

제 6장. 분포에 관한 추론

모평균에 관한 추론 : 모분산을 모르는 경우

- ▶ 정규모집단 $N(\mu, \sigma^2)$ 에서 크기 n 인 랜덤 표본이 선택되었을 때, 모분산을 모르는 경우 스튜던트화 된 표본평균의 분포

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

- ▶ 모분산을 모르는 경우, 모평균의 $100(1-\alpha)\%$ 신뢰구간

$$\left(\bar{x} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right)$$

: 단, 표본크기가 큰 경우에는 정규모집단의 가정이 완화될 수 있으며, t 분포 대신 표준정규분포가 사용될 수 있다.

- ▶ 모분산을 모르는 경우, 모평균의 유의성 검정

: 귀무가설이 $H_0: \mu = \mu_0$ 인 경우, 검정통계량은 다음과 같다

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t(n-1)$$

모평균에 관한 추론 : 모분산을 모르는 경우

- ▶ 예제 6.1 : 두 물리학자는 일정한 거리만큼 떨어진 곳의 거울을 이용하여, 빛이 7400m의 거리를 움직이는 시간을 측정하였다. 그 결과 64개의 자료로부터 표본평균은 27.750, 표본표준편차는 5.083임을 얻었다. 이를 이용하여 모평균에 대한 99% 신뢰구간을 구하시오

- ▶ 예제 6.2 : 전구를 생산하는 한 회사에서 현재 생산하는 전구의 평균수명은 1950시간으로 알려져 있다. 새로 개발중인 전구의 평균 수명이 기존보다 더 길다고 할 수 있는지를 확인하기 위해 9개의 시제품을 생산하여 조사한 결과가 아래와 같다.

2000 1975 1900 2000 1950 1850 1950 2100 1975

수명의 분포가 정규분포라는 가정하에 유의수준 5%에서 가설검정을 실시하시오.

모평균에 관한 추론 : 모분산을 모르는 경우

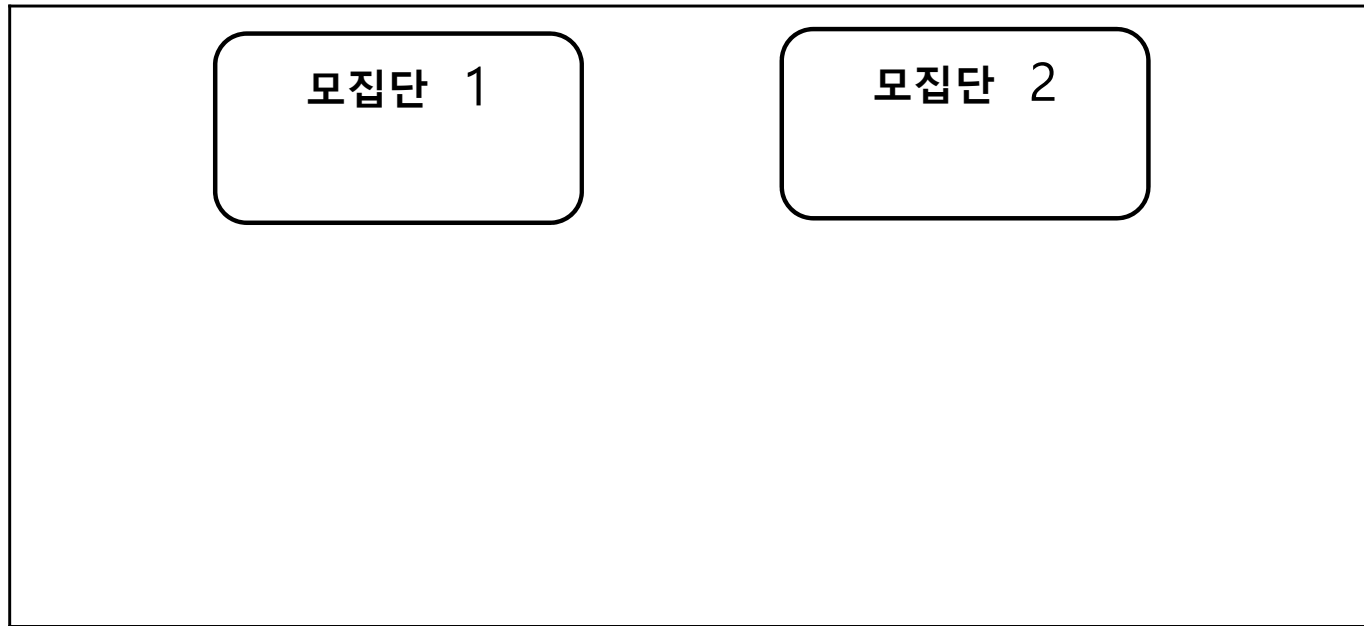
- ▶ 예제 6.3 : 반도체의 전기적 특성은 그 제조과정에서 첨가되는 불순물의 양에 따라 정해진다고 한다. 특정한 용도에 사용되는 실리콘 다이오드는 0.60볼트의 가동전압이 요구되며 이러한 목표에서 벗어나면 불순물의 양을 조절하는 장치에 대한 조정이 필요하다고 한다. 120개의 랜덤 표본을 조사한 결과 평균이 0.62, 표준편차가 0.11로 조사되었다. 유의수준 5%에서 가설을 검정하여라.

