S_04 Statistical Test

Statistical Analysis

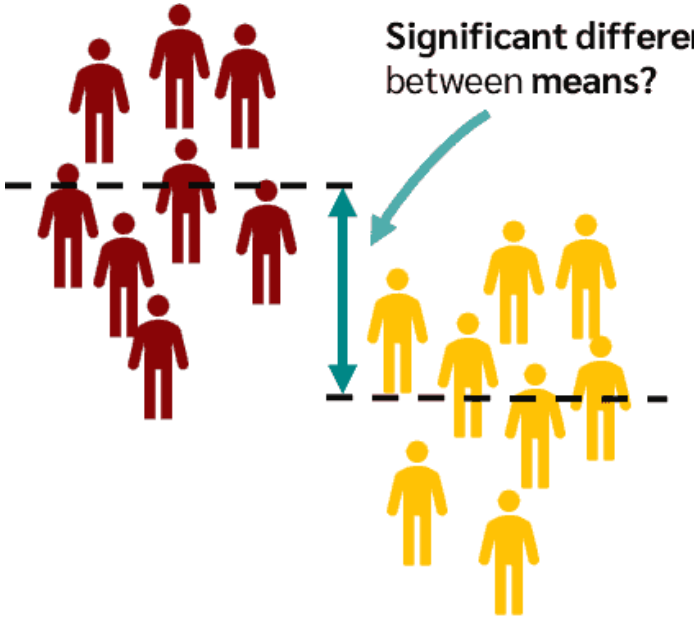
t-Test

- 두 그룹의 평균 간에 유의미한 차이가 있는지 테스트하는 statistical test
- Types of t-Test
  - One sample t-test
  - Independent sample t-test
  - Paired-sample t-test

One sample
t-test

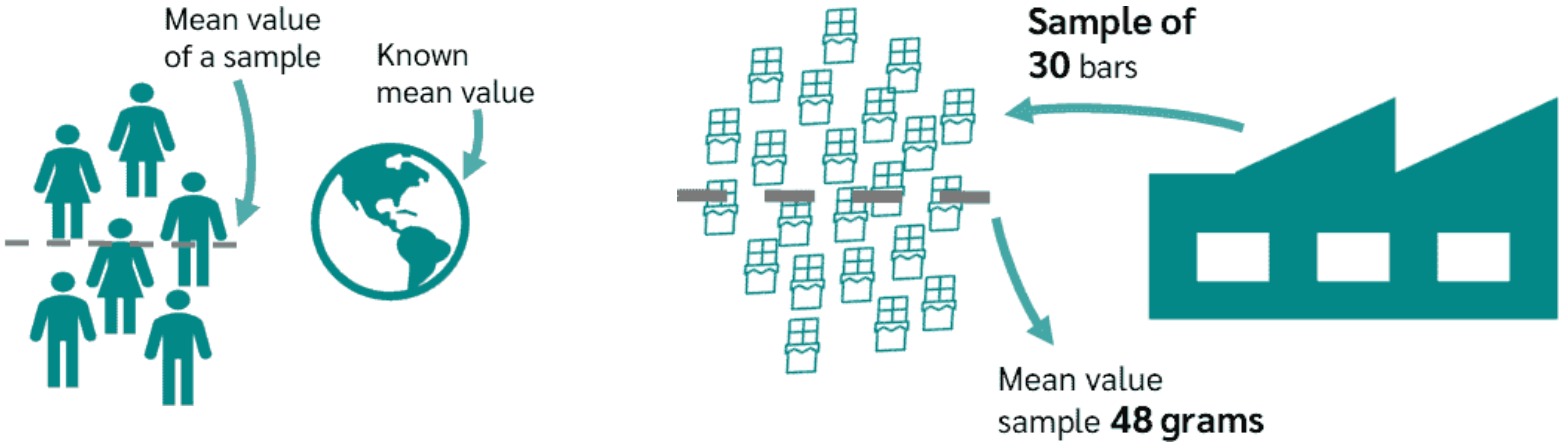
Independent
samples t-test

Paired
samples t-test



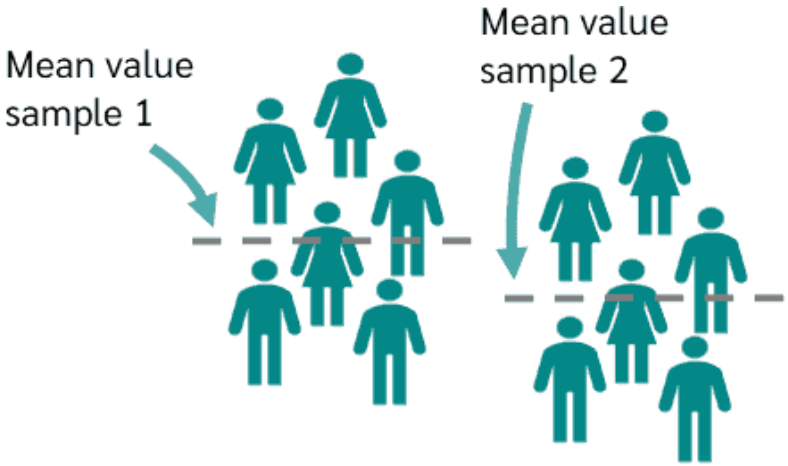
One Sample t-Test

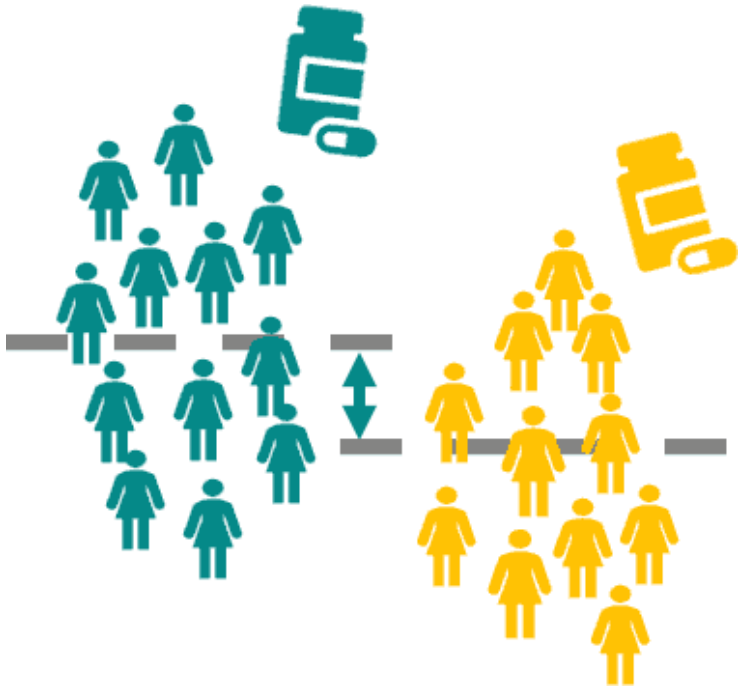
- To compare the mean of a sample with a **known** (population's) reference mean
- ex) 한 초콜릿 바 제조업체가 자사 초콜릿 바의 평균 무게가 50g이라고 주장합니다. 이를 확인하기 위해 30개의 바 샘플을 채취하여 무게를 측정합니다. 이 샘플의 평균값은 48그램입니다.
- One Sample t-Test
  - Known population mean: 50g
  - Sample mean:48g
  - H0: Sample mean is not different from the known population mean.



Independent Sample t-Test

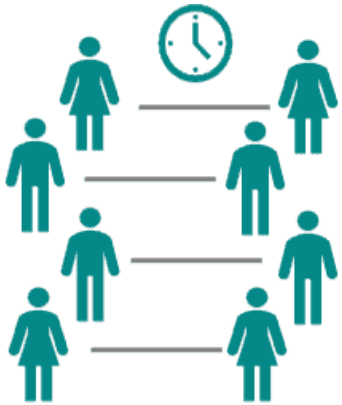
- To compare the means of two independent groups or samples
- ex) 진통제 A와 B의 효능을 비교
  - Group A: 30명, 진통제 A 복용
  - Group B: 30명, 진통제 B 복용
  - H0: 양쪽 group 간의 진통제 효능의 차이가 없다.

