제 6 장. 분포에 관한 추론

6.1 모평균에 관한 추론

$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, μ: 모평균, σ: 모표준편차

- 모표준편차 σ를 모를 때 모평균 μ에 관한 추론: $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ 임을 이용 ($\bar{X}$: 표본평균, S^2: 표본분산)
- μ의 $100(1-\alpha)\%$ 신뢰구간:

$$\bar{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}$$

[참고]

t분포: $X_1, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$에서 추출된 임의표본일 때

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

이때 $t(n-1)$: 자유도 $n-1$인 t분포

1. t분포의 일반적인 정의 (스튜던트 t분포): $Z \sim N(0,1)$, $V \sim \chi^2(k)$, Z와 V는 서로 독립일 때 $T = \frac{Z}{\sqrt{V/k}}$는 자유도 k인 t분포를 따른다.
2. $\bar{X}$를 표준화할 때 σ 대신에 S를 사용하는 것을 스튜던트화(Studentize)라 한다.

일표본 t-검정

귀무가설 $H_0 : \mu = \mu_0$

검정통계량 $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$

유의수준 α 에서 기각역:

$$H_1 : \mu > \mu_0 \Rightarrow T \geq t_\alpha(n-1)$$
$$H_1 : \mu < \mu_0 \Rightarrow T \leq t_\alpha(n-1)$$
$$H_1 : \mu \neq \mu_0 \Rightarrow |T| \geq t_{\alpha/2}(n-1)$$

예 1) 어느 대학의 신입생 가운데 랜덤하게 15명을 뽑아 심리검사를 실시한 결과 책임감에 대한 점수가 다음 표와 같았다. 이 대학 신입생의 평균점수가 40점 이상이라고 할 수 있는지 유의수준 $\alpha = 0.05$에서 검정해 보자.

점수
22 25 34 35 41 41 46 46 46 47 49 54 54 59 60

[풀이] 주어진 문제에 대한 가설은 다음과 같다.

$$H_0 : \mu = 40 \qquad H_1 : \mu > 40$$

일표본 t-검정 (One-sample t-test)을 위해서는 `scipy.stats` 의 `ttest_1samp` 함수를 사용한다. 검정을 원하는 자료와 귀무가설에서 지정된 모수 값을 설정하여 검정을 수행한다.

In [ ]:

```
from scipy.stats import ttest_1samp
ttest_1samp(a, popmean)
```

입력 값:

- a: 검정을 원하는 자료
- popmean: 귀무가설에서 지정된 모수 값

출력 값:

- statistic: t-값
- pvalue: 양측 유의확률

학생들의 책임감에 대한 점수를 score 변수에 입력하면 주어진 가설에 대한 t-검정은 다음과 같이 수행할 수 있다.

In [1]:

```
import numpy as np
from scipy.stats import ttest_1samp
score = np.array([22, 25, 34, 35, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60])
ttest_1samp(score, popmean=40)
```

Out[1]:

```
Ttest_1sampResult(statistic=1.354469698602765, pvalue=0.1970442368445378)
```

In [2]:

```
## 통계량 혹은 유의확률만 따로 보고 싶으면
T = ttest_1samp(score, popmean=40).statistic
pvalue = ttest_1samp(score, popmean=40).pvalue
print(T)
print(pvalue)
```

```
1.354469698602765
0.1970442368445378
```

이 때 pvalue=0.1970는 양측 검정에 해당하는 유의확률임에 주목하자. 대칭 분포에서 단측 유의확률은 양측 유의확률을 2로 나눈 값이 되며, 이 경우 단측 유의확률은 0.1970/2 = 0.09852이다.

검정통계량은 1.3545이고 유의확률은 0.09852 이므로 주어진 유의수준보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 신입생의 평균 점수는 40점 이상이라고 말할 수 없다.

예 2) B 회사 제품인 어느 통조림은 내용물 함량이 400g으로 표시되어있다. 이를 검사하기 위하여 이 회사 제품 10개를 시중에서 임의로 추출하여 조사한 결과가 다음과 같다. 내용물은 올바르게 표시되어 있는지 유의수준 5%에서 검정하고 평균 함량의 95% 신뢰구간을 구하여라.

함량

함량
408 405 397 405 395 415 389 403 397 390

신뢰구간을 구하기 위해 t분포의 분위수가 필요하므로 scipy.stats 의 t 모듈을 불러와 분위수 함수 ppf를 이용한다.

