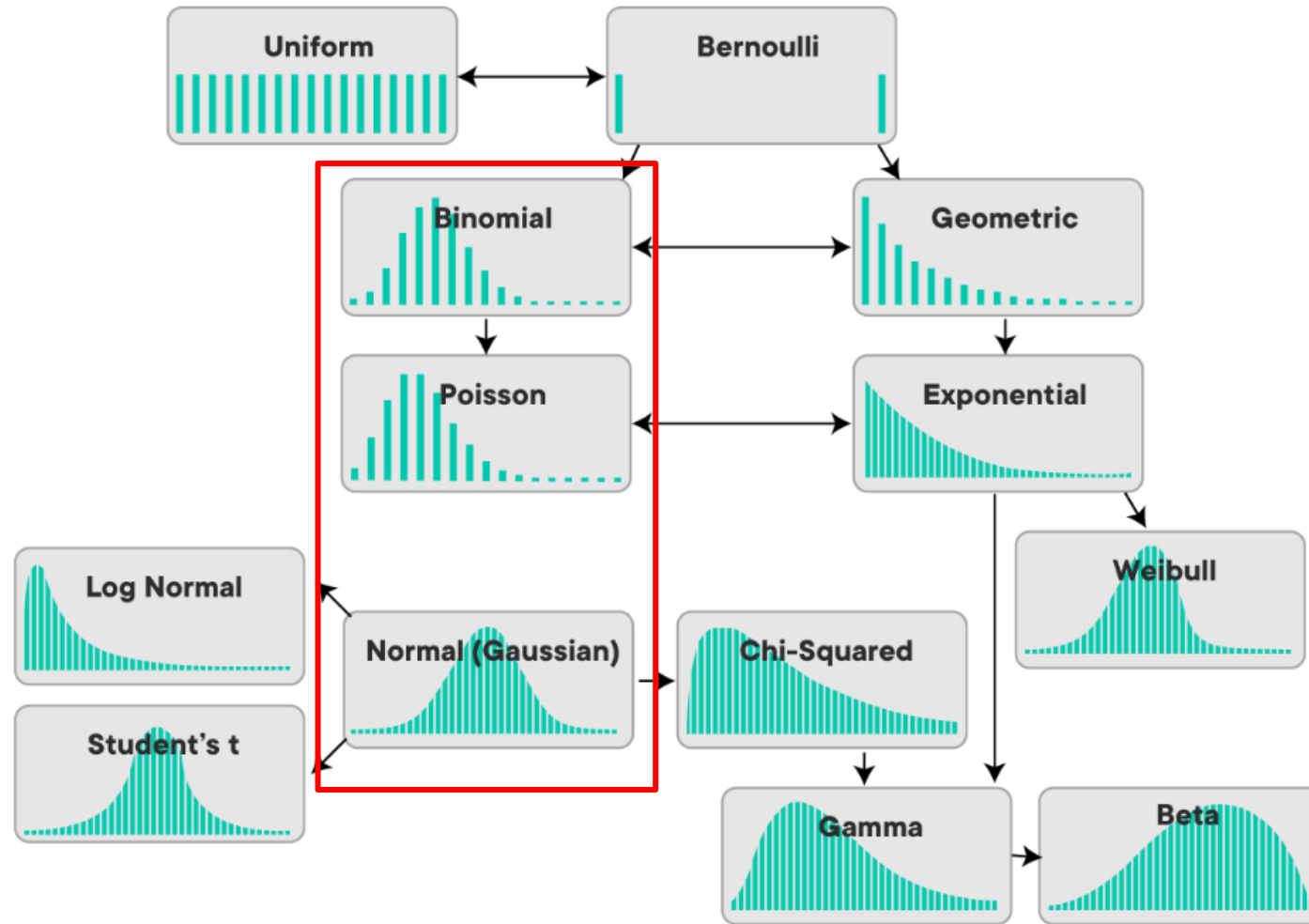


연속확률분포

Continuous



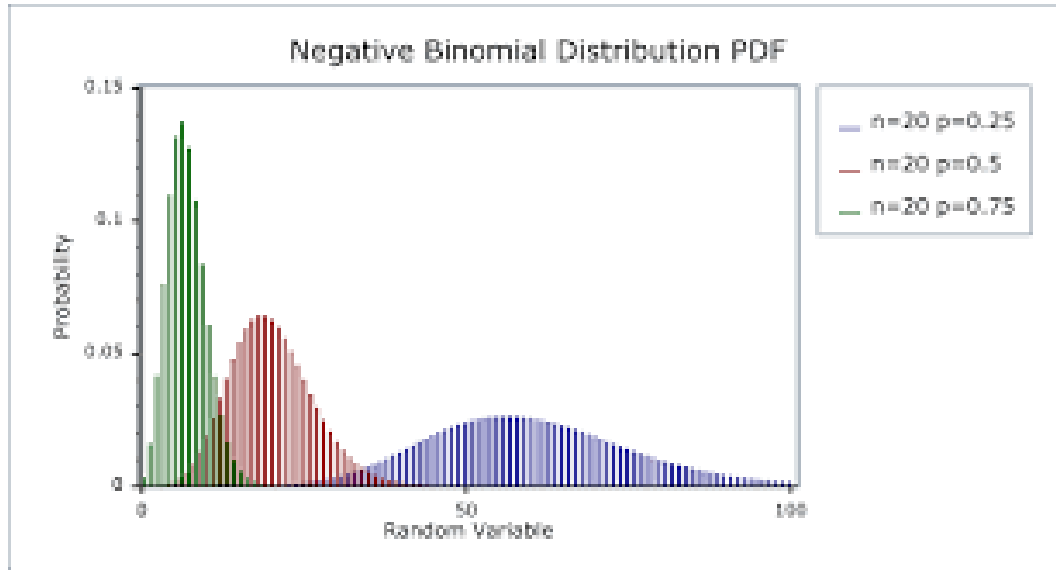
# 04 Probability distribution





# 04 Probability distribution (1) binomial distribution

연속된 n번의 독립적 시행에서 각 시행이 확률 p를 가질 때의 이산 확률 분포

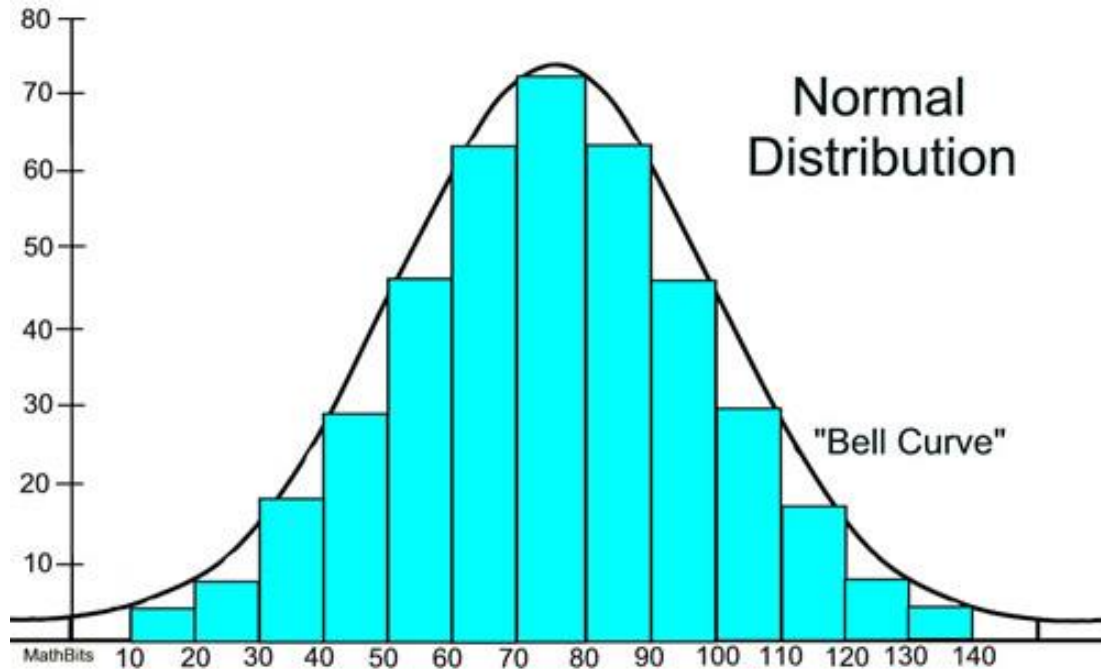