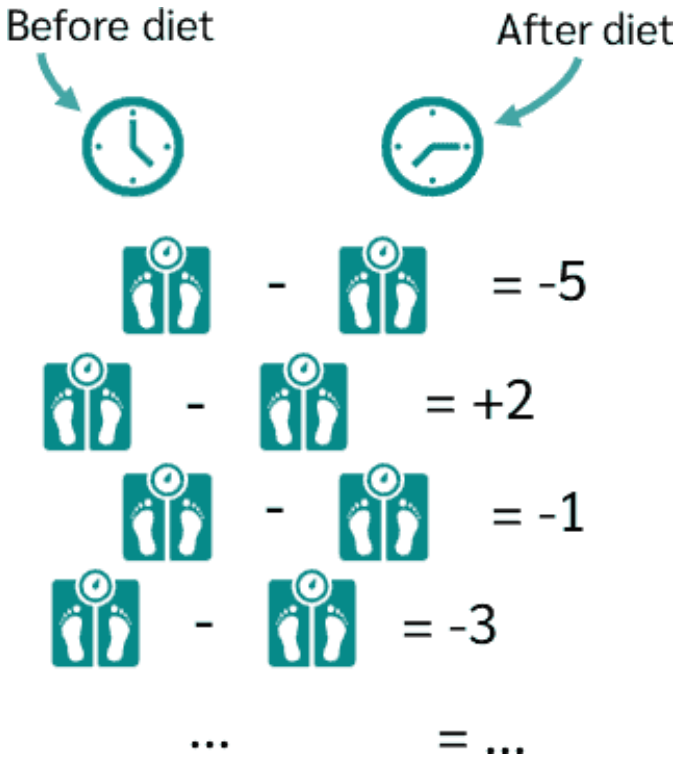




Paired Sample t-Test

- Dependent samples is used to compare the means of two dependent groups.
- Dependent group은 보통 같은 사람들로 구성된 한 group이 두 개의 다른 조건으로 실험하는 경우가 많음 (within group, within subject)
- ex) Diet의 효과 측정을 위해 30명의 group을 선택
  - 두 group: 같은 사람들을 Diet 전 체중 측정, Diet 후 체중 측정
  - H0: Before group과 After group의 체중에는 차이가 없다

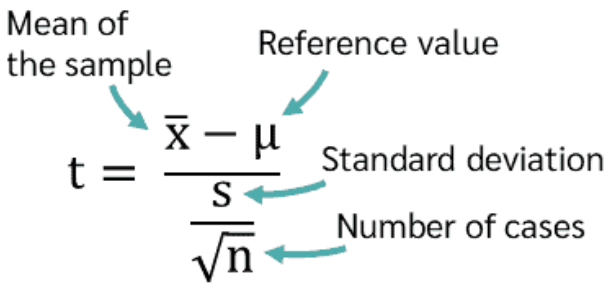




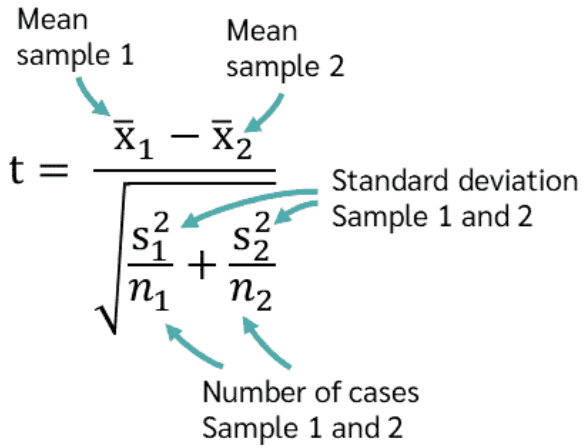
Computing t

$$t = \frac{\text{Difference between mean values}}{\text{Standard deviation from the mean (Standard error)}}$$

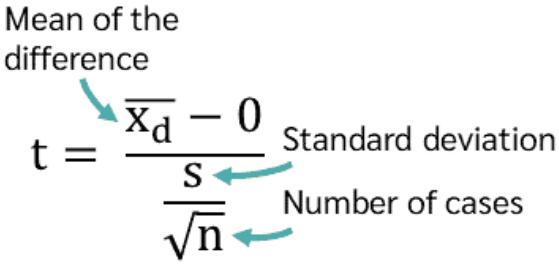
One sample t-test:



Independent Sample t-test:



Paired Sample t-test:



Critical t-Value

- Critical t-Value
  - H0를 reject할 수 있는 최대 t 값
  - 즉, t 값이 critical t-Value보다 작으면 H0를 reject할 수 있다.
  - t-value table에서 df (Degrees of freedom) 와 $(1 - \alpha)$ 값으로 결정

Table t-value

Area two-tailed

Significance level: 5%

df	0	0.5	0.6	0.7	0.8	0.9	0.95	0.98	0.99	0.998	0.999
1	0	1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0	0.816	1.061	1.386	1.886	2.92	4.303	6.965	9.925	22.327	31.599
3	0	0.765	0.978	1.25	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0	0.741	0.941	1.19	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	0	0.727	0.92	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0	0.718	0.906	1.134	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	0	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0	0.706	0.889	1.108	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	0	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	0	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.93	4.318

Degrees of freedom: 9

Degrees of freedom

One sample t-test	Independent samples t-test	Paired samples t-test
$\mathrm{df} = n - 1$	$\mathrm{df} = n_1 + n_2 - 2$	$\mathrm{df} = n - 1$

p-Value from t-Value

- t-distribution에서 t의 값을 가지는 확률
- $p\_value < \alpha$ 이면 H0를 reject할 수 있음

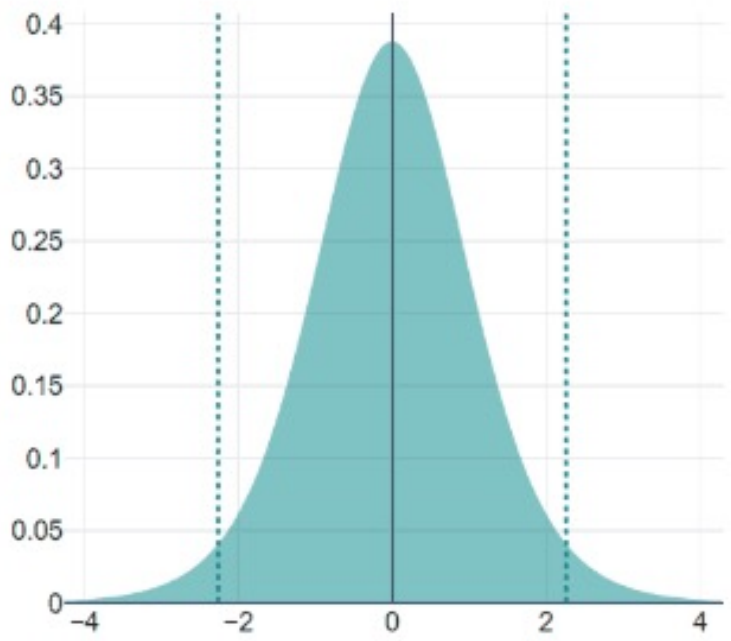
One-tailed | Two-tailed

Probability of error

t-Value	df	p-Value
2.5	9	= 0.0339

Critical t-Value

alpha	df	t-Value
0.05	9	= 2.262



CODE (1/2)

- https://github.com/iklee99/StatCode
- 05_t_Test.py

  - **One-sample t-test, t_test_one_sample.csv**
    - t-statistic: 1.29, p-value: 0.207
    - H0 retained (p-value > 0.05), 즉, Sample mean이 Known population mean (50)과 유의미하게 다르지 않음

Sample
54.483570765056164
51.308678494144075
55.23844269050346
59.61514928204013
50.82923312638332
50.829315
59.896064
55.837173645764544
49.65262807032524
54.712800217929825
49.68291153593769
49.671351232148716
53.20981135783017

CODE (2/2)

- **Independent samples t-test**
  - t-statistic: 3.49, p-value: 0.00094
  - H0 rejected (p-value < 0.05), 즉, 두 독립된 집단의 mean이 유의미하게 다름

- **Paired samples t-test**
  - t-statistic: -4.84, p-value: 0.00004
  - H0 rejected (p-value < 0.05), 즉, 두 연관된 집단(사전 테스트와 사후 테스트)의 mean이 유의미하게 다름

Group	Value
Group1	51.99146694
Group1	64.26139092
Group1	54.93251388
...	...
Group2	47.60412881
Group2	49.07170512
Group2	44.46832513
Group2	44.01896688
...	...

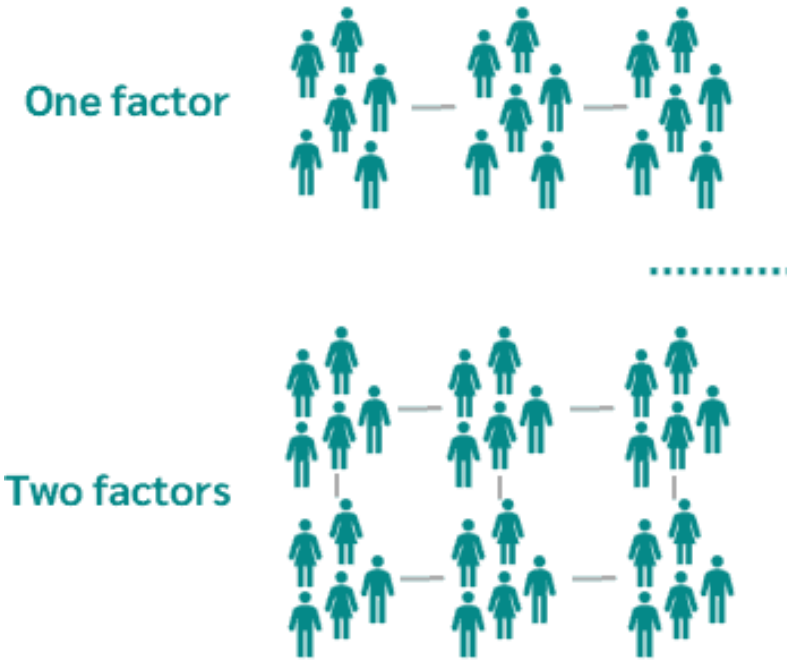
t_test_independent_samples.csv

PreTest	PostTest
50.48538775	54.06745164
54.84322495	55.02445004
46.48973453	51.29532315
48.36168927	47.55798714
48.03945923	51.21317342
42.68242526	49.06333651
51.48060139	51.49952874
51.30527636	52.1726809
50.02556728	52.22487001

t_test_paired_samples.csv

ANOVA

- **AN**alysis **O**f **VA**riance의 준말
- 3 개 이상의 Sample 간에 통계적으로 유의미한 차이가 존재하는지 여부를 테스트
- One-way ANOVA (단방향)
- Two-way ANOVA (양방향)

