통계적 가설 검정이란?

기술 통계와 추론 통계

- **기술 통계(Descriptive statistics)**
  - 데이터를 요약해 설명하는 통계 기법
  - ex) 사람들이 받는 월급을 집계해 전체 월급 평균 구하기
- **추론 통계(Inferential statistics)**
  - 단순히 숫자를 요약하는 것을 넘어 어떤 값이 발생할 확률을 계산하는 통계 기법
  - ex) 수집된 데이터에서 성별에 따라 월급에 차이가 있는 것으로 나타났을 때, 이런 차이가 우연히 발생할 확률을 계산

- **추론 통계(Inferential statistics)**
  - 이런 차이가 우연히 나타날 확률이 작다
    - -> 성별에 따른 월급 차이가 통계적으로 유의하다(statistically significant)고 결론
  - 이런 차이가 우연히 나타날 확률이 크다
    - -> 성별에 따른 월급 차이가 통계적으로 유의하지 않다(not statistically significant)고 결론
  - 기술 통계 분석에서 집단 간 차이가 있는 것으로 나타났더라도 이는 우연에 의한 차이일 수 있음
    - 데이터를 이용해 신뢰할 수 있는 결론을 내리려면 유의확률을 계산하는 통계적 가설 검정 절차를 거쳐야 함

통계적 가설 검정

- **통계적 가설 검정(Statistical hypothesis test)**
  - 유의확률을 이용해 가설을 검정하는 방법
- **유의확률(Significance probability, p-value)**
  - 실제로는 집단 간 차이가 없는데 우연히 차이가 있는 데이터가 추출될 확률
  - 분석 결과 유의확률이 크게 나타났다면
    - '집단 간 차이가 통계적으로 유의하지 않다'고 해석
    - 실제로 차이가 없더라도 우연에 의해 이 정도의 차이가 관찰될 가능성이 크다는 의미
  - 분석 결과 유의확률이 작게 나타났다면
    - '집단 간 차이가 통계적으로 유의하다'고 해석
    - 실제로 차이가 없는데 우연히 이 정도의 차이가 관찰될 가 능성이 작다, 우연이라고 보기 힘들다는 의미

t 검정(t-test)

- 두 집단의 평균에 통계적으로 유의한 차이가 있는지 알아볼 때 사용하는 통계 분석 기법

T-Test(T 검정) 란

T-Test (Two-Sample Test / Student's T Test) 란 두 집단의 평균을 비교하는 통계적 검정 방법이다. 단순히 차이의 존재 여부를 떠나 두 집단의 비교가 통계적으로 유의미한가를 검정한다. 다른 말로는, 이 두 모집단의 차이가 우연에 의해서인지 아닌지를 검정한다.

Example 1)
감기에 걸렸을 때 약을 먹었을 때 치유되는 기간과 먹지 않았을 때 자연적으로 치유되는 기간을 비교할 수 있다.
두가지 case 의 차이가 유의미한지 검정하기 위해서는 여러 명의 사람에게 평균적으로 비슷한 기간이 걸리는지 테스트를 할 필요가 있다. (약을 먹었을 때 평균 3 일, 자연치유가 평균 5 일이 걸렸을 시 이 결과가 repeatable 한가? 우연에 의해서 혹은 다른 요인에 의해 차이가 나는 것은 아닌지)

Example 2)
타이타닉 data 를 사용하여 비교할 때, 1 등석/2 등석/3 등석의 평균 생존률을 T-Test 를 사용하여 생존률과 좌석 class 의 차이가 연관이 있는지 검정할 수 있다

T-Score

ratio between the difference between two groups and the difference within the groups
- t score 이 클수록, 두개의 그룹은 차이가 나고 t score 이 작을수록 두개의 그룹이 비슷하다고 볼 수 있다.
- T-Score 이 클수록 테스트 결과가 repeatable 하다

T-Value and P-Value

모든 T-Value 는 P-Value (probability) 를 가지고 있다. P-Value 는 데이터가 우연에 의해 일어났을 확률이다.
P-Value 가 유의수준(통상 5%) 과 같거나 적다면 두 모집단이 유의미한 차이가 있다고 생각한다. (They indicate your data did not occur by chance)
0 에 가까울수록 좋은 p-value 이다.