In [3]:

```
from scipy.stats import t
mass = np.array([408, 405, 397, 405, 395, 415, 389, 403, 397, 390])
n = len(mass)
xbar = mass.mean()
width = t.ppf(0.975, n-1) * np.std(mass, ddof=1) / np.sqrt(n)
print(xbar-width, xbar+width)
```

394.508731970426 406.2912680295739

참고: np.var() 혹은 np.std()에서 ddof=1을 설정하면 표본분산 혹은 표본표준편차를 출력한다.

np.var(X, ddof) = $\frac{1}{n-ddof}\sum_{i=1}^{n}(X_i-\bar{X})^2$

6.2 대응비교를 통한 모평균의 비교

대응비교 또는 **쌍체비교** (paired comparison):

실험단위를 동질적인 쌍으로 묶고, 각 쌍에 두 처리를 랜덤하게 적용한 다음, 각 쌍에서 얻은 관측값의 차로 두 모평균의 차인 $\mu_1-\mu_2=\delta$ 에 대한 추론 문제를 다루는 방법이다.

관측값: $(X_1,Y_1),\cdots,(X_n,Y_n)$

차: $D_i=X_i-Y_i, \quad i=1,\cdots,n$

$\mu_1-\mu_2=\delta$에 관한 추론문제: $D_1,\cdots,D_n$에 기초

D_i들의 표본평균과 표본분산을 $\bar{D}, S_D^2$이라 할 때

1. $\mu_1-\mu_2=\delta$ 의 $100(1-\alpha)\%$ 신뢰구간:
$$\bar{D}\pm t_{\alpha/2}(n-1)\cdot\frac{S_D}{\sqrt{n}}$$
2. $H_0:\mu_1-\mu_2=\delta_0$에 대한 검정통계량:
$$T=\frac{\overline{D}-\delta_0}{S_D/\sqrt{n}}$$

예 1) (paired.txt) 색감의 차이가 감정변화에 미치는 영향을 연구하기 위하여 14명을 랜덤으로 선택하여 이들을 60초 간격으로 보라색과 초록색에 반복적으로 노출시키는 실험을 6분간 지속하였다. 각 색이 변할 때마다 최초 12초간 피부에 나타나는 전기반응을 측정하여, 각 색 별로 평균을 취한 후, 이것을 최종 자료로 선택하였다. 다음 자료를 이용하여 보라색과 초록 색 사이에 감정변화에 미치는 영향이 존재하는지를 유의수준 5%에서 검정하시오. 단, 자료는 모두 정규분포 가정을 만족한다고 하자.

색깔														
보라 (X)	3.1	3.7	4.0	3.2	3.6	3.5	4.2	3.8	3.7	3.4	3.6	3.8	3.4	3.4
초록 (Y)	2.2	2.7	3.1	2.9	3.3	2.6	2.9	2.8	3.2	2.5	3.5	3.1	2.3	3.5

[풀이] $D_i = X_i - Y_i$ 라고 정의하면 주어진 문제에 대한 가설은 다음과 같다.

$$H_0 : \mu_D = 0 \qquad H_1 : \mu_D \neq 0$$

In [4]:

```
import pandas as pd
paired = pd.read_csv("paired.txt", sep=" ")
```

대응비교를 위해서는 scipy.stats 의 ttest_rel 함수를 사용한다.

In [ ]:

```
from scipy.stats import ttest_rel
ttest_rel(a, b)
```

입력 값:

- a, b: 검정을 원하는 자료

출력 값:

- statistic: t-값
- pvalue: 양측 유의확률

주어진 데이터에 대한 대응 비교 실행은 다음과 같다.

In [5]:

```
from scipy.stats import ttest_rel
ttest_rel(paired['purple'], paired['green'])
```

Out[5]:

```
Ttest_relResult(statistic=6.3380731434065325, pvalue=2.5838913496640584e-05)
```

검정통계량의 값은 6.3381이고 유의확률은 0.00002584로 매우 작기 때문에 귀무가설을 기각할 수 있다. 따라서 보라색과 초록색 사이에는 감정변화에 미치는 영향이 존재한다고 볼 수 있다.

예 2) 다음은 20명의 학생들에게 특정 수업을 받기 전과 후의 시험 성적을 비교해 놓은 자료이다. 수업 이수가 학생들의 시험 성적 향상에 영향을 끼쳤다고 말할 수 있는지 유의수준 5% 에서 이를 검정하시오.

