비모수 통계 개념

비모수 통계란?(nonparametric method = free distribution)

- 통계학 분석 기법들은 크게 "모수적" 분석과 비모수적 분석 방법으로 분류됨
- 모수적 분석 : 데이터의 특정 분포를 가정함
- 비모수적 분석 : 데이터의 특정 분포를 가정하지 않음

비모수 통계가 왜 필요한지?

- 우리가 보통 알고 있고 사용하는 분석방법은 대부분 모수적 방법들이다
- 데이터가 모수적 분석방법이 가정한 특성을 만족하지 못할 때는 모수적 분석이 아니라 이에 해당하는 비모수적 방법을 적용해야 함
- 데이터 특성을 고려하지 못하고 그대로 적용하면 잘못된 결론을 내리게 됨
  -> 비모수적 방법은 모집단의 분포형태에 대한 가정을 완화하여 이론을 전개하기 때문에 가정이 만족되지 않음으로써 오류의 가능성이 적고 또한 계산이 간편하고 직관적으로 이해하기 쉽다는 장점이 있다.

비모수 통계를 언제 사용해야 하는지?

1. 평균 값이나 표준편차를 알 수 없을 때
2. 통계자료가 정규분포가 아닐 때
3. 주로 서열척도나 명목척도일 때 *(모수통계는 적어도 등간척도)*
4. 표본 수가 적을 때

비모수 통계 장 단점

장점:

1. 모집단의 분포가 어떻든지 사용할 수 있다.
2. 모수적 방법에 비해 계산이 간단하며, 적용하기 쉽다.
3. 관측값에 신뢰성이 적어 단순히 순위를 통해 검정하고 싶을 때 이용하면 좋다.

단점:

1. 모수적 방법으로 검정할 수 있는 데이터에 비모수적 방법을 이용하면, 효율성이 떨어진다.
2. 표본의 크기가 커질 수록 비모수적 방법의 계산량은 늘어난다.

기존의 정보가 정확할수록 모수통계방법이 신뢰도가 높고 기존의 정보가 없거나 부정확하다면 비모수통계방법이 더 신뢰도가 높습니다

비모수 통계 종류

사용목적	비모수 통계분석 방법	목적이 유사한 모수 통계 방법
적합도 검정	단일표본 카이스퀘어 검정	
	단일 표본 콜모고프 스미르노프 검정	
	이항분포 검정	
무작위성 검정	연의 검정 Run test	
두 변수의 비교	부호검정	대응표본 T 검정
	윌콕슨 부호 – 순위 검정	
	맥네마르 검정 McNemr Test	
세 변수의 비교	프리드만 검정 Friedman test 대응 K 표본	MANOVA
	켄탈의 일치 계수 (Kendall W test) 대응 K 표본	
	코크란 큐 검정 Cochran Q test	
두 집단의 비교	맨 휘트니 검정 윌콕슨 순위합검정	T 검정
	콜모고프 스미르노프 검정	
	월드 월포비치 검정 Wald-Wolfwiz Test	
	중앙값 검정	
세 집단 이상의 비교	중앙값 검정	anova
	클루스칼 윌리즈 검정 (Kruskal- wallis h test)	
변수 간의 상관분석	스피어만 순위 상관분석	correlation
	카이자승 분석 =교차분석	

적합도 검정

모집단이 일정한 확률분포 형태를 가진다고 가정할 경우 표본에서 얻어진 분포가 모집단에서 가정하고 있는 분포에 적합한지를 검정하는 방법이다.

카이제곱 검정

관찰된 빈도가 기대되는 빈도와 유의하게 다른지를 검정

cf) 카이제곱검정에는 두 가지 형태가 있으며, 같은 카이제곱 통계량과 분포를 사용하지만 다른 목적을 가짐

1) 적합도 검정: 관찰된 비율 값이 기대값과 같은지 조사하는 검정
(어떤 모집단의 표본이 그 모집단을 대표하는지)
2) 동질성 검정: 두 집단의 분포가 동일한지
3) 독립성 검정: 어떤 범주형 확률변수 x가 다른 범주형 확률변수 y와 독립인지 상관관계를 가지는지 검정하는 데 사용

눈의 수 x	1	2	3	4	5	6	계
나온 눈의 수 x	9	6	14	13	5	13	60

위에서 주사위를 60 던졌을 때 눈이 나올 확률이 1/6인지 검정해보아라.