T-Test in Python

파이썬 라이브러리를 사용하여 간단하게 T-Test 를 할 수 있다.

SciPy stats 의 ttest_ind 명령어를 사용한다

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

파이썬 예)

```
import numpy as np
from scipy import stats
# to get consistent result
np.random.seed(1)

# before treatment : mean 60, standard deviation 5
beforeWeights = [60 + np.random.normal(0, 5) for _ in range(20)]

# after treatment : mean 0.99-fold decrease, standard deviation 0.02
afterWeights = [w * np.random.normal(0.99, 0.02) for w in beforeWeights]

# perform paired t-test
tTestResult = stats.ttest_rel(beforeWeights, afterWeights)

print("The T-statistic is %.3f and the p-value is %.3f" % tTestResult)
```

실행결과:

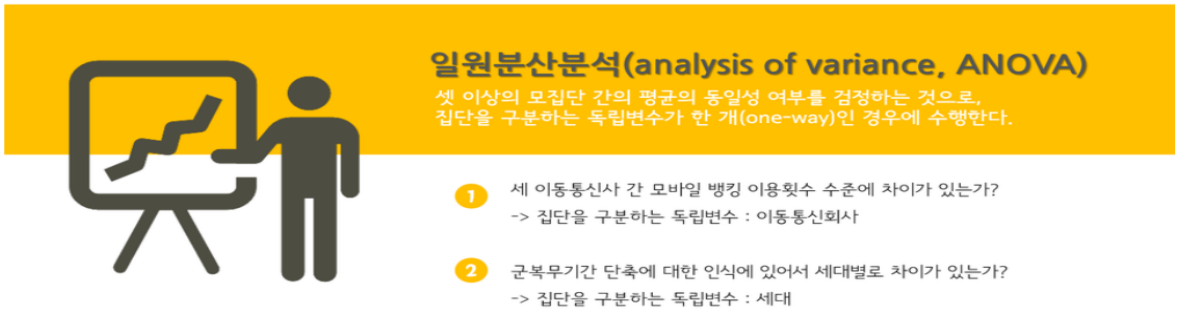
The T-statistic is 2.915 and the p-value is 0.009

 # p-value < 0.05 에서 귀무 가설을 기각한다.
 # 즉, 다이어트 약 복용 전/후에 체중 차이는 유의미하다고 할 수 있다

독립표본 t test 의 두 정규분포의 분산값이 같은 경우와 같지 않은 경우에 사용하는 검정 통계량이 다르기 때문에
equal_var 을 사용하여 지정해줘야 한다. (default 값은 True)

참고 : http://www.statisticshowto.com/probability-and-statistics/t-test/

ANOVA



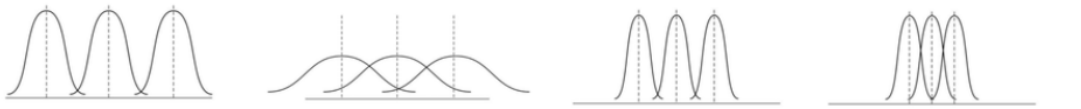
일원분산분석이란, 집단을 구분하는 **독립변수(요인)**이
한 개일때 수행합니다.
요인이 2개라면 뒤에서 다룰 이원분산분석이 되겠죠?!

분산분석에는 3가지의 전제조건의 충족되어야 합니다
1. 독립성 : 무작위 표본으로부터 선정
2. 정규성 : 모집단은 정규분포를 함
3. 분산동일성 : 모집단 분산은 모두 동일함

ONE-WAY ANOVA

집단 간 평균의 차이를 검정하는데 '분산'을 분석하는 이유는?

집단의 평균이 다르다는 말은, 각 집단의 평균이 떨어져 있어 즉 분산이 크다는 것을 의미한다.
따라서 집단평균 분산이 클수록 집단 간 평균이 서로 다를 가능성도 크다.
하지만 집단평균의 분산이 크더라도 각 집단 내의 분산도 크다면 서로 겹치는 영역이 커서 분포가 명확히 구분이 되지 않아 평균이 다르다고 주장하기 힘들다(그래프2)