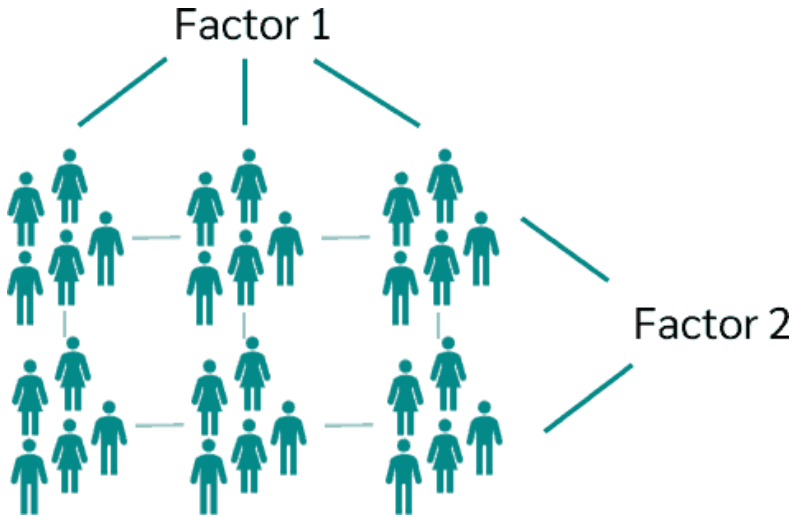


Why not Multiple t-Tests?

- 3개 이상의 group이라 해도 2개씩 짝을 짓는 모든 조합에 대해 t-Test를 하면 되지 않나?
- Statistical test의 error 비율이 약 5%
- 즉 20번의 test 마다 한 번은 정확하지 않다는 것
- 따라서 test 수를 줄이는 것이 error를 막는 방법

One-way vs Two-way ANOVA

- One-way ANOVA
  - Independent variable이 dependent variable의 metric에 영향을 미치는가 만을 check
  - ex) 주거지가 수입에 영향을 미치는가?
- Two-way ANOVA
  - 두 independent variable을 고려하여 분석
  - ex) 주거지와 성별이 수입에 영향을 미치는가?



ANOVA with/without Repeated Measures

- Sample이 independent인지 dependent인지에 따라 without/with repeated measures ANOVA 방식이 선택됨
- Paired Samples t-Test와 유사하나 group이 3개 이상
- ex) 같은 사람을 서로 다른 시간에 인터뷰 하는 것
  - Sample은 dependent
  - ANOVA는 repeated measure를 적용

Post-hoc test

- ANOVA 이후에 수행되는 추가적인 Statistical test
- ANOVA 결과가 유의미하다고 판단될 때 구체적으로 어떤 그룹 간의 차이가 있는지를 확인

- **Tukey's HSD (Honestly Significant Difference) Test**:
  - 가장 많이 사용되는 Post-hoc 검정 중 하나로, 모든 가능한 그룹 간의 쌍별 비교를 수행
  - 각 그룹의 평균 차이를 비교하여 유의미한 차이가 있는지를 판단
- **Bonferroni Correction**:
  - 각 비교의 유의수준을 조정하여 다중 비교로 인한 오류를 통제
  - 매우 보수적인 방법으로, significance level α를 비교의 수로 나누어 각 비교에 적용
  - ex) Post-hoc test에서 두 개씩 짝 지은 비교의 수가 10개라 하면
    - $\alpha = 0.05$ 일 때 $\alpha' = \frac{0.05}{10} = 0.005$ 로 significance level을 낮춤

One-way ANOVA

- Assumptions of One-way ANOVA
  - Level of scale
    - Independent variable은 nominal, Dependent variable은 metric (quantitative)
  - Independence
    - Measurement는 독립적이어야 함. (한 group의 measured value가 다른 group에 영향을 미쳐서는 안됨)
  - Homogeneity
    - 각 group의 variance는 거의 같아야 함 (Levene test로 측정)
  - Normality
    - Group들의 data는 normally distributed 되어 있어야 함
- Assumption을 만족하지 못할 때
  - Kruskal-Wallis test (Nonparametric) 이용
    - Dependent variable이 metric이 아닐 때, Normality를 만족하지 않을 때
  - ANOVA with repeated measures 이용
    - Independence를 만족하지 않을 때

Calculating One-way ANOVA (1/3)

1. Variance

   ○ Total Sum of Squares, SST (총 제곱 합): 전체 데이터의 변동성

$$SST = \sum_{i=1}^{N} (X_i - \overline{X})^2$$

   여기서 N: 전체 샘플 수, X_i:각 data point, $\bar{X}$ 는 total mean

   ○ Sum of Squares Between groups, SSB (그룹 간 제곱 합): 그룹 간 mean의 차이로 인한 변동성

$$SSB = \sum_{j=1}^{k} n_j (\overline{X}_j - \overline{X})^2$$

   여기서 k는 그룹 수, n_j 는 그룹 j의 샘플 수, $\bar{X}_j$ 는 그룹 j의 mean

Calculating One-way ANOVA (2/3)

- Sum of Squares Within groups, SSW (그룹 내 제곱 합): 그룹 내 데이터의 변동성

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$$

여기서 X_{ij}: 그룹 j의 i번째 데이터 포인트

2. Degree of Freedom (df)
   - 그룹 간 자유도: $df_{between} = k - 1$
     - 각 group의 평균을 정한다면, 마지막 한 group의 평균은 전체 평균에 의해 제한되기 때문
   - 그룹 내 자유도: $df_{within} = N - k$
     - 각 group마다 1개씩의 자유도가 제한되기 때문

3. Mean Squares (평균 제곱)
   - MSB (Between: 그룹 간 평균 제곱) = $SSB/df_{between}$: Group 간 차이에 의해 설명되는 변동 크기
   - MSW (Within: 그룹 내 평균 제곱) = SSW/df_{within} : Group 내에서 개별 데이터 간의 변동성

Calculating One-way ANOVA (3/3)

4. F-value 계산
   - F-value는 Group 간 변동성을 Group 내 변동성으로 나눈 값

$$F = \frac{MSB}{MSW}$$

5. p-value 계산 및 significance 판단
   - F-distribution을 사용, p-value 계산
   - significance level (ex. 0.05) 보다 작으면 group간에 유의미한 차이가 있다고 판단

Effect Size for One-way ANOVA

- Eta-square and partial Eta-square:

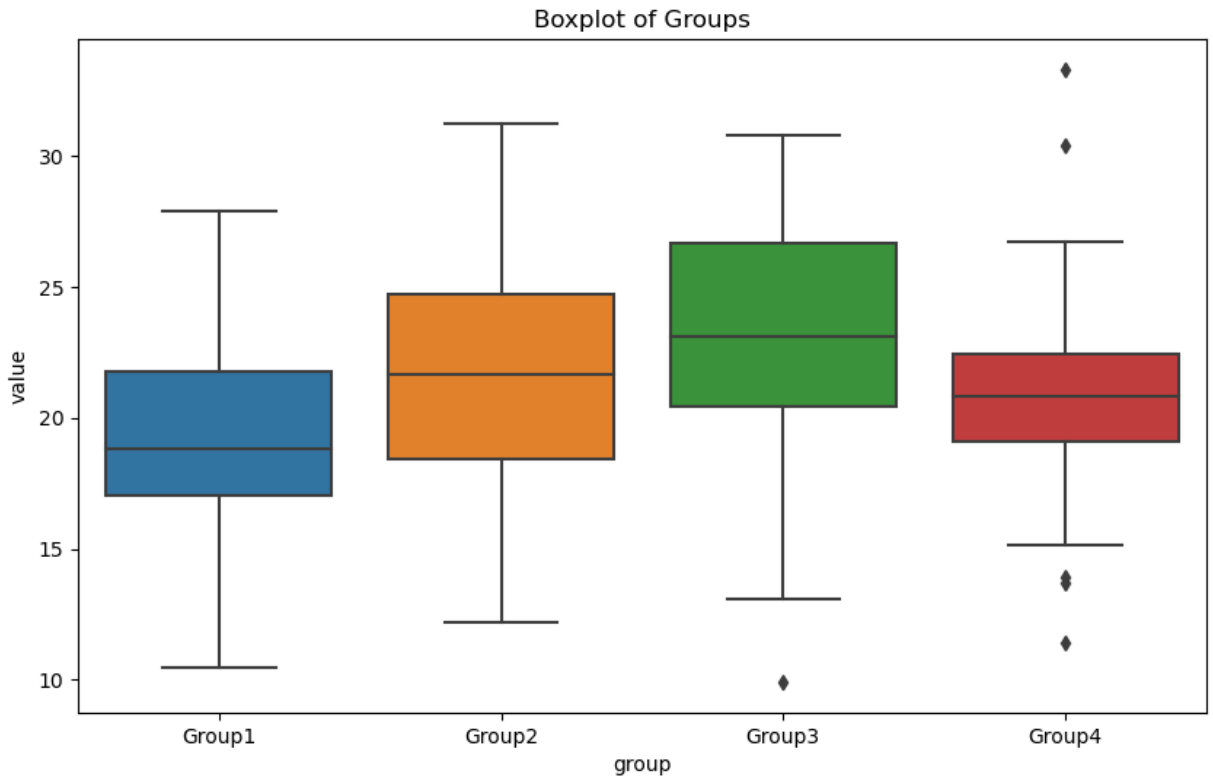
$$\eta^2 = \frac{SS_{btw}}{SS_{tot}} \qquad \eta_p^2 = \frac{SS_{btw}}{SS_{btw} + SS_{wi}}$$

- Cohen's f:

$$f = \sqrt{\frac{\eta_p^2}{1 - \eta_p^2}}$$

CODE (1/3)

- https://github.com/iklee99/StatCode
- 06_ANOVA_oneway.py, anova_oneway_data.csv



Value	Group
22.48357077	Group1
19.30867849	Group1
23.23844269	Group1
...	...
18.99146694	Group2
31.26139092	Group2
21.93251388	Group2
...	...
20.60412881	Group3
22.07170512	Group3
17.46832513	Group3
...	...
21.48538775	Group4
25.84322495	Group4
17.48973453	Group4
...	...

