제 6장. 분포에 관한 추론

6.1 모평균에 관한 추론

- σ를 모를 때 모평균 μ에 관한 추론문제 : $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$ 임을 이용
- σ를 모를 때 모평균 μ의 $100(1-\alpha)\%$ 신뢰구간

$$\overline{x} \pm t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}} = \left(\overline{x} - t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}},\quad \overline{x} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}} \right)$$

(참고)

t-분포 : $X_1, X_2, \cdots, X_n$ 이 정규분포 $N(\mu, \sigma^2)$에서 추출된 임의표본일 때

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

: 자유도 $n-1$ 인 t-분포

참고. t-분포

① t-분포의 일반적인 정의 (스튜던트 t-분포)

$Z \sim N(0,1)$, $V \sim$ 자유도 k인 카이제곱 분포, Z와 V는 서로 독립일 때

$T = \dfrac{Z}{\sqrt{V/k}}$ 는 자유도 k인 t-분포를 따른다

② $\overline{X}$를 표준화시킬 때 σ 대신에 S를 사용하는 것을 스튜던트화(Studentize)

<u>일표본 t-검정</u>

귀무가설 : $H_0: \mu = \mu_0$

검정통계량 : $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

유의수준 α에서 기각역 : $H_1: \mu > \mu_0$ 일 때 $T \geq t_\alpha(n-1)$

$H_1: \mu < \mu_0$ 일 때 $T \leq -t_\alpha(n-1)$

$H_1: \mu \neq \mu_0$ 일 때 $|T| \geq t_{\alpha/2}(n-1)$

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.

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

[풀이] 주어진 문제에 대한 가설을 다음과 같다.

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

t-검정을 위해서는 t.test() 함수를 사용한다. 검정을 원하는 자료와 귀무가설에서 지정된 모수 값, 그리고 대립가설의 형태를 지정하여 검정을 수행한다. 학생들의 책임감에 대한 점수를 score 변수에 입력했다고 하면 주어진 가설에 대한 t-검정은 다음과 같이 수행할 수 있다.

```
> t.test (score, mu=40, alternative="greater")
```

- ▶ alternative에는 "greater", "less" 또는 "two.sided"를 지정할 수 있고 기본값은 "two.sided"이다.
- ▶ 유의수준을 변경하고자 할 때에는 conf.level을 사용한다. 기본값은 0.95 (유의수준 0.05)이다.

다음은 t-검정 실행 결과이다.

```
>
> t.test(score,mu=40, alternative="greater")

 One Sample t-test

data:  score
t = 1.3545, df = 14, p-value = 0.09852
alternative hypothesis: true mean is greater than 40
95 percent confidence interval:
 38.81855      Inf
sample estimates:
mean of x
 43.93333
```

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

408	405	397	405	395	415	389	403	397	390

6.2 대응 비교에 의한 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 방법

관측값 : (X_1, Y_1), $\cdots$, (X_n, Y_n)

차 : $D_i = X_i - Y_i$, $i = 1, \cdots, n$

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

① $\mu_1 - \mu_2 = \delta$의 100(1-α)% 신뢰구간 :

$$\overline{d} \pm t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}$$

② H_0 : $\mu_1 - \mu_2 = \delta_0$의 검정 : (일표본 t-검정)

$$검정통계량 : T = \frac{\overline{D} - \delta_0}{S_D/\sqrt{n}}$$

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 색별로 평균을 취한 후, 이것을 최종 자료로 선택하였다. 다음 자료를 이용하여 보라색과 초록색 사이에 감정변화에 미치는 영향이 존재하는지를 유의수준 5%에서 검정하시오. 단, 자료는 모두 정규분포 가정을 만족한다고 하자.

사람	1	2	3	4	5	6	7	8	9	10	11	12	13	14
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

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

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

대응비교 역시 t.test() 함수를 사용할 수 있다. 다만 t.test()의 옵션에서 paired 옵션을 지정해주면 된다. 주어진 데이터에 대한 대응 비교 실행은 다음과 같다.

```
> t.test(paired$purple, paired$green, paired=T)
```

▶ paired 옵션의 기본값을 False이다. 따라서 대응비교를 원하는 경우에만 True 값으로 지정해주면 된다.