집단 간 분산이 큰경우
▶ 집단의 평균이 서로 다름

집단간 분산이 크면서 집단 내 분산이 큰경우
▶ 각 집단의 분포가 명확히 구분되지 않음 집단 간 평균이 다르다고 주장하기 어려움

집단 내 분산이 작은 경우
▶ 집단 간 분포가 명확히 구분됨

집단 내 분산이 작으면서 집단 간 분산이 작은 경우
▶ 각 집단의 분포가 명확히 구분되지 않음 집단 간 평균이 다르다고 주장하기 어려움

그런데, 평균의 동일성 여부를 검정하는데 왜 분산을 분석할까요?
이유는 집단의 평균이 다르다는 것이
집단평균의 분산이 크다는 것을 의미하기 때문입니다.
그래프에서 겹치는 영역이 많을수록
집단 간 평균이 다르다고 보기힘들어지게 됩니다.

F-VALUE

F 통계량(F-value)

집단 간 분산이 클수록, 집단 내 분산은 작을수록 집단평균이 다를 가능성이 증가하여 두 종류의 분산이 갖는 값의 상대적 크기에 의해 집단 간 평균의 동일성 여부가 결정된다고 할 수 있다. **집단 간 분산**과 **집단 내 분산의 비**를 F 통계량이라고 한다.

- 집단개수-1, 표본크기-집단개수 를 자유도로 갖는 F분포 따름
- 항상 양수 값, 양(+)의 영역에서 분포, 두 개의 자유도에 의해 분포모양 결정됨

$$F = \frac{\text{집단 간 분산}}{\text{집단 내 분산}}$$

$$\text{집단 간 분산} = \frac{\text{집단 간 제곱합}}{\text{자유도}}$$

$$\text{집단 간 분산} = \frac{\sum_g \left[(\bar{X}_g - \bar{X})^2 \times n_g \right]}{g-1}$$

여기서, g는 집단의 수, $\bar{X}_g$는 g집단의 표본평균, $\bar{X}$는 전체표본의 평균, n_g는 g집단의 표본크기

$$\text{집단 내 분산} = \frac{\text{집단 내 제곱합}}{\text{자유도}}$$

$$\text{집단 내 분산} = \frac{\sum_g \sum_i (X_{ig} - \bar{X}_g)^2}{\sum_g (n_g - 1)} = \frac{\sum_g [s_g^2 \times (n_g - 1)]}{\sum_g (n_g - 1)} = \frac{\sum_g [s_g^2 \times (n_g - 1)]}{n - g}$$

여기서, g는 집단의 수, s_g는 g집단의 표준편차, n_g는 g집단의 표본크기
X_{ig}는 g집단의 i번째 관측값, $\bar{X}_g$는 g집단의 표본평균, n은 전체 표본크기

집단 간 분산과 집단 내 분산의 비를 나타내는 것이
F통계량, 즉 F값입니다.
집단 간 분산(분자)이 클수록, 그리고 집단 내 분산(분모)이 작을수록
F통계량이 커지며, 이 F통계량이 커질수록
집단간 평균 차가 존재한다고 할 수 있습니다.

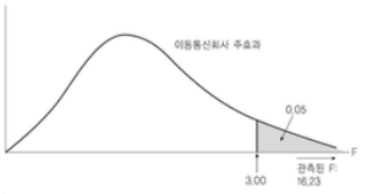
TWO-WAY ANOVA

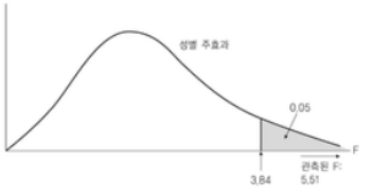
이원분산분석

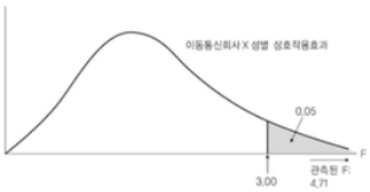
집단을 구분하는 독립변수(요인)이 두 개인 경우
이원분산분석(two-way ANOVA)를 이용하여 여러 모
집단 간의 평균의 동일성 여부를 검정한다.

- 이동통신사 간 및 남녀 간의 모바일 뱅킹 이용횟수에 차이가 존재하는가?
  -> 집단을 구분하는 독립변수 : 이동통신회사, 성별

- 승하차시스템의 유형 및 놀이시설의 유형은 대기시간에 영향을 치는가?
  -> 집단을 구분하는 독립변수 : 승하차시스템의 유형, 놀이시설의 유형







ANALYSIS OF VARIANCE

일원배치분산분석
One-way ANOVA

셋 이상의 모집단 간의 평균의 동일성 여부를 검정하는 것으로, 집단을 구분하는 독립변수가 한 개(one-way)인 경우이다.