CODE (2/3)

levene_stat = 0.3677123324399591 levene_p = 0.7764376639547123 (> 0.05 이므로 homogeneity pass)

shapiro_p_values = [0.6868111491203308, 0.9129548072814941, 0.365369588136673, 0.1564980000257492] (각 group 당 > 0.05 이므로 normality 모두 pass)

anova_result = F_onewayResult(statistic=3.739905332007017, pvalue=0.013124823365697496)
(p-value < 0.05 이므로 H0 reject, 즉, "group들의 value 평균들은 유의미하게 차이가 있다")

Effect size (eta_squared) = 0.08819164473195337
Effect size (partial_eta_squared) = 0.08819164473195337
Effect size (cohen_f) = 0.31100110871325126

CODE (3/3)

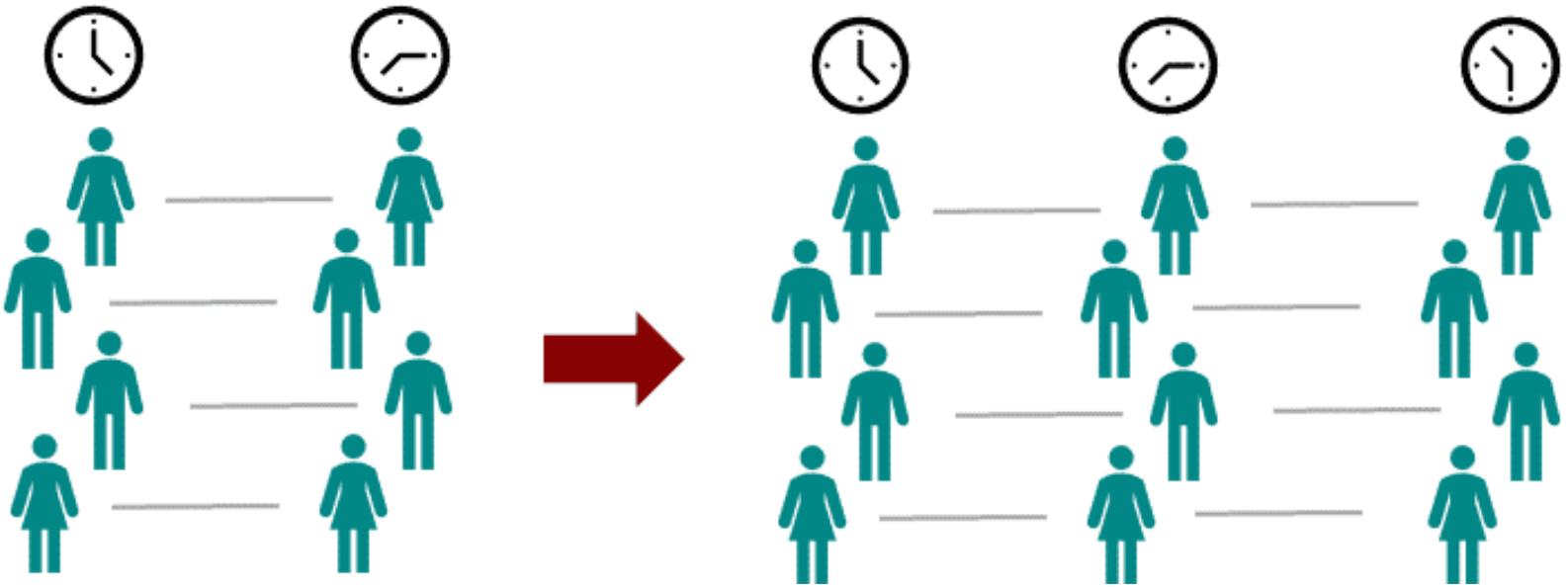
bonferroni_result
Multiple Comparison of Means - Tukey HSD, FWER=0.05
==========
group1 group2 meandiff p-adj lower upper reject

Group1 Group2 2.3349 0.2185 -0.8073 5.4771 False
Group1 Group3 4.0052 0.0065 0.863 7.1473 True
Group1 Group4 1.8395 0.4268 -1.3027 4.9817 False
Group2 Group3 1.6702 0.5096 -1.4719 4.8124 False
Group2 Group4 -0.4954 0.9 -3.6376 2.6467 False
Group3 Group4 -2.1657 0.2803 -5.3079 0.9765 False

Post-hoc tests에서 (Group1, Group3) 의 비교만 value의 mean이 유의미하게 차이가 있다.
나머지 조합들의 차이는 모두 유의미하지 않다.

One-way ANOVA with Repeated Measures

- t-Test for dependent samples (paired samples t-Test) 의 확장

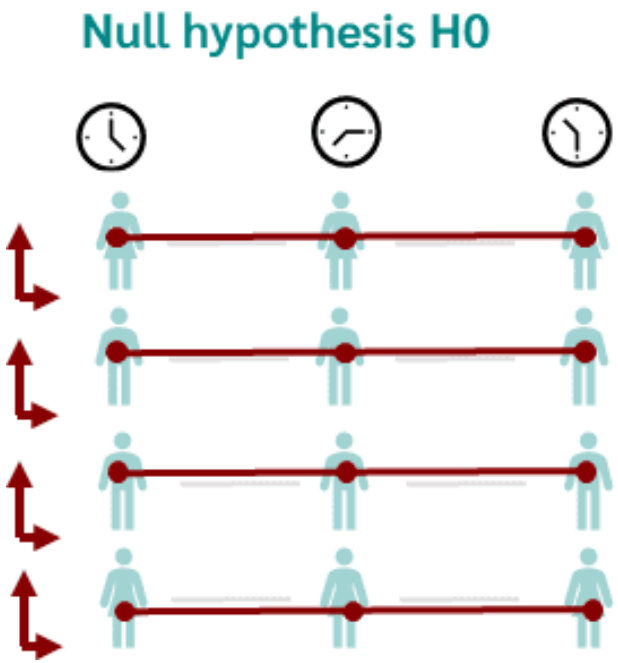


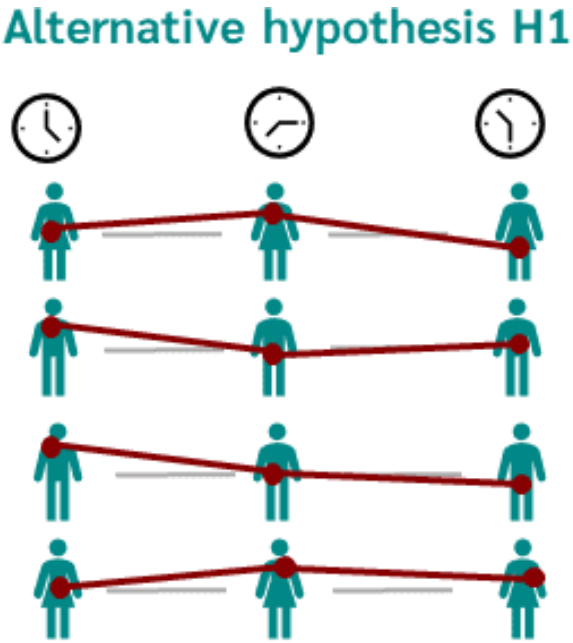
Paired samples t-Test

One-way ANOVA with Repeated Measures

H0 and H1 for One-way ANOVA with R.M.

- Null hypothesis (H0): there are no significant differences between the dependent groups.
- Alternative hypothesis (H1): there is a significant difference between the dependent groups.





Assumptions for ANOVA with R. M.