$$P(X = k) = \frac{n!}{k!(n - k)!} p^k (1 - p)^{n-k}$$

이 때 성공한 횟수를 X라는 확률변수로 나타내고, k번 성공할 확률은 위와 같이 나타냅니다.



## 04 Probability distribution (2) normal distribution



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

$\mu$  = Mean

$\sigma$  = Standard Deviation

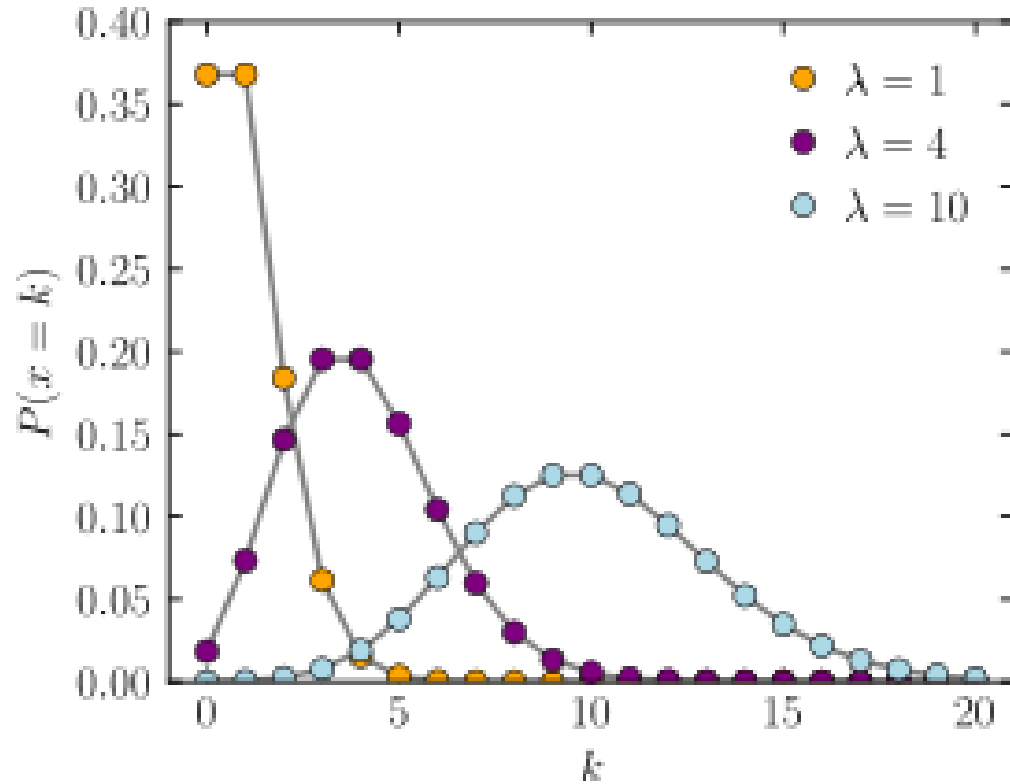
$\pi \approx 3.14159 \dots$

$e \approx 2.71828 \dots$

연속 확률 분포의 하나입니다. 정규분포는 수집된 자료의 분포를 근사하는 데에 자주 사용되며, 이것은 중심극한정리에 의하여 독립적인 확률변수들의 평균은 정규분포에 가까워지는 성질이 있기 때문입니다.



## 04 Probability distribution (3) Poisson distribution



$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률 분포