Ho : p = 1/6
H1 : p != 1/6

In [75]:

```
x = [9,6,14,13,5,13]
```

In [84]:

```
stats.chisquare(x, 10, ddof = 5)
```

Out[84]:

```
Power_divergenceResult(statistic=7.6000000000000005, pvalue=nan)
```

In [87]:

```
from scipy.stats import chi2

chi2.ppf(0.95, df=5)
```

Out[87]:

```
11.070497693516351
```

11.07 > 7.6 이므로 귀무가설은 기각된다.

콜모고로프-스미르노프 검정 (K-S 검정)

콜모고로프-스미르노프 D 검정이라고도 한다.

표본의 분포가 가정한 분포와 적합한지 검정하는 방법이다
표본분포의 누적확률과 이론분포의 누적확률 분포의 차를 이용하는 검정

$$D = \max \cdot |F_o(X) - S_n(X)|$$

$F_o(X)$: 기대되는 상대적 누적도수

$S_n(X)$: 관찰된 상대적 누적도수

통계량 D는 표본의 누적확률분포와 가설로 설정된 누적 확률분포와의 최대 차이를 의미한다. D가 클수록 귀무가설을 기각한다.

만족도 조사를 위해 40명 고객을 대상으로 설문조사를 진행했다. 해당 설문조사 답안을 가지고 가설을 검증하기 전, 해당 데이터가 정규분포성을 띠는지 검정하여라.

귀무가설: 고객들의 고객만족도점수는 정규분포를 따른다.
연구가설: 고객들의 고객만족도점수는 정규분포를 따르지 않는다.

In [4]:

```
from statsmodels.stats.diagnostic import kstest_normal

import numpy as np

#만족도 조사를 위해 40명 고객들을 대상으로 설문을 응답 받음
x = [88, 75, 79, 84, 68, 51, 70, 75, 88, 90,
     92, 88, 63, 72, 94, 80, 78, 98, 81, 67,
     85, 87, 79, 81, 85, 48, 79, 86, 53, 100,
     87, 80, 80, 32, 60, 75, 62, 82, 40, 57]

x = np.array(x)

kstest_normal(x, dist='norm') #dist는 norm/exp 설정 가능
```

Out[4]:

```
(0.16508249090030575, 0.007856999983881514)
```

귀무가설을 기각하여 고객 만족도 점수는 정규분포를 따른다.

무작위성 검정

표본의 배열이 무작위로 구성되어 있는지 검정

연의 검정 (Run test)

런이란 동일한 관측값이 연속적으로 이어진 것을 말한다.

ex) 동전 반복 던졌을 때 '11001011100' 과 같이 나타났을 경우 11/00/1/0/111/00 으로 구분되어 6개의 런이라고 말할 수 있다.

런검정의 런의 수를 판단하여 무작위성을 검정하게 되는데, 런의 수가 매우 많거나 매우 적으면 관찰치 간의 연관성 있다고 할 수 있다.

A 쇼핑몰은 새로운 브랜드 런칭 이벤트를 지원하기 위하여 매장 방문 고객에게 상품 1만원권을 배포하였다. 매장 오픈 후 최초 20명의 방문이력을 조사한 결과 아래의 순서로 멤버십을 소지한 사람(1)과 소지하지 않은 사람(0)이 방문하였다. A 쇼핑몰의 CRM 팀에서는 이러한 마케팅 행사가 한쪽에 치우치지 않고 공정하게 이루어졌는지를 판단하기 위해 무작위성 검정을 진행하고자 한다.