▶ alternative 옵션의 기본값은 양측 검정(two.sided)이므로 별도로 지정해 줄 필요는 없다.

실행 결과는 다음과 같다.

```
> t.test(paired$purple, paired$green, paired=T)

 Paired t-test

data:  paired$purple and paired$green
t = 6.3381, df = 13, p-value = 2.584e-05
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 0.461401 0.938599
sample estimates:
mean of the differences 
                    0.7 
```

검정통계량의 값은 6.3381 이고 검정통계량은 자유도 13의 $t-$분포를 따른다. 유의확률은

0.00002584 로 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.

학생	1	2	3	4	5	6	7	8	9	10
수업 전	18	21	16	22	19	24	17	21	23	18
수업 후	22	25	17	24	16	29	20	23	19	20

학생	11	12	13	14	15	16	17	18	19	20
수업 전	14	16	16	19	18	20	12	22	15	17
수업 후	15	15	18	26	18	24	18	25	19	16

6.3 이표본에 의한 모평균의 비교

두 모집단의 분포 :

$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$을 가정

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

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

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

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

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \text{ 대략 } t(\phi) \text{ 임을 이용}$$

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

. $\mu_1 - \mu_2$**의 100(1-α)% 신뢰구간** :

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

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

검정통계량 : ① 등분산을 가정할 때 : $T = \dfrac{\overline{X}_1 - \overline{X}_2 - \delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

② 이분산을 가정할 때 : $T = \dfrac{\overline{X}_1 - \overline{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$

기각역 :

① $H_1: \mu_1 - \mu_2 > \delta_0$ 일 때 $T \ge t_\alpha(df)$

② $H_1: \mu_1 - \mu_2 < \delta_0$ 일 때 $T \le -t_\alpha(df)$

③ $H_1: \mu_1 - \mu_2 \ne \delta_0$ 일 때 $|T| \ge t_{\alpha/2}(df)$

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

6.4 두 모분산에 관한 추론

F-분포

S_1^2 : $N(\mu_1, \sigma_1^2)$에서 크기 n_1인 임의표본의 표본분산

S_2^2 : $N(\mu_2, \sigma_2^2)$에서 크기 n_2인 임의표본의 표본분산 (두 표본은 서로 독립)

$F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ ~ 자유도 (n_1-1, n_2-1)인 F-분포

기호 : $F(n_1-1, n_2-1)$

참고. F-분포의 정의 (자유도 k인 카이제곱 분포를 기호 $\chi^2(k)$로 나타냄)

$V_1 \sim \chi^2(k_1)$이고 $V_2 \sim \chi^2(k_2)$이며 V_1과 V_2는 서로 독립일 때

$$F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$$

F-분포의 성질 : $F \sim F(k_1, k_2)$일 때 $\dfrac{1}{F} \sim F(k_2, k_1)$

왼쪽 꼬리의 F_α**값** : $F_{1-\alpha}(k_1, k_2) = \dfrac{1}{F_\alpha(k_2, k_1)}$

예. $F_{0.95}(5,10) = \dfrac{1}{F_{0.05}(10,5)} = \dfrac{1}{4.74} = 0.21$

두 모분산의 비에 관한 검정

등분산성의 검정 (F-검정) :

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

검정통계량 : $F = \dfrac{S_1^2}{S_1^2} \sim F(k_1, k_2)$, k_1: 분자의 자유도, k_2: 분모의 자유도

기각역 : $F > F_{\alpha/2}(k_1,k_2)$ or $F < F_{1-\alpha/2}(k_1,k_2)$

예 1) (paint.txt) 한 페인트 제조회사에서는 새 상품의 유성페인트를 개발하여 기존의 페인트와의 건조속도를 비교하고자 한다. 이를 확인하기 위해 시중에서 가장 인기 있는 상품과 새 상품을 각각 5종류의 벽에 칠한 후 건조시간을 측정하였다. 새 상품은 기존의 상품보다 건조시간이 더 빠르다고 할 수 있는가? 유의 수준 5%에서 검정해보자.