시험 성적																				
수업 전	18	21	16	22	19	24	17	21	23	18	14	16	16	19	18	20	12	22	15	17
수업 후	22	25	17	24	16	29	20	23	19	20	15	15	18	26	18	24	18	25	19	16

In [6]:

```
score_bef = np.array([18, 21, 16, 22, 19, 24, 17, 21, 23, 18, 14, 16, 16, 19, 18, 20, 12, 22, 15, 17
score_aft = np.array([22, 25, 17, 24, 16, 29, 20, 23, 19, 20, 15, 15, 18, 26, 18, 24, 18, 25, 19, 16

ttest_rel(score_bef, score_aft)
```

Out[6]:

```
Ttest_relResult(statistic=-3.231252665580312, pvalue=0.004394965993185664)
```

6.3 이표본에 의한 모평균의 비교

두 모집단의 분포는 $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$을 가정한다. 아래에서 $\bar{X}_i$, S_i^2는 i번째 모집단의 표본의 표본평균, 표본분산을 의미한다.

1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (등분산)일 때

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

단, $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$: 공통분산 σ^2의 합동추정량

2. $\sigma_1^2 \neq \sigma_2^2$ (이분산)일 때

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(w)$$

단, $w = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$

$\mu_1 - \mu_2$의 $100(1-\alpha)\%$ 신뢰구간

1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (등분산)일 때

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2)S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

2. $\sigma_1^2 \neq \sigma_2^2$ (이분산)일 때

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}(w)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

귀무가설 $H_0 : \mu_1 - \mu_2 = \delta_0$의 검정법

검정통계량:

1. 등분산을 가정할 때

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

2. 이분산을 가정할 때

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

유의수준 α에서 기각역

$$H_1 : \mu_1 - \mu_2 > \delta_0 \Rightarrow T \geq t_\alpha(df)$$
$$H_1 : \mu_1 - \mu_2 < \delta_0 \Rightarrow T \leq t_\alpha(df)$$
$$H_1 : \mu_1 - \mu_2 \neq \delta_0 \Rightarrow |T| \geq t_{\alpha/2}(df)$$

단, $df = n_1 + n_2 - 2$ (등분산) 또는 $df = w$ (이분산)

독립 이표본 검정에서 등분산 가정 여부는 두 모분산에 대한 검정 결과를 이용할 수 있다.

6.3.1 두 모분산에 관한 추론

F-분포

$S_1^2 : N(\mu_1, \sigma_1^2)$에서 크기 n_1 인 임의표본의 표본분산

$S_2^2 : N(\mu_2, \sigma_2^2)$에서 크기 n_2 인 임의표본의 표본분산 (두 표본은 서로 독립)

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

: 자유도 $(n_1 - 1, n_2 - 1)$인 F-분포

[참고]

1. F-분포의 일반적인 정의: $V_1 \sim \chi^2(k_1)$이고 $V_2 \sim \chi^2(k_2)$이며 V_1과 V_2는 서로 독립일 때
$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$$
2. F-분포의 성질: $F \sim F(k_1, k_2)$일 때 $1/F \sim F(k_2, k_1)$
3. 왼쪽 꼬리의 F_α 값: $F_{1-\alpha}(k_1, k_2) = \frac{1}{F_\alpha(k_2, k_1)}$ 예. $F_{0.95}(5, 10) = \frac{1}{F_{0.05}(10,5)} = \frac{1}{4.74} = 0.21$

두 모분산의 비에 관한 검정

등분산성의 검정 (F-검정)

가설 $H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2$

검정통계량: $F = \frac{S_1^2}{S_2^2}$

유의수준 α에서 기각역

$$F > F_{\alpha/2}(n_1 - 1, n_2 - 2) \quad \text{or} \quad F < F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$$

예 1) (paint.txt) 한 페인트 제조회사에서는 새 유성페인트 상품을 개발하여 기존 페인트와의 건조속도를 비교하고자 한다. 이를 확인하기 위해 시중에서 가장 인기 있는 상품과 새 상품을 각각 10종류의 벽에 칠한 후 건조시간

을 측정하였다. 새 상품은 기존의 상품보다 건조 시간이 더 빠르다고 할 수 있는가? 유의 수준 5%에서 검정해보자.

	건조 시간 (분)
기존 상품	49 44 47 44 46 40 48 45 45 42
새 상품	44 41 45 44 43 39 42 40 40 42

[풀이] 기존 상품의 건조시간의 모평균을 μ_1, 새 상품의 건조시간의 모평균을 μ_2라고 하면 검정을 위한 가설은 다음과 같다.

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 > 0$$

독립 이표본 검정의 자료구조는 대응표본과는 다르게 그룹을 나타내는 변수(group, 1=기존 상품, 2=새 상품)와 검정 대상이 되는 변수 (time, 건조 시간)로 구성되어 있다.