1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 0 0 0 0 0

이 문제를 위한 가설을 설정하고 검정하시오.

1/0 0 0 0 /1 1 1 1/ 0 0 0 0/ 1 1/ 0 0 0 0 0

Ho: 멤버심 소지 고객과 비소지 고객의 방문은 무작위로 이루어졌다.
H1: 멤버십 소지 고객과 비소지 고객의 방문은 무작위로 이루어지지 않았다.

In [31]:

```
from statsmodels.sandbox.stats.runs import Runs
#statsmodels.sanbox.stats.runs import runstest_1samp

import numpy as np


x= np.array(x)

Runs(x).runs_test()
```

Out[31]:

(-1.8277470669267506, 0.06758752074917526)

유의수준 0.5 수준에서는 귀무가설을 기각하지 못해 무작위로 이루어졌다고 볼 수 있다.

두 변수의 비교

paired t-test와 유사

윌콕슨 부호-순위 검정

쌍체 표본 t-검정에서 표본의 수가 30개 미만일 경우 활용된다.
부호로 되어있거나 서열 검정에 쓰인다.

전자회사 C사는 기존의 물류 알고리즘보다 개선되었다고 알려진 새로운 물류 경로 최적화 알고리즘을 도입해 상품의 배송시간을 단축하고자 한다. 이에 전국 7개의 물류센터에 실험적으로 적용해보고 실제로 얼마나 더 나은 성과를 보이는지 검증해보고자 한다. 전국 7개의 물류센터에서 새로운 알고리즘의 적용 전 평균 배송시간과 적용 후의 평균 배송시간은 다음과 같다.

Ho: 기존 물류 알고리즘과 신규 알고리즘간을 통한 평균 배송시간은 차이가 없다.
H1: 기존 물류 알고리즘과 신규 알고리즘간을 통한 평균 배송시간은 차이가 있다.

In [33]:

```
from scipy.stats import wilcoxon

x = [10, 30, 9, 21, 35, 12, 17]
y = [8, 27, 16, 25, 30, 13, 11]


wilcoxon(x,y)
#wilcoxon(x-y)
```

Out[33]:

```
WilcoxonResult(statistic=12.0, pvalue=0.8125)
```

귀무가설을 기각하지 못해 알고리즘으로 개선이 되지 못한 것을 알 수 있다.

A 쇼핑 마케팅 팀에서는 새로운 로열티 프로그램을 제공하며 멤버십 기능을 강화하였다. 로열티 프로그램 만족도의 변화가 통계적으로 유의한 지 알아보기 위해 도입 전과 후의 고객만족도에 대한 검정을 수행하고자 한다.

사용데이터 : Ashopping.csv

Ho: 로열티 프로그램 제공 전 만족돠 후 만족도 차이는 없다.
H1: 로열티 프로그램 제공 전 만족돠 후 만족도 차이는 있다.

In [32]:

```
from scipy import stats

data = pd.read_csv("C:WWUsersWWjWWDesktopWWtestWWAshopping.csv", encoding = 'CP949')

# ttest_1samp
stats.ttest_rel(data["멤버쉽_프로그램_가입후_만족도"], data["멤버쉽_프로그램_가입전_만족도"])
```

Out[32]:

Ttest_relResult(statistic=29.560410783358122, pvalue=1.7319140513197275e-138)

귀무가설이 기각되어 로열티 프로그램 전과 후에는 만족도 차이가 있다.

맥네마르 검정

이항변수로 되어있는 두 변수간의 분포의 차이를 검정할 때 이용된다.
자료가 명목변수와 순위변수로 이루어져있을 때 적절하다.