F통계량
F statistic, F-value, F값

집단 간 분산과 집단 내 분산의 비 F값이 클수록 집단 간 평균차가 존재한다고 할 수 있다.

$$\frac{\text{집단 간 분산}}{\text{집단 내 분산}}$$

이원배치분산분석
Two-way ANOVA

집단을 구분하는 독립변수가 두 개 (two-way)인 경우이다.

F-검정은 언제 사용할까?

F-검정은 모집단의 분산의 차이가 있는가를 검정할 때 사용한다.
(집단의 평균의 차이가 존재하는 가가 아니라 분산의 차이가 있는가를 검정한다.)

F-검정의 특징

F-검정 값은 항상 1 보다 같거나 크다. (두 표본집단의 분산 값을 나눈 것이므로, 큰 것이 분모, 작은 것이 분자)
F 값이 클수록 두 집단간의 분산의 차이가 존재하는 것을 의미한다.

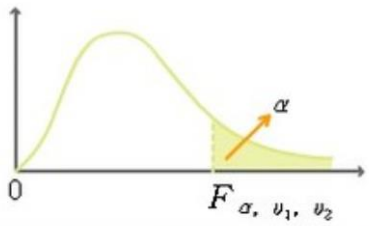
예 1.

예를들어, 어느 중학교에서 1 학년 학생들의 성적의 차이(분산)이 2 학년이 되면 더 커질 것이라고 예상된다. 실제로 그런가 검정해보자. 1 학년에서 7 명을 뽑고, 2 학년에서 9 명을 뽑아서 각각의 성적의 분산을 조사해 봤더니, 1 학년의 분산은 9.0 이었고, 2 학년의 분산은 19.8 이었다. 두 모집단의 분산은 같다고 볼 수 있을까? 알파=0.05 에서 검정해보자.
F(8,6) = 4.15 이다. (자유도는 개체 크기에서 1 씩 뺀 값이며 2 개가 사용된다. F 분포표에서 찾아보자.)
F = 19.8 / 9 = 2.2 이다. 2.2 < 4.15 이므로 F=2.2 는 기각역 안에 있으며, 귀무가설을 기각할 수 없다.
즉, 2 학년학생의 성적 차이가 1 학년 학생의 성적차이보다 크다고 할 수 없다.

F-분포표

F 검정에 필요한 F 분포표를
F 검정표는 두 개의 자유도 값을 사용한다. (행, 열에 두 표본의 자유도가 사용된다.)

<부표-5> F 분포표(α=0.05)



υ_2 \ υ_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9	243.9	245.9	248	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.4	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.51
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.59	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.10	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.56	4.53	4.50	4.46	4.43	4.40
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.71	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.14	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.49	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.71	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.9	2.85	2.79	2.72	2.63	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.42	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13

15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.18	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.15	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.23	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.25	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.13	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.64	1.70
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.68	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.88	1.75	1.68	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.34	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