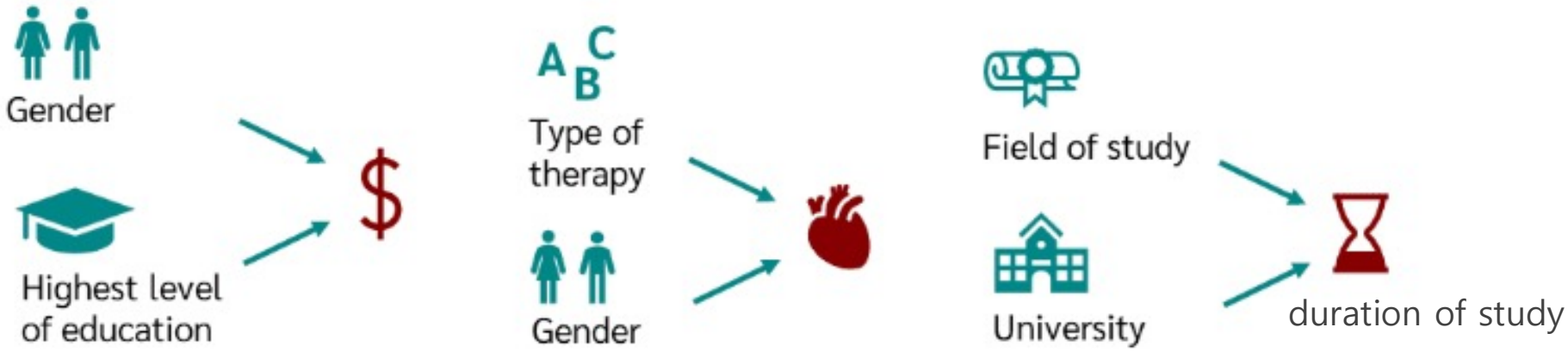
- Dependent Samples
- Normality
- Homogeneity of Variance

CODE: Python Library

```python
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import AnovaRM

# Example data
data = pd.DataFrame({
    'subject': ['S1', 'S1', 'S1', 'S2', 'S2', 'S2', 'S3', 'S3', 'S3'],
    'condition': ['A', 'B', 'C', 'A', 'B', 'C', 'A', 'B', 'C'],
    'value': [10, 15, 14, 9, 12, 13, 11, 14, 15]
})

# Repeated measures ANOVA
aovrm = AnovaRM(data, 'value', 'subject', within=['condition'])
res = aovrm.fit()

print(res)
```

Two-way ANOVA

- Dependent variable의 값을 결정하는데 있어 두 개 이상의 요인 (independent variable) 들이 관여하는 3개 이상의 group을 비교



Example Data

- Two independent variables (factors) are gender (male or female) and study (yes or no)
- Dependent variable is attitude toward retirement planning, where 1 means "not important" and 10 means "very important."

Groups for Factor A: Not-Studied, Studied
Groups for Factor B: Male, Female

Factor A: Study status

Factor B: Gender

	Not Studied	Studied	
Male	6	4	
	4	5	
	7	6	
	9	7	
	3	5	
Mean	5.8	5.4	5.6
Female	8	3	
	3	5	
	5	9	
	8	2	
	6	3	
Mean	6	4.4	5.2
	5.9	4.9	**5.4**

H0 for Two-way ANOVA

Null hypotheses
There is no significant difference between the groups of the independent variable Studied in relation to the dependent variable Attitude.
There is no significant difference between the groups of the independent variable Gender in relation to the dependent variable Attitude.
There is no significant interaction between the two variables Studied and Gender in relation to the dependent variable Attitude.

Assumptions

- Independent variable은 categorical, Dependent variable은 metric (quantitative)
- Independence
  - Measurements는 독립적으로 이루어져야 함
- Homogeneity
  - 각 그룹들의 variance들은 거의 유사해야 함 (Levene's test)
- Normality
  - 데이터는 normal distribution을 이루어야 함

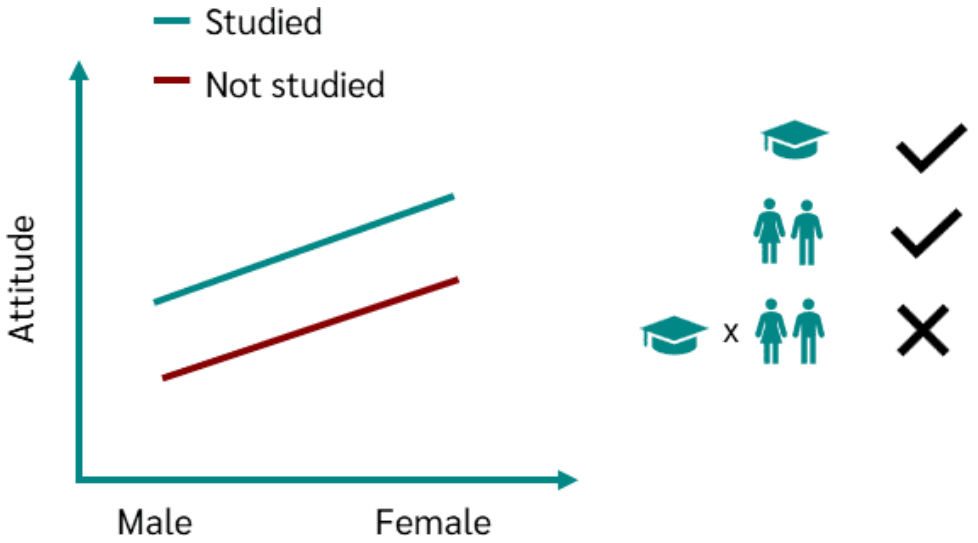
Result of Two-way ANOVA

	Type III Sum of Squares	df	Mean Squares	F	p
Corrected Model	7.6	3	2.53	0.53	.671
Intercept	583.2	1	583.2	120.87	<.001
Studied	5	1	5	1.04	.324
Gender	0.8	1	0.8	0.17	.689
Studied x Gender	1.8	1	1.8	0.37	.55
Error	77.2	16	4.83		
Total	668	20			
Corrected total variation	84.8	19			

Interpreting Two-way ANOVA

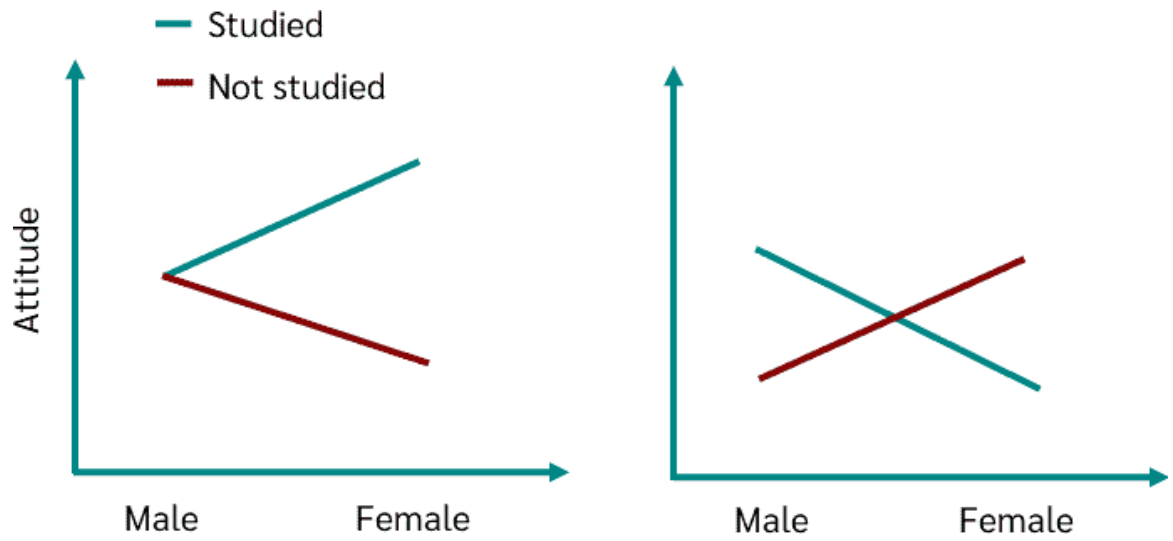
- 3가지 p-value가 모두 > 0.05 이므로 3개의 N0를 모두 reject 할 수 없다.
- "Studied" 여부는 "은퇴계획에 대한 자세" 에 유의미한 영향을 주지 않는다.
- "성별" 차이는 "은퇴계획에 대한 자세"에 유의미한 영향을 주지 않는다.
- "은퇴계획에 대한 자세"에 대한 "Studied" 와 "Gender" 간의 유의미한 interaction 은 없다.

Interaction Effect (1/3)



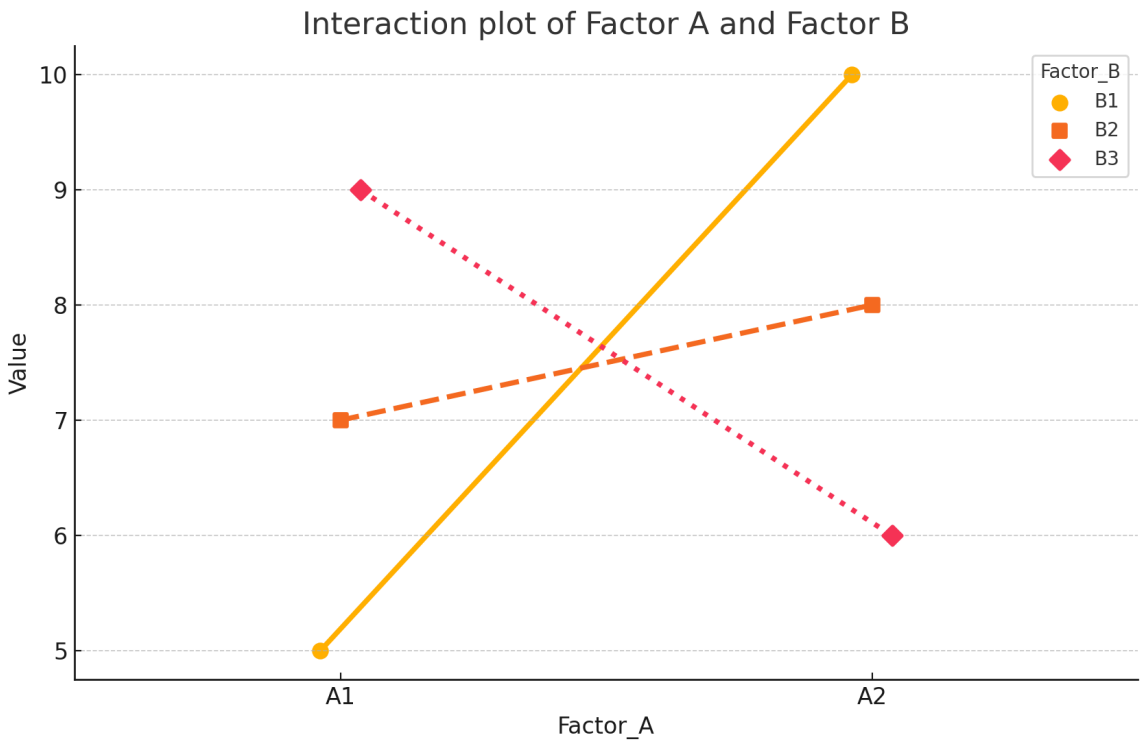
- Line의 양 끝점은 두 factor들의 group mean value들, e.g. male and not studied
- 이 graph의 해석:
  - Female의 attitude가 Male 보다 높다
  - Studied의 attitude가 Not studied 보다 높다
  - Gender와 Studied의 두 요인의 interaction은 거의 없다. (유의미하지 않다)

Interaction Effect (2/3)



- 두 경우 모두 두 요인 Gender와 Studied 여부가 유의미하게 interaction하고 있다.
- 즉, Male인지 Female인지가 정해지면, 다른 요인인 Studied 여부에 따라 Attitude가 완전히 다른 방향으로 갈 수 있다.