A대학에서는 새로운 강의 방법을 도입하고자 한다. 이를 위해 새로운 강의법과 기존의 강의법을 이용하여 한 달간 각각 수업을 실시한 후 무작위로 추출된 10명의 학생들에게 찬성(=1) 및 반대(=2) 의사를 물었다. 두 강의법에 대한 학생들의 찬반 의견이 일치하는가를 검정하라

학생	1	2	3	4	5	6	7	8	9	10
새 강의법	1	1	2	2	1	2	2	1	1	1
기존 강의법	2	2	1	1	2	1	1	2	2	2

Ho : 찬반 의견은 일치한다.
H1 : 찬반 의견은 일치하지 않는다.

In [24]:

```
df = pd.DataFrame({"강의 종류": ["새 강의법","새 강의법","새 강의법","새 강의법","새 강의법","새 강의법","새 강의법","새 강의법","새 강의법","새 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법","기존 강의법"], "찬반":[1,1,2,2,1,2,2,1,1,1,2,2,1,1,2,1,1,2,2,2]})
```

In [26]:

```
df = pd.crosstab(index = df["강의 종류"], columns = df["찬반"])
df
```

Out[26]:

찬반	1	2
강의 종류		
기존 강의법	4	6
새 강의법	6	4

In [30]:

```
from statsmodels.stats.contingency_tables import mcnemar

print(mcnemar(df.values, exact=False))

#mcnemar(data, exact= False , correction = False)
#exact = True -> binomial distribution 사용
# FALSE = > chi2
# If true, then a continuity correction is used for the chisquare distribution (if exact is fals
e.
```

```
pvalue      0.7728299926844475
statistic   0.08333333333333333
```

0,77로 0.05보다 크므로 새로운 강의법과 기존강의법에 대한 찬반의사는 같다라는 귀무가설이 채택된다.

이산형 분포를 연속형인 정규분포 등으로 근사시켜 생기는 오차를 보정해줄 수 있는데, 이를 연속성 수정 (continuity correction)이라 한다. 이는 prop.test() 함수에 옵션으로 correct = TRUE 값을 추가함으로써 보정할 수 있다(default : FALSE)

cf) <bunch containing results, print to see contents>
print로 감싸주기

세 변수의 비교

MANOVA 와 유사

프리드만 검정

2개의 명목형 독립변수로 구분되는 세부 그룹 간의 차이를 검정한다는 관점에서 이원분산분석에 대한 비모수 통계분석 기법이라고 볼 수 있다.

A 쇼핑에서는 VIP 고객들을 대상으로 새로운 혜택을 제공하고자 한다. 샘플증정, 포인트 추가, 무료배송, 할인쿠폰 등 4가지 혜택에 대한 5개 지역별 고객들에 대한 사전 선호도 조사를 실시한 결과 지역별 서비스에 대한 서열은 아래 표와 같이 정리되었다. 혜택 별 고객 선호도에 차이가 있는지를 검정을 통해 알아보자.

지역	샘플증정	포인트추가	무료배송	할인쿠폰
서울경기	1	3	2	4
강원	2	3	4	1
충청	1	3	4	2
경상	1	2	4	3
전라	2	1	3	4

Ho: 혜택별 고객의 선호도 평가에 차이가 없다.
H1: 혜택별 고객의 선호도 평가 차이가 있다.

In [34]:

```
from scipy.stats import friedmanchisquare


a = [1,2,1,1,2]
b = [3,3,3,2,1]
c = [2,4,4,4,3]
d = [4,1,2,3,4]

friedmanchisquare(a,b,c,d)
```

Out[34]:

```
FriedmanchisquareResult(statistic=6.359999999999999, pvalue=0.09535032301698126)
```

즉, VIP고객들에게 큰 메리트는 없기 때문에, 새로운 혜택을 고려해보는 것이 필요할 수도 있음을 의미한다.

https://www.reneshbedre.com/blog/manova-python.html (https://www.reneshbedre.com/blog/manova-python.html)

공장 종류(A,B,C,D)에 따라서 공장 설비(높이, 캐노피 면적)에 차이가 발생하는 지 알아보자.

Ho: 공장 종류와 설비간에는 차이가 없다.
H1: 공장 종류와 설비간에는 차이가 있다.