두 모집단에 관한 추론

- ▶ 두 모집단의 평균 비교의 예
 - ▶ 사무직과 생산직의 임금 비교
 - ▶ 수도권과 지방 고등학생들 간의 학력 비교
 - ▶ 두 치료제의 효능 비교
 - ▶ 두 가지 방법에 따른 생산성 비교
- ▶ 두 모집단의 모비율 비교의 예
 - ▶ 두 생산공장의 불량률 비교
 - ▶ 남녀의 성별에 따른 특정 정당 지지율 비교
- ▶ 두 모집단의 모분산 비교의 예
 - ▶ 두 개의 포트폴리오의 변동성 비교

두 모평균에 관한 추론

- ▶ 예 : 두 가지 교육방법의 효과를 비교하는 경우



- ▶ 관심모수 :
- ▶ 이표본에 의한 비교에서 자료의 구조
 - ▶ 각 그룹에서의 관측값들은 각 모집단에서의 랜덤 표본이다
 - ▶ 서로 다른 그룹에서의 관측값들은 독립적으로 관측된 것이다

$(\bar{X}_1 - \bar{X}_2)$ 의 표본분포

- ▶ $(\bar{X}_1 - \bar{X}_2)$ 의 기대값과 분산

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) - 2Cov(\bar{X}_1, \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- ▶ $(\bar{X}_1 - \bar{X}_2)$ 의 표본분포
: 정규성과 독립성 가정 하에

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- ▶ 두 모집단의 평균의 차이에 대한 추론은 다음과 같은 경우로 나누어진다.
 - I. 모분산을 아는 경우
 - II. 모분산을 모르는 경우
 - II-1. [이분산 가정]
 - II-2. [등분산 가정]

이표본 문제 : 모분산을 아는 경우

- ▶ 검정 통계량

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

- ▶ 모평균 차이에 대한 $(1-\alpha) \times 100\%$ 신뢰구간

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

- ▶ 실제로는 모분산을 아는 경우가 드물기 때문에 Z-검정을 하는 경우는 거의 없음

이표본 문제 : 이분산 가정 (표본의 크기가 크지않은 경우)

- ▶ 두 모분산이 알려져 있지 않은 경우
: 각각의 표본을 이용하여 모분산을 각각 추정하게 된다

$$\hat{\sigma}_1^2 = S_1^2, \quad \hat{\sigma}_2^2 = S_2^2$$

- ▶ $H_0 : \mu_1 - \mu_2 = \delta_0$ 의 검정통계량
: 표본의 크기가 크지 않은 경우, 정규모집단 가정 하에서

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(df), \quad df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

: 자유도가 자연수가 아닌 경우에는 가까운 값을 이용하거나 보간법을 사용한다

- ▶ 모평균 차이에 대한 $(1-\alpha) \times 100\%$ 신뢰구간

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}(df) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}(df) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

이표본 문제 : 이분산 가정 (표본의 크기가 크지않은 경우)

- ▶ 예제 6.7 : 음식물을 통한 질산 칼륨의 과다 섭취는 갑상선 호르몬의 감소 또는 피부의 흑청색화 등 유해한 영향을 끼친다고 한다. 이러한 질산 칼륨의 과다 섭취가 성장을 저해하는 증거가 있는지를 알아보기 위하여 16마리의 쥐를 대상으로 실험하였다. 이들 중 9마리를 랜덤추출하여 2000ppm의 질산칼륨을 섭취하게 하고, 나머지 7마리는 일상적인 식사를 하게 하였다. 일정기간 후의 이들의 체중 증가율을 조사한 결과가 다음의 표와 같다. 질산칼륨의 과다섭취가 성장을 저해하는 증거가 있는가를 유의수준 5%에서 검정하고, 유의확률도 구하여라. 단, 모집단은 정규분포를 따르고 두 모분산은 서로 다르다고 가정하자.

	질산칼륨 섭취군	규정식 섭취군
표본크기	9	7
평균	15.07	19.27
표준편차	3.56	8.05

이표본 문제 : 이분산 가정 (표본의 크기가 큰 경우)

- ▶ $H_0 : \mu_1 - \mu_2 = \delta_0$ 의 검정통계량 : n_1, n_2 가 충분히 큰 경우,

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

- ▶ 모평균 차이에 대한 $(1-\alpha) \times 100\%$ 신뢰구간

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

이표본 문제 : 이분산 가정 (표본의 크기가 큰 경우)

- ▶ 예제 6.6 : 지역 환경에 따라 학력에 차이가 있는가를 알아보기 위하여, 두 도시의 중학교 1학년 학생 중에서 각각 90명과 100명을 랜덤추출하여 동일한 시험을 시행한 결과가 다음과 같았다.