Interaction Effect (3/3)



- Main effect
  - Factor A: A1과 A2의 값이 다르면 Factor A에 주 효과가 있다
  - Factor B: 각 line color (style) 의 높이가 다르면 B에 주 효과가 있다
- Interaction of Factor A & B
  - Line들이 교차하면 A와 B가 interaction 가능성이 매우 높다 (왼쪽 graph에서 노랑색 라인 (Factor B1) 의 경우, A1에서 value가 낮고 A2에서 높은데, B3 line의 경우 반대로 A1에서 가장 높고 A2에서 가장 낮다. 즉, Factor B가 무엇이냐가 Factor A의 효과에 영향을 많이 미친다)
  - Line들이 평행하지 않으면 interaction 효과가 어느정도 있다.
  - Line들이 평행하면 interaction 효과가 별로 없다.

CODE (1/5)

- https://github.com/iklee99/StatCode
- 07_ANOVA_twoway.py, datasets/anova_twoway_data.csv

CODE (2/5)

passed

```
Levene's test for homogeneity: stat=0.2452042440967302, p-value=0.6223432008475429
Shapiro-Wilk test for normality: stat=0.9880377650260925, p-value=0.8232324719429016
Two-way ANOVA results:
                            sum_sq    df          F    PR(>F)
C(Factor_A)               0.151433   1.0   0.092280  0.762466
C(Factor_B)               9.240388   2.0   2.815435  0.068693
C(Factor_A):C(Factor_B)  11.566196   2.0   3.524081  0.036431
Residual                 88.615244  54.0        NaN       NaN
```

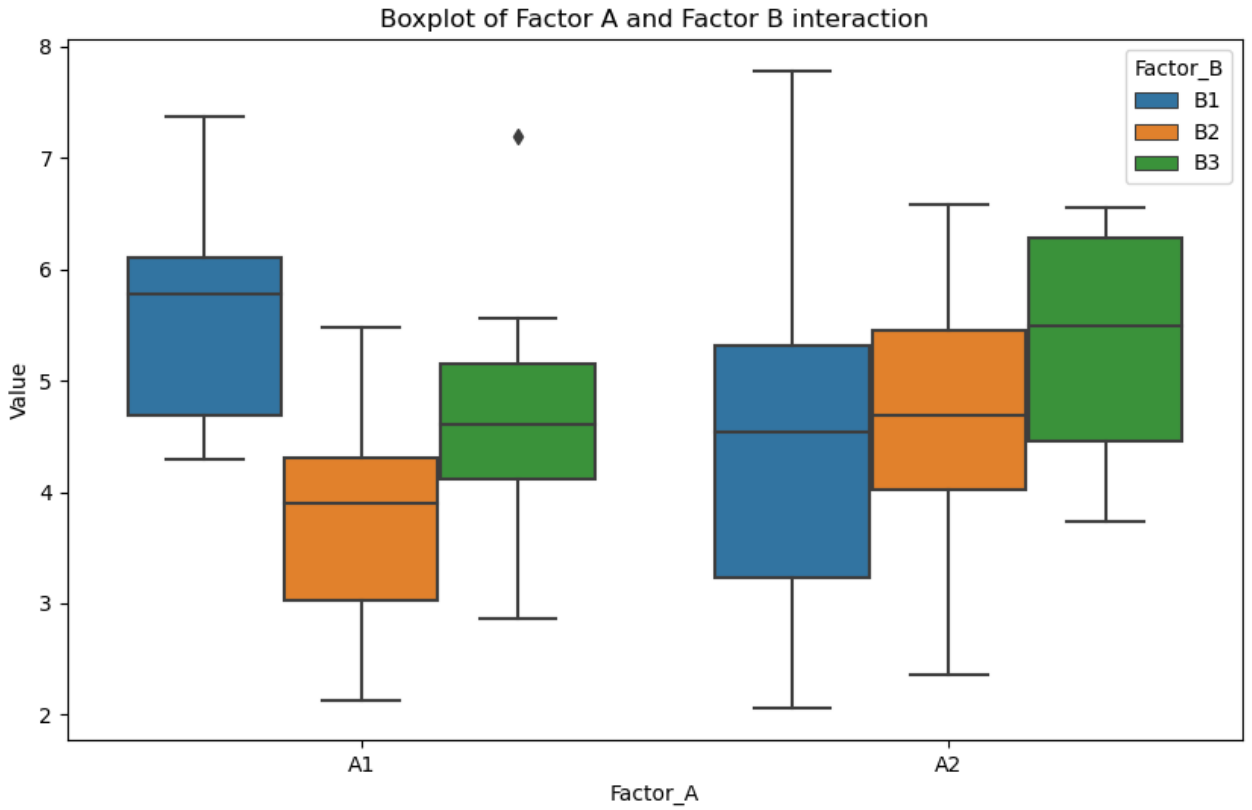
Factor A의 Group들이 가지는 Value는 유의미한 차이가 없다.
Factor B의 Group들이 가지는 Value는 유의미한 차이가 없다.
Factor A와 B는 Value값에 대하여 유의미한 interaction을 가진다.

CODE (3/5)

```
Bonferroni post-hoc test results:
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
  A1B1   A1B2  -1.8581 0.0236 -3.5507 -0.1655   True
  A1B1   A1B3  -1.0049 0.5025 -2.6974  0.6877  False
  A1B1   A2B1  -1.1373 0.3649 -2.8298  0.5553  False
  A1B1   A2B2  -1.0513 0.4543 -2.7439  0.6413  False
  A1B1   A2B3  -0.3729    0.9 -2.0655  1.3197  False
  A1B2   A1B3   0.8532 0.6514 -0.8394  2.5458  False
  A1B2   A2B1   0.7208 0.7815 -0.9718  2.4134  False
  A1B2   A2B2   0.8068 0.6971 -0.8858  2.4993  False
  A1B2   A2B3   1.4851  0.117 -0.2074  3.1777  False
  A1B3   A2B1  -0.1324    0.9  -1.825  1.5602  False
  A1B3   A2B2  -0.0465    0.9 -1.7391  1.6461  False
  A1B3   A2B3   0.6319 0.8688 -1.0607  2.3245  False
  A2B1   A2B2   0.0859    0.9 -1.6067  1.7785  False
  A2B1   A2B3   0.7643 0.7388 -0.9283  2.4569  False
  A2B2   A2B3   0.6784 0.8232 -1.0142   2.371  False
-----------------------------------------------------
```

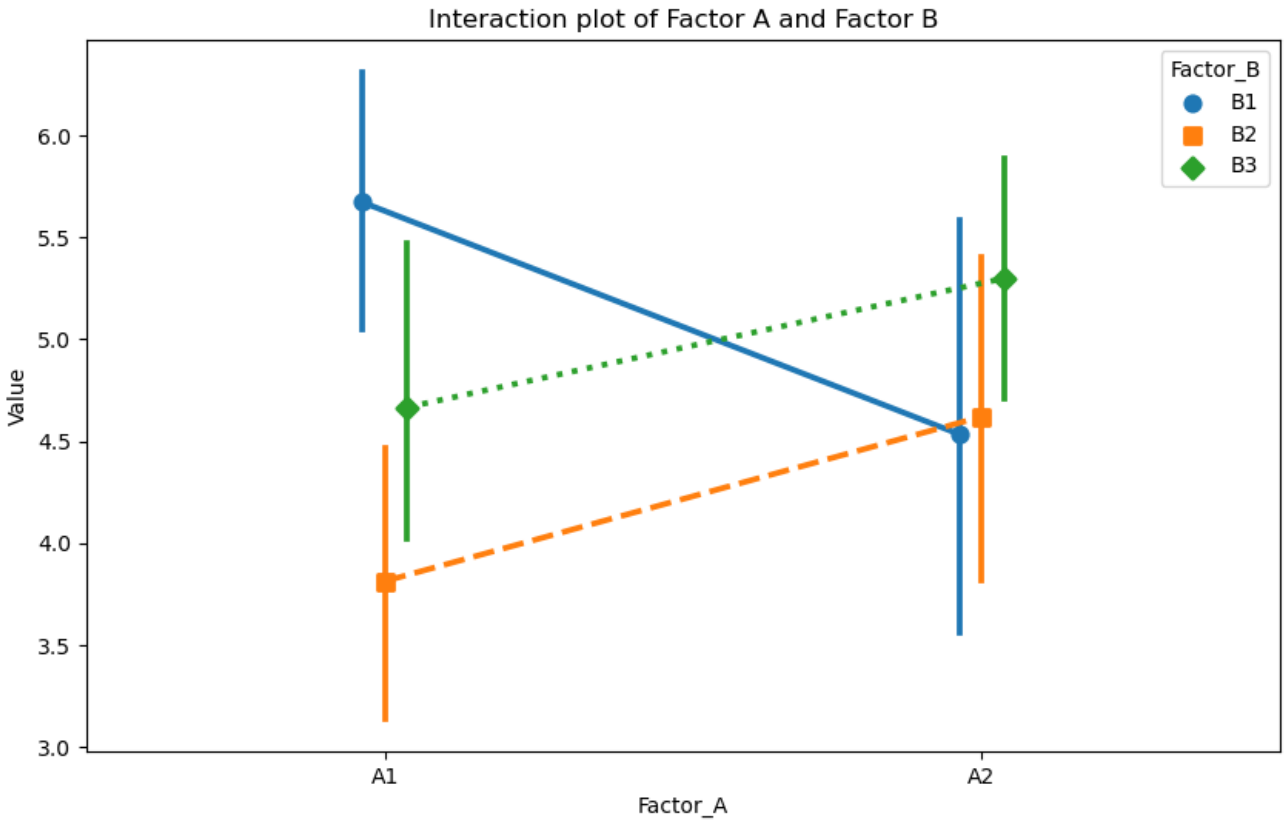
CODE (4/5)