In [35]:

```
#해당 코드를 실행하면 데이터가 로드됩니다.

df=pd.read_csv("https://reneshbedre.github.io/assets/posts/ancova/manova_data.csv")
```

In [36]:

```
from statsmodels.multivariate.manova import MANOVA

model = MANOVA.from_formula('height + canopy_vol ~ plant_var', data=df)
print(model.mv_test())
```

```
                  Multivariate linear model
============================================================

------------------------------------------------------------
       Intercept         Value  Num DF  Den DF F Value  Pr > F
------------------------------------------------------------
          Wilks' lambda  0.0374 2.0000 35.0000 450.0766 0.0000
         Pillai's trace  0.9626 2.0000 35.0000 450.0766 0.0000
 Hotelling-Lawley trace 25.7187 2.0000 35.0000 450.0766 0.0000
    Roy's greatest root 25.7187 2.0000 35.0000 450.0766 0.0000
------------------------------------------------------------

------------------------------------------------------------
       plant_var         Value  Num DF  Den DF F Value  Pr > F
------------------------------------------------------------
          Wilks' lambda  0.0797 6.0000 70.0000  29.6513 0.0000
         Pillai's trace  1.0365 6.0000 72.0000  12.9093 0.0000
 Hotelling-Lawley trace 10.0847 6.0000 44.9320  58.0496 0.0000
    Roy's greatest root  9.9380 3.0000 36.0000 119.2558 0.0000
============================================================
```

Pillai's Trace 검정 통계량은 통계적으로 유의하며 공장 종류에 따라 공장설비가 유의한 연관성을 가지고 있음을 알 수 있다.

- Pillai's Trace : 양 값의 통계량이며, 이 값의 증가는 처리 효과가 모형에 미치는 영향이 큰 것임을 의미한다.

*- Wilks's Lambda : 0과 1사이의 값을 가지는 통계량이며, 이 통계량 값이 작다는 것은 처리효과가 모형에 미치는 영향이 더 크다는 것을 의미한다.

*- Hotelling's Trace: 검정 행렬의 고유값의 합으로 양의 값을 가진 통계량이다. 이 값이 높으면 처리 효과가 모형에 미치는 영향이 크다는 것을 의미한다.

- Roy's Largest Root : 검정 행렬의 가장 큰 고유값을 의미한다. 때문에, 양의 값을 가진 통계량이며, 이 값이 증가하면 처리 효과가 모형에 미치는 영향이 더 크다는 것을 의미한다.

켄달의 W 검증 (켄달의 일치 계수)

Kendall의 w검증은 Friedman 검증과 거의 유사한 계산 과정을 거치게 된다. Friedman 검증에 Kendall의 일치도 계수가 추가되었다고 봐도 된다.

<일치도 계수>
완전일치는 1, 완전불일치는 0

cf) Kendall tau는 두 데이터 집합이 같은 방향으로 변화하는 경향이 있는지 여부를 정량화하려고 하는 반면, 켄달의 W는 두 데이터 집합이 실제로 동일한지 여부를 정량화하려고 합니다.

롯데시네마에서 각 회원 5명을 뽑아 최근 영화 4개에 대해 평가를 내리게 했다. 가장 재밌었던 영화는 1등, 재미없었던 영화는 4등으로 평가를 하게했다고 가정하자.(즉 서열 척도로 평가를 한 것이다.) 각 회원 5명의 영화에 대한 평가는 어느 정도로 일치하는지 검정해보자.

구분	암살	베테랑	사도	검은사제
회원1	1	2	4	3
회원2	2	1	3	4
회원3	1	3	4	2
회원4	1	3	2	4
회원5	2	3	4	1
서열척도의 합	7	12	17	14

Ho: 다섯명의 평가 순위는 일치하지 않을 것이다.
H1: 다섯명의 평가 순위는 일치할 것이다.(비슷할 것이다)

In [98]:

```
from scipy.stats import friedmanchisquare

a = [1, 2, 4, 3]
b = [2, 1, 3, 4]
c = [1, 3, 4, 2]
d = [1, 3, 2, 4]
e = [2, 3, 4, 1]

friedmanchisquare(a,b,c,d,e)
```

Out[98]:

```
FriedmanchisquareResult(statistic=0.4242424242424346, pvalue=0.9804444208267745)
```

0.98로 귀무가설을 기각하지 못해 순위는 일치하지 않을 것이다.