	도시 1	도시 2
표본크기	90	100
평균	76.4	81.2
표준편차	8.2	7.6

두 도시의 중학교 1학년 학생 전체의 평균성적에 차이가 있는지를 유의수준 1%에서 검정하고 유의확률을 구하여라

이표본 문제 : 등분산 가정

- ▶ 두 모분산은 미지이나 같다고 가정하는 경우 ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)
 - : 추가 전제조건하에서 더욱 효율적인 추론이 가능해짐
 - : 공통 모분산을 가정한 경우이므로, 공통 분산 추정을 위해 **합동표본분산**(pooled sample variance)를 이용한다

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

- ▶ $H_0 : \mu_1 - \mu_2 = \delta_0$ 의 검정통계량

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

- ▶ 모평균 차이에 대한 $(1 - \alpha) \times 100\%$ 신뢰구간

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}(n_1 + n_2 - 2)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}(n_1 + n_2 - 2)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

이표본 문제 : 등분산 가정

- ▶ 예제 : 비타민 B_1 이 버섯의 성장을 촉진시킨다는 가설을 확인하기 위하여 성장상태가 비슷한 20개의 버섯을 임의 추출하였다. 이 가운데 10개를 다시 임의로 뽑아 처리그룹으로 분류한 뒤 비타민을 투여하고 나머지 10개는 대조그룹으로 분류하여 비타민을 투여하지 않았다. 일정기간이 지난 후에 버섯의 무게를 측정한 결과 다음의 자료를 얻었다. 이를 이용하여 비타민 B_1 이 버섯의 성장을 촉진한다고 할 수 있는가? 유의수준 5%에서 이를 검정하시오. (단, 모집단은 정규분포를 따르고 등분산임을 가정하자.)

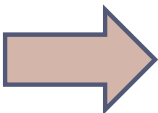
처리그룹	12	18	16	22	23	19	20	22	25	20
대조그룹	15	17	19	16	13	19	18	16	12	18

대응 비교

- ▶ 두 모평균 비교의 예
 - ▶ 새로 개발한 신약은 고혈압 환자의 혈압을 낮추는 데 효과가 있는가?
 - ▶ 두 가지 다른 소재로 만들어진 신발 중 내구성이 더 뛰어난 신발은?
 - ▶ 새로 도입한 생산 공정이 기존의 방법보다 더 효과적이라고 할 수 있는가?
- ▶ 이러한 경우, 독립적인 두 집단에 대해 실험을 실시하게 된다면 우리의 관심분야 외의 다른 요인들이 실험의 결과에 영향을 미치게 되어 올바른 결론을 얻을 수 없게 된다
- ▶ 따라서 이와 같은 경우에는 조건이 비슷한 표본들을 하나의 쌍(pair)으로 묶어서 각 쌍에 대하여 실험을 실시하는 것이 더 효과적인 실험이 될 수 있다
- ▶ 이와 같은 실험에서 각 쌍은 조건이 비슷하므로 X와 Y는 독립이 아니며, 각 쌍들 사이에서는 독립이 유지된다
- ▶ 이와 같이 쌍으로 주어진 자료를 이용하여 두 모평균을 비교하는 방법을 **대응비교 (paired comparison)** 또는 **쌍체비교**라고 한다.

대응 비교

- ▶ 예 6.1 : 두 종류의 진통제에 대한 상대적 효과의 척도로서 복용 후 숙면할 수 있는 정도를 비교하려고 한다. 이러한 실험에 참여하기로 한 환자 중에서 소수의 환자를 랜덤 추출하여 조사하기로 하였으나, 이들 환자들의 제반 건강상태에 상당한 차이가 있음을 알고 있다. 따라서 이들 중 6명의 환자를 랜덤 추출하여 각 환자에게 두 종류의 진통제를 각각 1회씩 복용하게 하여 숙면시간의 차이를 이용하여 두 진통제의 효과를 비교하기로 하였다.
- ▶ 진통제 A와 진통제 B에 따른 숙면 시간은 서로 독립이 아니기 때문에 기존의 모평균 비교 방법을 사용할 수 없음 : 추정량의 표본분포가 달라졌기 때문
- ▶ 따라서 새로운 변수 $D(= \text{진통제 A} - \text{진통제 B})$ 를 정의한 후, D 의 모평균이 0인지를 검정하는 문제로 바꾸어서 생각하기로 함 (일표본 방법을 적용)