- Boxplot for the factors



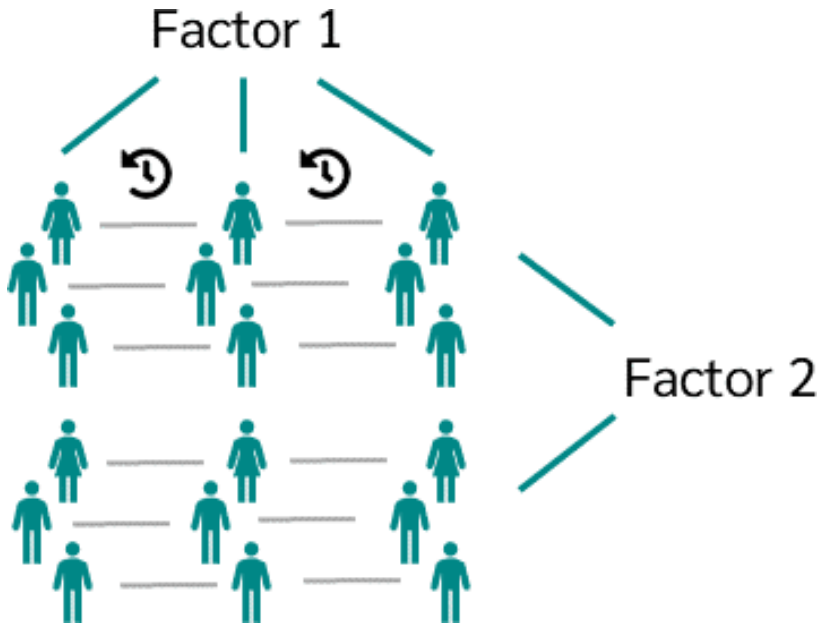
CODE (5/5)

- Interaction plot
- Factor A와 B 사이에는 어느정도 interaction 효과가 있다
- Cross 하거나 평행하지 않음
- 그러나 Factor B2, B3 과 Factor A 간에는 interaction 효과가 거의 없다. (평행하다)



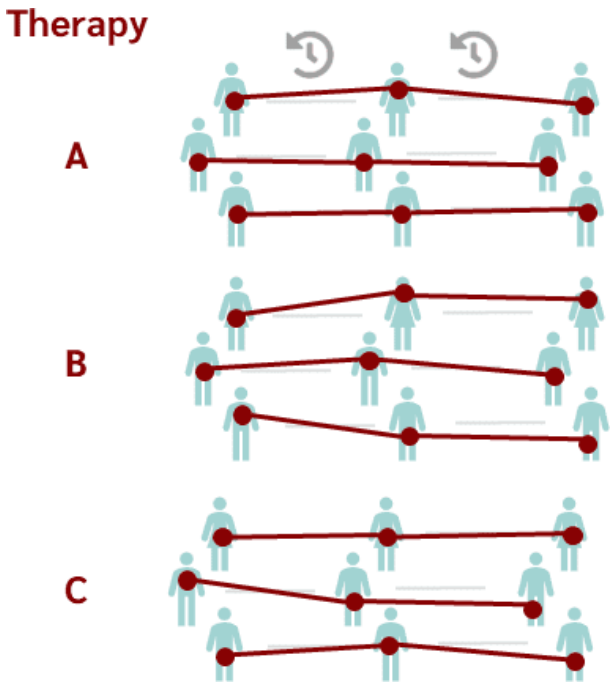
Two-way ANOVA with measurement repetition

- Two-way ANOVA 에 비교하여, 두 개의 factor 중 하나가 서로 다른 시간에 반복 측정됨
- 즉, 하나의 factor는 dependent sample 이라는 뜻



Example of Two-way ANOVA with M. R.

- Independent variable:
  - Factor 1: 같은 Sample에 대해 측정한 시간 (Therapy 1일 후, Therapy 1주일 후, Therapy 2주일 후)
  - Factor 2: Therapy 의 종류 (A, B, C)
- Dependent variable:
  - 혈압 (Blood Pressure)



H0 for Two-way ANOVA with M. R.

- The mean values of the different measurement times do not differ (There are no significant differences between the "groups" of the first factor).
- The mean values of the different groups of the second factor do not differ.
- One factor has no influence on the effect of the other factor

CODE: Python Library

```python
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.anova import AnovaRM
import matplotlib.pyplot as plt
import seaborn as sns

# 데이터 로드
data = pd.read_csv('path_to_your_data.csv')  # 실제 데이터 파일 경로로 변경

# 데이터 구조 확인
print(data.head())

# 반복 측정을 포함한 이원 분산분석 수행
aovrm = AnovaRM(data, 'Value', 'Subject', within=['Factor_A', 'Factor_B'])
res = aovrm.fit()

# 결과 출력
print(res)

# 상호작용 플롯 생성
plt.figure(figsize=(10, 6))
sns.pointplot(x='Factor_A', y='Value', hue='Factor_B', data=data, dodge=True
plt.title('Interaction plot of Factor A and Factor B')
plt.show()
```

MANOVA (Multivariate Analysis of Variance)

- 종속변수의 개수
  - MANOVA: 종속변수가 2개 이상인 경우 사용
  - Two-way ANOVA: 종속변수가 1개인 경우 사용
- 분석 목적
  - MANOVA: 여러 종속변수 간의 상관관계를 고려하여 독립변수의 효과를 종합적으로 분석
  - Two-way ANOVA: 개별 종속변수에 대한 독립변수의 효과를 각각 분석
- 통계적 검정력
  - MANOVA: 여러 종속변수를 동시에 고려하므로 검정력이 높음
  - Two-way ANOVA: 개별 종속변수를 분석하므로 상대적으로 검정력이 낮음

Correlation Test

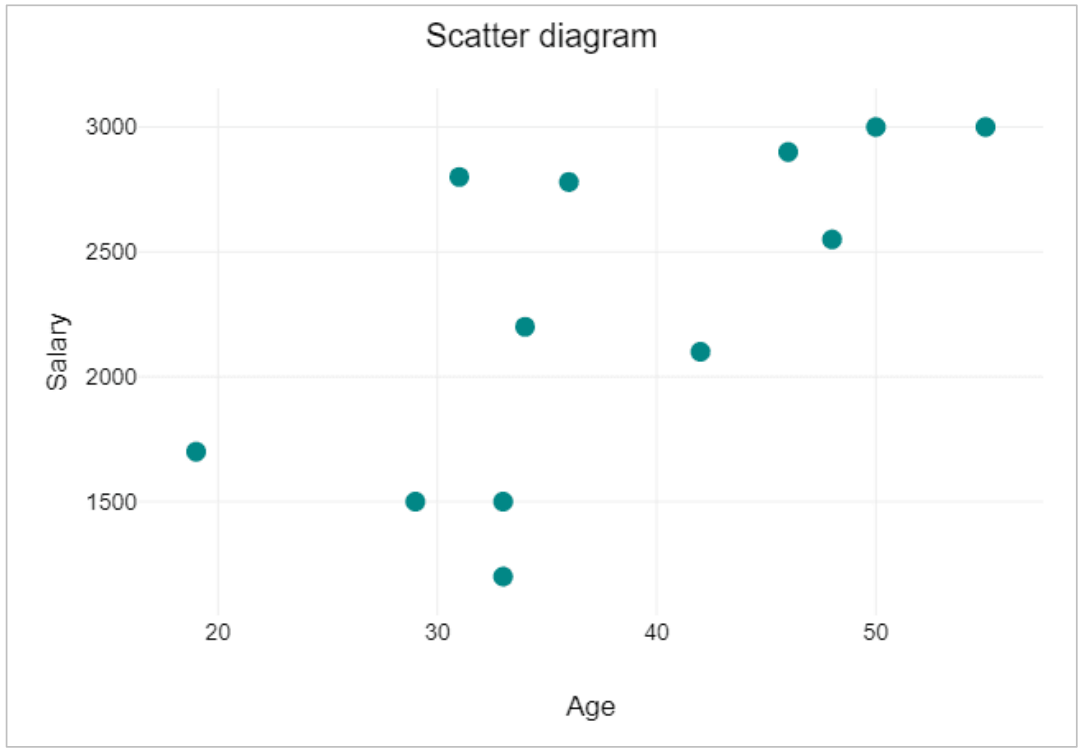
- Variable 간의 Relationship에 대한 정보를 제공하는 통계 기법
- Correlation Coefficient r: from -1 to +1