일치도 계수 구하는 공식

검정통계량 W(a)를 구하는 공식은 다음과 같다.

$$W=\frac{S}{\frac{1}{12}K^2(N^3-N)}$$

(S는 각 영화 서열척도에 대한 평균편차, K는 평가인의 명수, N은 평가대상의 개수 즉 영화의 개수)

우선 S를 구하는 공식은 서열척도 합에 대한 평균 $\overline{R}$를 구한다.

$$\overline{R}=\frac{\sum R}{N}=\frac{7+12+17+14}{4}=12.5$$

$$S=\sum(R_i-\overline{R})^2$$
$$=(7-12.5)^2+(12-12.5)^2+(17-12.5)^2+(14-12.5)^2$$
$$=53$$

$$W=\frac{S}{\frac{1}{12}K^2(N^3-N)}=\frac{53}{125}=0.424$$

코크란 큐 검정

이항변수로 되어있는 3개 이상의 변수간 비율차이를 검정하는 방법이다.
변수 2개면 맥네마르

K화장품 회사에서는 3가지 판매전략을 구사하고 있다. 이 판매전략들의 효과에 차이가 있는지를 조사하기 위해 판매사원 13명에게 판매시 사용하고 있는 판매전략에 대해 기입하도록 하였다. (1=판매성공, 2=판매실패) 판매전략 종류에 따라 판매효과에 차이가 있는지를 검정하라.

학생	1	2	3	4	5	6	7	8	9	10	11	12	13
판매전략1	1	2	1	2	2	1	2	1	2	1	1	1	2
판매전략2	1	1	2	1	2	1	1	2	1	2	1	1	1
판매전략3	2	2	2	1	1	1	2	2	2	1	2	1	1

Ho: 판매전략별 효과 차이가 없다.
H1: 판매전략별 효과 차이가 있다.

In [38]:

```
data = pd.read_csv("C:₩₩Users₩₩j₩₩Desktop₩₩test₩₩sell.csv")
```

In [48]:

```
df = pd.crosstab(index = data["판매종류"], columns = data["효과여부"])
```

In [49]:

```
df.values
```

Out[49]:

```
array([[7, 6],
       [9, 4],
       [6, 7]], dtype=int64)
```

In [51]:

```
from statsmodels.stats.contingency_tables import cochrans_q

print(cochrans_q(df.values))
```

```
df          1
pvalue      0.31731050786291115
statistic   1.0
```

귀무가설을 기각하지 못하므로 판매전략별 차이가 없다는 걸 알 수 있다.

두 집단의 비교

t-test와 유사

맨 휘트니 검정

두 개의 독립된 집단간의 특정 값의 평균을 비교하는 검정이다.
두 개의 표본 집단간의 차이를 검정한다.

V 반도체 회사는 공장 A와 공장 B 2개의 공장에서 반도체를 생산하고 있다. 이 때, 2개 공장의 생산 효율성에 차이가 있는지를 Mann-Whitney U 검정을 통해 살펴본다.

Ho: 공장 A와 공장 B의 생산효율성은 동일하다.
H1: 공장 A와 공장 B의 생산효율성은 다르다.