anova_ex01.py

```
# 분산분석에는 여러 종류가 있지만, 여기서는 일원분산분석(One-way ANOVA)과 이원분산분석(Two-way ANOVA)를 알아보겠다.
# 1) 일원분산분석(One-way ANOVA)
#  종속변인은 1 개이며, 독립변인의 집단도 1 개인 경우다.
#  한 가지 변수의 변화가 결과 변수에 미치는 영향을 보기 위해 사용된다.
# 파이썬에서 One-way ANOVA 분석은 scipy.stats 이나 statsmodel 라이브러리를 이용해서 할 수 있다.
# statsmodel 라이브러리가 좀 더 많고 규격화된 정보를 제공한다.

# 예제) 22 명의 심장 우회 수술을 받은 환자를 다음의 3 가지 그룹으로 나누었다.
#  Group I  : 50% 아산화 질소(nitrous oxide)와 50%의 산소(oxygen) 혼합물을 24 시간 동안 흡입한 환자
#  Group II : 50% 아산화 질소와 50% 산소 혼합물을 수술 받는 동안만 흡입한 환자
#  Group III : 아산화 질소 없이 오직 35-50%의 산소만 24 시간 동안 처리한 환자
#  그런 다음 적혈구의 엽산 수치를 24 시간 이후에 측정하였다.

import pandas as pd
import matplotlib.pyplot as plt

import numpy as np
import urllib  # url 로 데이터 얻어오기
```

```
url = 'https://raw.githubusercontent.com/thomas-haslwanter/statsintro_python/master/ipynb/Data/data_altman/altman_910.txt'
data = np.genfromtxt(urllib.request.urlopen(url), delimiter=',')

# Sort them into groups, according to column 1
group1 = data[data[: ,1 ]==1 ,0]
group2 = data[data[: ,1 ]==2 ,0]
group3 = data[data[: ,1 ]==3 ,0]

# matplotlib plotting
plot_data = [group1, group2, group3]
print(plot_data)
ax = plt.boxplot(plot_data)
plt.show()

# Boxplot 에서 볼 수 있듯이, 평균값의 차이가 실제로 의미가 있는 차이인지, 분산이 커서 그런 것인지 애매한 상황이다.
# 이런 상황에서 분산분석을 통해 통계적 유의성을 알아 볼 수 있다.
# Scipy.stats 으로 일원분산분석 : 아래와 같은 코드로 분산분석을 할 수 있다.

import scipy.stats as stats
f_statistic, pval = stats.f_oneway(group1, group2, group3)
print('Altman 910 데이터의 일원분산분석 결과 : F={0:.1f}, p={1:.5f}'.format(f_statistic, pval))  # F=3.7, p=0.04359

if pval < 0.05:
    print('P-value 값이 유의수준 보다 작음으로 인해 그룹의 평균값이 통계적으로 유의미하게 차이가 남.')



# 이번에는 pandas 와 statsmodels 라이브러리를 사용해서 분산분석을 수행.
# Statsmodel 을 사용한 일원분산분석

import statsmodels.formula.api as smf
import statsmodels.api as sm
#import warnings  # 경고 메세지 무시하기
#warnings.filterwarnings('ignore')

df = pd.DataFrame(data, columns=['value', 'treatment'])
# the "C" indicates categorical data

model = smf.ols('value ~ C(treatment)', df).fit()
print(sm.stats.anova_lm(model))  # statsmodels 을 사용하면 간편하게 결과를 얻을 수 있다.
```

print()

```
#                   df          sum_sq        mean_sq          F          PR(>F)
# C(treatment)     2.0     15515.766414     7757.883207   3.711336       0.043589
# Residual        19.0     39716.097222     2090.320906        NaN            NaN
```

anova_ex02.py

```
# 이원분산분석(two-way ANOVA)  ---------------------------
# 독립변인의 수가 두 개 이상일 때 집단 간 차이가 유의한지를 검증하는 데 사용한다.
# 상호작용효과(Interaction effect), 한 변수의 변화가 결과에 미치는 영향이 다른 변수의 수준에 따라 달라지는지를 확인하기 위해 사용된다.
# 예제 데이터(altman_12_6) 설명
# 태아의 머리 둘레 측정 데이터다. 4 명의 관측자가 3 명의 태아를 대상으로 측정함.
# 이를 통해서 초음파로 태아의 머리 둘레측정 데이터가 재현성이 있는지를 조사함.

url_base = 'https://raw.githubusercontent.com/thomas-haslwanter/statsintro_python/master/ipynb/Data/data_altman/'
url = url_base + 'altman_12_6.txt'
data = np.genfromtxt(urllib.request.urlopen(url), delimiter=',')

# Bring them in dataframe-format
df = pd.DataFrame(data, columns=['head_size', 'fetus', 'observer'])
print(df.tail())

#     head_size  fetus  observer
# 31       12.7    3.0       3.0
# 32       12.5    3.0       3.0
# 33       13.0    3.0       4.0
# 34       12.9    3.0       4.0
# 35       13.8    3.0       4.0

# 태아별 머리 둘레 plot 만들기
df.boxplot(column = 'head_size', by='fetus' , grid = False)
plt.show()

# 태아(fetus) 3 명의 머리둘레는 차이가 있어 보이나 이것이 관측자와 상호작용이 있는 것인지 분석을 통해 알아 보겠다.
# 분산분석으로 상관관계 확인: statmodels 라이브러리를 사용해 계산

formula = 'head_size ~ C(fetus) + C(observer) + C(fetus):C(observer)'
```

```
lm = smf.ols(formula, df).fit()  # 선형회귀를 이용
print(sm.stats.anova_lm(lm))
```