In [7]:

```
paint = pd.read_csv("paint.txt", sep=" ")
```

먼저 각 그룹별 건조시간의 평균을 비교해보자.

In [8]:

```
print("기존 상품(group 1) 건조시간의 평균:", paint[paint['group'] == 1]['time'].mean())
print("새 상품(group 2) 건조시간의 평균:", paint[paint['group'] == 2]['time'].mean())
```

```
기존 상품(group 1) 건조시간의 평균: 45.0
새 상품(group 2) 건조시간의 평균: 42.0
```

독립 이표본 평균 검정에 앞서 등분산 여부에 관한 모분산 검정을 먼저 시행한다. 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \sigma_1^2/\sigma_2^2 = 1 \quad \text{vs} \quad H_1 : \sigma_1^2/\sigma_2^2 \neq 1$$

주어진 두 그룹의 분산 검정은 다음과 같이 시행가능하다.

In [9]:

```
from scipy.stats import f

def var_test(sample1, sample2):
    n1 = len(sample1)
    n2 = len(sample2)
    S1 = np.var(sample1, ddof=1)
    S2 = np.var(sample2, ddof=1)
    dfn = n1 - 1
    dfd = n2 - 1
    F = S1 / S2
    pval = 2*min(f.cdf(F, dfn = dfn, dfd = dfd),1-f.cdf(F, dfn=dfn, dfd=dfd))

    print("F test to compare two variances: F = %s, num df = %s, denom df = %s, p-value = %s"
          % (round(F,5),dfn,dfd,round(pval,5)))       # 등분산 검정을 위한 함수
```

In [10]:

```
group1 = paint[paint['group'] == 1]['time']
group2 = paint[paint['group'] == 2]['time']

var_test(group1, group2)
```

```
F test to compare two variances: F = 1.83333, num df = 9, denom df = 9, p-value = 0.
37999
```

등분산 여부 검정 결과, 검정 통계량은 1.833이고 유의확률은 0.38이었다. 따라서 주어진 유의수준 5%에서 두 모집단의 분산이 같다는 귀무가설을 기각할 수 없다.

따라서 등분산을 가정한 독립 이표본 평균 검정 결과는 다음과 같다. 독립 이표본 검정을 위해서는 `scipy.stats`의 `ttest_ind` 함수를 사용한다.

In [ ]:

```
from scipy.stats import ttest_ind
ttest_ind(a, b, equal_var=True)
```

입력 값:

- a, b: 검정을 원하는 자료
- equal_var: 옵션을 이용하여 독립 이표본 검정의 등분산 가정 여부를 선택한다. equal_var 옵션의 기본값은 True이다.

출력 값:

- t: t-값
- pvalue: 양측 유의확률

In [11]:

```
from scipy.stats import ttest_ind
ttest_ind(group1, group2, equal_var=True)
```

Out[11]:

```
Ttest_indResult(statistic=2.8180093098831724, pvalue=0.0113883036492929)
```

이 때 pvalue=0.0114는 양측 유의확률이다. 대칭 분포에서 단측 유의확률은 양측 유의확률을 2로 나눈 값이므로, 이 경우 단측 유의확률은 0.005694이다. 검정통계량은 2.818이고 유의확률은 0.005694로 나타났다. 이는 주어진 유의수준 0.05보다 작기 때문에 귀무가설을 기각할 수 있다. 따라서 새 페인트의 건조시간은 기존 페인트의 건조시간 보다 더 빠르다고 말할 수 있다.

예 2) 다음은 두 집단에서 조사한 체질량 지수의 자료이다. 집단 별로 체질량지수는 차이가 있다고 볼 수 있는가? 유의수준 5%에서 이를 검정하시오.

	체질량 지수
그룹 1	22 23 25 26 27 19 22 28 33 24
그룹 2	21 25 36 24 33 28 29 31 30 32 33 35

In [12]:

```
bmi1 = np.array([22, 23, 25, 26, 27, 19, 22, 28, 33, 24])
bmi2 = np.array([21, 25, 36, 24, 33, 28, 29, 31, 30, 32, 33, 35])
```

In [13]:

```
var_test(bmi1, bmi2)
```

```
F test to compare two variances: F = 0.7267, num df = 9, denom df = 11, p-value = 0.
64207
```

등분산 검정 결과를 바탕으로 등분산성을 가정할 수 있다. (p-value = 0.64207 > 0.05)

In [14]:

```
ttest_ind(bmi1, bmi2, equal_var=True)
```

Out[14]:

```
Ttest_indResult(statistic=-2.643712672303319, pvalue=0.01557793053055382)
```