In [7]:

```
import pandas as pd
from scipy.stats import mannwhitneyu

x = [12, 11, 13, 14, 15]
y = [16, 15, 17, 19, 20]

print(mannwhitneyu(x, y))
```

```
MannwhitneyuResult(statistic=0.5, pvalue=0.015970696353780123)
```

In [8]:

```
xy = pd.DataFrame(x+y)
xy['생산량 순위'] = xy.rank(ascending=False)
xy['공장이름'] = ["A", "A", "A", "A", "A", "B", "B", "B", "B", "B"]
print(xy.groupby('공장이름').mean())
```

```
         0  생산량 순위
공장이름
A     13.0     7.9
B     17.4     3.1
```

즉, A와 B의 공장 생산의 효율성은 다르며, A의 효율성이 떨어지는 것을 볼 수 있다.

윌콕슨 순위합 검정

이 검정방법은 독립적인 두 표본으로 검정하며, 두 표본의 모집단의 중앙값이 동일한지를 검정한다.
이 검정은 만-위트니 U검정(Mann-Whitney U-test)와 일치하며, independent two sample t-test의 대안법이다.

계산은 두 표본의 결과를 오름차순으로 정렬하고 순위를 부여 한다. 순위 부여 시 결과가 같으면 해당 순위의 평균값을 동일하게 적용한다. 그리고 표본의 순위들을 각각 합한다. 이 표본들의 순위합과 표본들의 개수를 사용하여 검정통계량을 계산하고 p-value를 확인하여 어떤 가설을 채택할 것인지 판단한다.

A사 닭가슴살 제품의 중량과 B사 닭가슴살 제품의 중량이 차이가 있는지 확인해보자.

Ho : A사 닭가슴살 중량의 중앙값과 B사 닭가슴살 중량의 중앙값은 차이가 없다.

H1 : A사 닭가슴살 중량의 중앙값과 B사 닭가슴살 중량의 중앙값은 차이가 있다.

In [63]:

```
data = pd.read_csv("C:₩₩Users₩₩j₩₩Desktop₩₩test₩₩ranksum.csv")
```

In [72]:

```
data_a = data[data.company == "A"]["weight"]
data_b = data[data.company == "B"]["weight"]
```

In [104]:

```
data_a.head(5)
```

Out[104]:

```
0     97.604131
1     95.963639
2    101.239447
3     99.734285
4     99.130079
Name: weight, dtype: float64
```

In [73]:

```
from scipy.stats import ranksums

ranksums(data_a, data_b)
```

Out[73]:

```
RanksumsResult(statistic=-4.861187873450645, pvalue=1.1668344724558955e-06)
```

귀무가설을 기각하여 두 그룹값의 중앙값의 차이는 0이 아니므로 A사 닭가슴살 중량의 중앙값은 차이가 있다.

In [109]:

```
print("a의 중앙값:", np.median(data_a))
print("b의 중앙값:", np.median(data_b))
```

```
a의 중앙값: 99.9981956567448
b의 중앙값: 109.943107053405
```

b사의 닭가슴살 중앙값이 a사 닭가슴살 중앙값보다 유의하게 더 크다.

세 집단 이상의 비교

anova와 유사

클루스칼 윌리즈 검정

3개 이상의 독립 표본집단간의 본포가 동일한지 비교하는 검정 방법이다.

연구배경: Q제철기업은 철근을 생산하는데 있어 3개의 공장을 운영하고 있다. 생산량이 가장 낮은 공장을 찾아 공장 설비 등을 보강시켜 생산량을 높일 계획이다.

가설검증
귀무가설: 3개의 공장의 철근 생산량은 모두 동일하다.
연구가설: 3개의 공장의 철근 생산량은 모두 동일하지는 않다.

In [15]:

```
import pandas as pd
from scipy.stats import kruskal

a = [35, 41, 45, 42, 33, 36, 47, 45, 31, 32, 40, 44]
b = [40, 38, 44, 48, 45, 46, 42, 39, 40, 41, 38, 47]
c = [30, 34, 38, 39, 40, 41, 38, 37, 40, 41, 39, 38]

# Kruskal-Wallis H 검정 분석
print(kruskal(a, b, c))

# 생산량 평균 순위 출력
data = pd.DataFrame(a+b+c)
data["생산량순위"] = data.rank(ascending=False)
data["공장이름"] = ''
data["공장이름"][0:12] = 1
data["공장이름"][12:24] = 2
data["공장이름"][24:36] = 3

print(data.groupby("공장이름").mean())
```

```
KruskalResult(statistic=6.047476974964328, pvalue=0.04861911622342764)
              0  생산량순위
공장이름
1      39.250000    19.0
2      42.333333    13.0
3      37.916667    23.5
```

즉, 3개의 철근 생산량을 모두 동일하지 않으며, 그 중에서도 공장 3의 평균 생산량 순위가 가장 낮음을 알 수 있다.