건조시간(단위: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

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

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

독립 이표본 검정의 자료구조는 대응표본과는 다르게 그룹을 나타내는 변수(group, 1=인기상품, 2=새상품)와 검정 대상이 되는 변수 (time, 건조시간)로 구성되어 있다.

먼저 각 그룹별 건조시간의 평균을 비교해보자. 이를 위해서는 tapply() 함수를 사용한다.

```
> tapply(paint$time, paint$group, mean)
```

▶ tapply(x, group, function) : 주어진 x 자료를 group 변수별로 나누어서 지정된 function을 시행하는 함수이다.

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

$$H_0 : \sigma_1^2/\sigma_2^2 = 1 \qquad vs \qquad H_1 : \sigma_1^2/\sigma_2^2 \neq 1$$

등분산 여부 검정을 위해서는 var.test() 함수를 사용한다. 주어진 두 그룹의 분산 검정은 다음과 같이 시행가능하다.

```
> var.test(paint$time ~ paint$group)
또는
> var.test(time ~ group , data=paint)
```

▶ var.test(x~y, data) : 주어진 data에 있는 x 변수를 y 변수에 입력된 그룹에 따라 나누어서 검정을 시행한다.

```
>
> var.test(paint$time ~ paint$group)

 F test to compare two variances

data:  paint$time by paint$group
F = 1.8333, num df = 9, denom df = 9, p-value =
0.38
alternative hypothesis: true ratio of variances is not equal t
o 1
95 percent confidence interval:
 0.4553741 7.3809893
sample estimates:
ratio of variances
          1.833333
```

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.

```
> t.test(paint$time ~ paint$group, var.equal=T, alternative="greater")
또는
> t.test(time ~ group, var.equal=T, alternative="greater", data=paint)
```

▶ var.equal 옵션을 이용하여 독립 이표본 검정의 등분산 가정 여부를 선택한다. var.equal 옵션의 기본값은 False이다.
▶ 주어진 자료의 검정 대립가설의 형태에 따라 alternative 옵션 값을 지정해준다.

```
> t.test(paint$time~paint$group, var.equal=T, alternative="greater")

 Two Sample t-test

data:  paint$time by paint$group
t = 2.818, df = 18, p-value = 0.005694
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.153948      Inf
sample estimates:
mean in group 1 mean in group 2
             45              42
```

검정 결과, 검정 통계량은 2.818이고 유의확률은 0.005694로 나타났다. 이는 주어진 유의수준 0.05보다 작기 때문에 귀무가설을 기각할 수 있다. 따라서 새 페인트의 건조시간은 기존 페인

트의 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.

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

6.5 자료를 이용한 예제

예제1. (textbooks.txt) 주어진 자료는 UCLA 내의 서점과 Amazon.com 에서 판매되는 교재들의 가격에 대한 자료이다. 2010년 봄학기에 개설된 UCLA의 강의 중에서 73개를 선택하여 각 강의에 쓰이는 교재의 온라인 판매 가격(amazNew) 과 오프라인의 판매 가격(uclaNew)을 조사하였다. 교재의 판매가격은 판매 장소 (온라인 또는 오프라인)에 따라 차이가 난다고 볼 수 있는가? 적절한 가설을 세우고 유의수준 5%에서 이를 검정하시오.

예제2. (run10samp.txt) 주어진 자료는 2012년 Washington, DC에서 열렸던 Cherry Blossom 10 mile run 경기에서 완주를 한 선수 100명의 자료이다. 주요 변수에 대한 설명은 다음과 같다.

변수명	설명
time	10 마일 달리기 완주 기록 (분)
age	선수 나이
gender	성별 (M=남성, F=여성)
state	출신 지역 (또는 국가)

성별에 따른 완주시간은 차이가 있는가? 적절한 가설을 세우고 유의수준 5%에서 이를 검정하시오.