$\|r\|$	Strength of correlation
0.0 < 0.1	no correlation
0.1 < 0.3	little correlation
0.3 < 0.5	medium correlation
0.5 < 0.7	high correlation
0.7 < 1	very high correlation

- Correlation Analysis의 사용
  - Correlation이 확인되면 Linear Regression을 통해 미래의 상태를 예측 가능
  - x 와 y가 correlation이 높으면 x, y가 혹은 y, x가 원인-결과 관계 일 수 있음

Scatter Plot

- Correlation을 가장 잘 표현해 주는 Scatter Plot



H0, H1 for Correlation Test

- H0: there is no correlation between the variables under consideration
- H1: there is a correlation between the variables under cosinderation

- t-value for testing the hypothesis:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

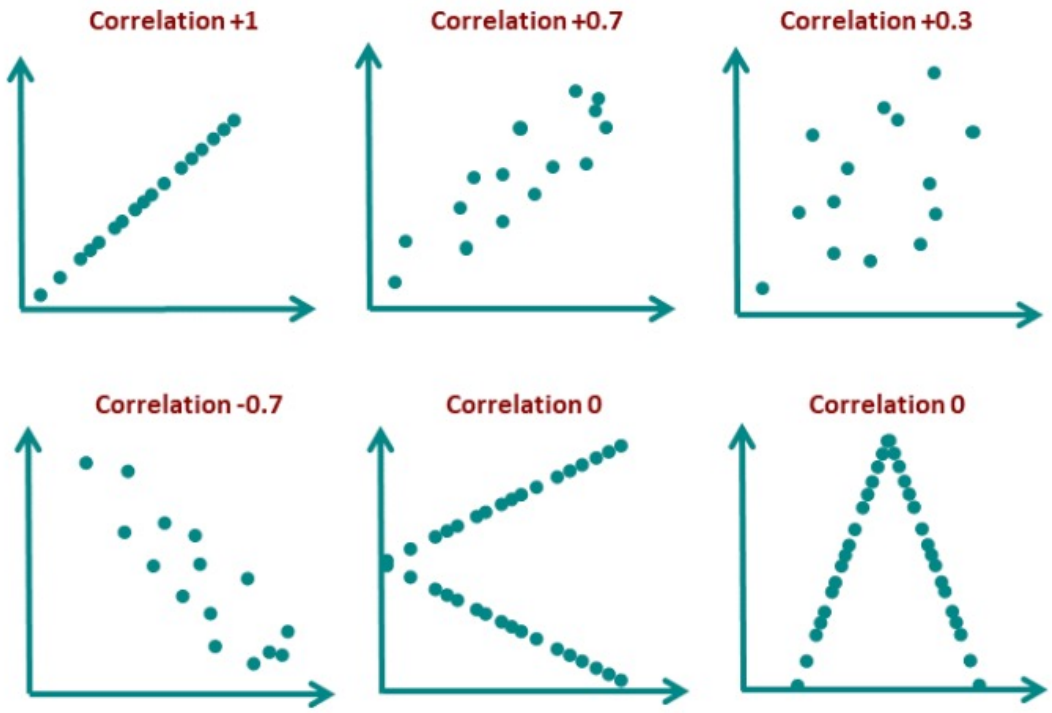
Directional and Non-directional Hypothesis

- Non-directional Hypothesis
  - Correlation이 없다 또는 있다는 hypothesis
  - r이 positive 또는 negative라는 것에 관심이 없음

- Directional Hypothesis
  - H0: Correlation이 없거나, 있어도 r이 negative (resp. positive) 이다.
  - H1: Correlation이 있으며, r이 positive (resp. negative) 이다

Pearson Correlation Analysis

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



$\lvert r \rvert$	Strength of correlation
0.0 < 0,1	no correlation
0.1 < 0,3	little correlation
0.3 < 0,5	medium correlation
0.5 < 0,7	high correlation
0.7 < 1	very high correlation

Pearson Correlation assumptions

- Normality: Variables must be normally distributed
- Normality를 만족하지 않으면 Spearman Correlation을 실행

Spearman Rank Correlation

- Ordinal 또는 Quantitative 인 두 variable 간의 correlation을 분석
- Nonparametric test

CODE (1/2)

- https://github.com/iklee99/StatCode
- 08_correlationTest.py, correlation_test_data.csv

- Result:

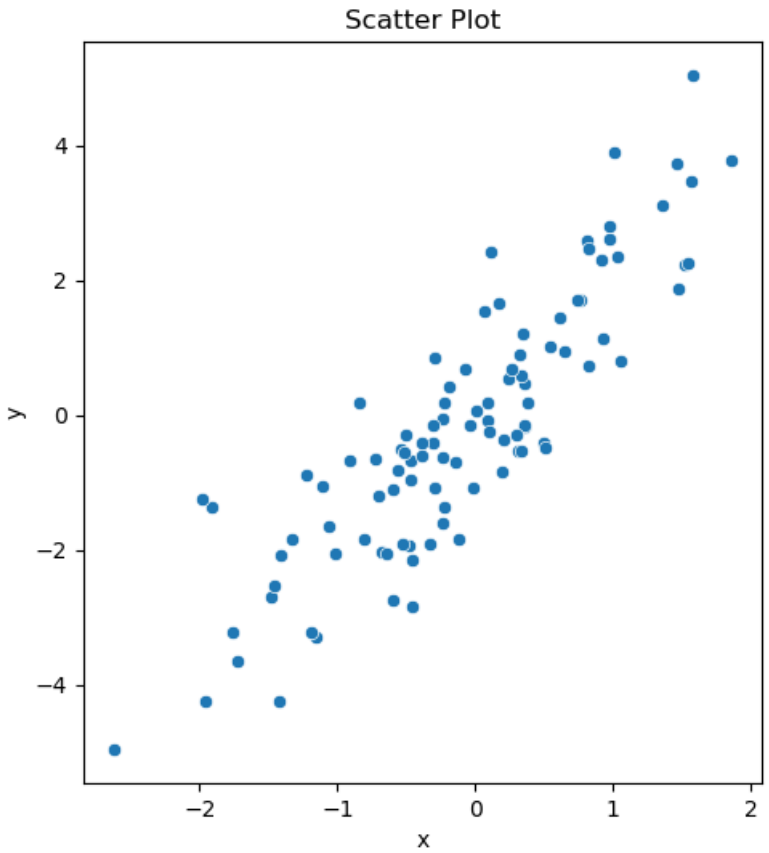
Shapiro Test for x: stat = 0.9899, p = 0.6552 (Normality passed)

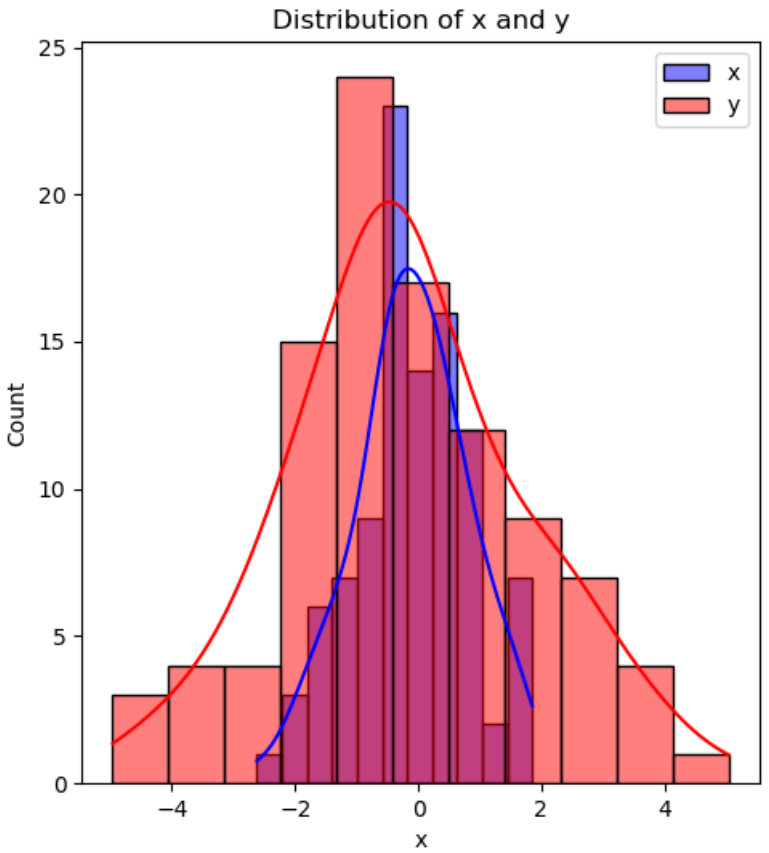
Shapiro Test for y: stat = 0.9906, p = 0.7144 (Normality passed)

Pearson Correlation Test: stat = 0.8724, p = 0.0000

(r = 0.8724, very strong correlation, p < 0.05, so H0 rejected)

CODE (2/2)


Scatter Plot
y
x
4
2
0
−2
−4
−2
−1
0
1
2


Distribution of x and y
Count
x
x
y
25
20
15
10
5
0
−4
−2
0
2
4