변수 간의 상관분석

correalation 과 유사

스피어만 순위 상관 분석

A 쇼핑은 1회 평균 매출액이 높은 고객 100명과 방문빈도가 높은 고객 100명을 선별하여 특별한 사은행사를 기획하고자 한다. 두 가지 변수를 기준으로 순위를 선정하였을 때 선별된 고객들이 동질적이라면 두 가지 기준으로 추출하지 않아도 될 것이다.

Ho : 1회 평균 매출액 순위와 방문빈도 순위는 연관성이 없다
H1 : 1회 평균 매출액 순위와 방문빈도 순위는 연관성이 있다

In [61]:

```
data = pd.read_csv("C:₩₩Users₩₩j₩₩Desktop₩₩test₩₩Ashopping.csv", encoding = 'CP949')

df = data[["1회_평균매출액", "방문빈도"]]
```

In [62]:

```
df.head(2)
```

Out[62]:

	1회_평균매출액	방문빈도
0	235711	17
1	226314	14

In [60]:

```
from scipy import stats

stats.spearmanr(df["1회_평균매출액"], df["방문빈도"])
```

Out[60]:

```
SpearmanrResult(correlation=-0.4988411248473936, pvalue=4.929293870381245e-64)
```

Kendall 서열상관 분석 (켄달타우)

스피어만 상관계수는 값에 순위를 매겨 그 순위에 대한 상관계수를 구하는 반면, 켄달 타우는 두 변수들 간의 순위를 비교하여 연관성을 계산한다.

켄달 타우는 샘플 사이즈가 작거나 데이터의 동률이 많을 때 유용하다.
스피어만은 데이터 내 편차와 에러에 민감하며 일반적으로 켄달 상관계수보다 높은 값을 가진다.

h 홈쇼핑에서는 최근 주력 판매 품목이었던 의류/패션 상품의 매출이 급락하여 시급한 대착마련이 필요하게 되었다. 패션을 포함한 소비재 상품을 취급하는 팀의 부장은 매출부진이 비교적 값 비싼 상품을 판매했기 때문이라고 지적하였고, CRM 팀에서는 이러한 지적이 사실인지 판단하기 위해 최근 판매된 5개 의류 브랜드의 가격과 판매량을 바탕으로 검증하기로 하였다.

브랜드	가격	판매량
A	5	4
B	2	1
C	4	3
D	1	2
E	3	5

Ho: H홈쇼핑의 의류브랜드 가격 서열과 판매량 서열은 연관성이 없다.
H1: H홈쇼핑의 의류브랜드 가격 서열과 판매량 서열은 연관성이 있다.

In [1]:

```
from scipy.stats import kendalltau

x= [5,2,4,1,3]
y= [4,1,3,2,5]

kendalltau(x,y)
```

Out[1]:

```
KendalltauResult(correlation=0.39999999999999997, pvalue=0.48333333333333334)
```

In [3]:

```
from scipy import stats
stats.spearmanr(x,y)
```

Out[3]:

```
SpearmanrResult(correlation=0.6, pvalue=0.28475697986529375)
```

p-value는 0.48로 귀무가설을 기각하지 못해 최근 의류 판매 브랜드의 판매 가격 서열과 판매량 서열은 서로 연관성이 없다고 해석해야 한다. 의류 판매량 부진은 판매가격이 아닌 다른 곳에서 찾아봐야 함..