#	df	sum_sq	mean_sq	F	PR(>F)
# C(fetus)	2.0	324.008889	162.004444	2113.101449	1.051039e-27
# C(observer)	3.0	1.198611	0.399537	5.211353	6.497055e-03
# C(fetus):C(observer)	6.0	0.562222	0.093704	1.222222	3.295509e-01
# Residual	24.0	1.840000	0.076667	NaN	NaN

p-value 가 0.05 보다 작다. 따라서 귀무가설을 기각할 수 없고,
측정자와 태아의 머리 둘레값에는 연관성이 없다고 할 수 있다. 측정하는 사람이 달라도 머리 둘레값은 일정하다는 얘기.

해설 : 초음파로 측정하는 태아의 머리 둘레값은 믿을 수 있다고 판단할 수 있다.
분산분석(ANOVA)은 전체 그룹간의 평균값 차이가 통계적 의미가 있는지 판단하는데 유용한 도구다.
하지만 정확히 어느 그룹의 평균값이 의미가 있는지는 알려주지는 않는다.
그러므로 그룹 간의 관계를 보기 위해 추가적인 사후분석(Post Hoc Analysis)이 필요하다.

상관관계 분석

상관분석(Correlation Analysis)은 확률론과 통계학에서 두 변수 간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법 . 상관관계의 정도를 파악하는 **상관계수(Correlation coefficient)는 두 변수 간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아님.** 두 변수 간에 원인과 결과의 인과관계가 있는지에 대한 것은 회귀 분석을 통해 인과관계의 방향, 정도와 수학적 모델을 확인해 볼 수 있음

피어슨 상관 계수 상관관계 분석에서는 기본적으로 피어슨 상관 계수(Pearson correlation coefficient)는 두 변수 간의 관련성을 구하기 위해 보편적으로 이용됩니다.

피어슨 상관계수를 해석: r 값은 x 와 y 가 완전히 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일하면 -1 의 값을 갖습니다.

r 값	관계
+0.7 ~ +1.0	강한 양적 상관관계
+0.3 ~ +0.7	뚜렷한 양적 상관관계
+0.1 ~ +0.3	약한 양적 상관관계
-0.1 ~ +0.1	상관관계 거의 없음
-0.3 ~ -0.1	약한 음적 상관관계
-0.7 ~ -0.3	뚜렷한 음적 상관관계
-1.0 ~ -0.7	강한 음적 상관관계

Python 에서의 상관분석 예)

idx	a	b	c	d
0	1	10	0	-1
1	2	15	0	-20
2	3	20	0	-30
3	4	25	0	-45
4	5	50	0	-50
5	6	55	0	-55
6	7	60	0	-70

다음과 같은 데이터 a, b, c, d 라는 변수가 있는 상황에서 각각이 어떤 관계를 갖는지 상관분석

```
import pandas as pd

lst = [[1,2,3,4,5,6,7],
       [10,15,20,25,50,55,60],
       [0,0,0,0,0,0,0],
       [-1,-20,-30,-45,-50,-55,-70]]

df = pd.DataFrame(lst).T
corr = df.corr(method = 'pearson')
print(corr)
```

list 를 pandas DataFrame 으로 변환한 뒤 corr 메소드를 실행 . 기본적으로 '피어슨 상관계수'를 사용합니다.

실행 결과 :

	0	1	2	3
0	1.000000	0.966282	NaN	-0.983120
1	0.966282	1.000000	NaN	-0.917002
2	NaN	NaN	NaN	NaN
3	-0.983120	-0.917002	NaN	1.000000

즉, 다음과 같은 결과를 확인 가능

	a	b	c	d
a	1.000	0.966	NaN	-0.983
b	0.966	1.000	NaN	-0.917
c	NaN	NaN	NaN	NaN
d	-0.983	-0.917	NaN	1.000

1 과 가까울수록 양의 상관관계이며 -1 과 가까울수록 음의 상관관계

a,b 는 양의 상관관계이고 a,b 와 d 는 음의 상관관계를 갖는 것을 알 수 있음