GROUP 1	GROUP 2		D
X_1	Y_1	$D_i = X_i - Y_i$ 	D_1
X_2	Y_2		D_2
X_3	Y_3		D_3
X_4	Y_4		D_4
X_5	Y_5		D_5
....
X_n	Y_n		D_n

대응 비교

- ▶ $\{D_1, D_2, \dots, D_n\}$ 에 대하여 다음과 같이 정의

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i, \quad S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$$

- ▶ \bar{D} 의 표본 분포 : 차이의 모집단(population of D)이 $N(\mu_D, \sigma_D^2)$ 의 분포를 따를 때,

$$\bar{D} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

- ▶ $H_0 : \mu_D = \delta_0$ 의 검정통계량

$$T = \frac{\bar{D} - \delta_0}{S_D / \sqrt{n}} \sim t(n-1)$$

- ▶ 모평균의 차이에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left[\bar{d} - t_{\alpha/2}(n-1) \frac{s_D}{\sqrt{n}}, \quad \bar{d} + t_{\alpha/2}(n-1) \frac{s_D}{\sqrt{n}} \right]$$

대응 비교

- ▶ 예제 6.4 : 아래의 조사 결과를 이용하여 두 진통제 A,B에 의한 평균 숙면시간의 차이가 존재하는지 유의수준 5%에서 이를 검정하시오. 검정을 위해 필요한 가정사항은 무엇인가?

환자	1	2	3	4	5	6
진통제 A	4.8	4.0	5.8	4.9	5.3	7.4
진통제 B	4.0	4.2	5.2	4.9	5.6	7.1
차이						

모분산에 관한 추론

- ▶ 모집단의 변동성 또는 퍼짐의 정도에 관심이 있는 경우, 모분산이 추론의 대상이 됨
- ▶ 예 : 일정 규격의 볼 베어링을 생산하는 경우, 제조 공정에서 지름의 변동성이 크면 평균적으로 알맞아도 불량품이 많아지는 문제가 발생한다. 따라서 베어링 지름의 변동성을 확인하여 허용범위를 넘지 않도록 생산공정을 조절해야 한다. 이때 베어링 지름의 변동성을 확인하는 방법으로 모분산에 대한 추론을 할 수 있다.
- ▶ 예 : 주식 투자하는 사람들은 주식 가격의 변동성(volatility)에 주목한다. 가격변동이 심한 주식은 위험(risk)이 높다.
- ▶ 예: Y 식품은 캔음료 생산 공장에서 350ml 캔음료를 생산하고 있다. 표기된 용량보다 적다는 소비자의 불만이 접수되었다. 평균 350ml가 주입되지만 모든 캔 음료가 정확히 350ml는 아니고, 만약 모분산이 일정 기준 이상으로 크다면 과소 또는 과다용량 제품이 많이 나타나게 될 것이다. 따라서 모분산에 대한 검정을 통해 생산공정을 조정할 수 있을 것이다.

모분산에 대한 추론

- ▶ 가정사항
: 모집단은 정규분포를 따른다.

- ▶ 관심 모수 : 모분산 (σ^2)
- ▶ 관심 모수의 추정량 : 표본 분산

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

- ▶ 표본분산의 표본분포

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

: chi-squared distribution with (n-1) degrees of freedom

모분산에 대한 추론

- ▶ 모분산에 대한 $(1-\alpha)\times 100\%$ 신뢰구간 : 정규모집단 가정 하에서,

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-(\alpha/2)}^2(n-1)}$$

- ▶ $H_0: \sigma^2 = \sigma_0^2$ 의 검정통계량

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

: 평균의 검정과는 달리 정규모집단 가정에 깊게 의존하고 있으므로 이러한 가정에 대한 검토가 적절하게 이루어져야 한다

모분산에 대한 추론

- ▶ 예제 6.9 : 플라스틱판을 생산하는 한 공장에서는 생산되는 판 두께의 표준편차가 1.5mm를 상회하면 공정에 이상이 있는것으로 간주한다. 어느날 점검에서 10개의 판을 랜덤추출하여 그 두께를 측정한 결과가 다음과 같을 때(단위:mm), 공정에 이상이 있는지를 유의수준 5%에서 검정하고 유의확률을 구하여라. 단, 과거의 기록에 의하면 판 두께의 분포는 정규분포라고 해도 무방하다.

226 228 226 225 232 228 227 229 225 230

두 모분산에 대한 추론

- ▶ 두 모집단의 분산 비교의 예
 - ▶ 서로 다른 투자 포트폴리오에 관련된 위험 비교
 - ▶ 두 모평균의 비교 시 등분산 가정이 적합한지에 대한 분석
- ▶ 가정사항
: 두 모집단은 정규분포를 따른다
- ▶ 관심모수와 추정량 : $\frac{\sigma_1^2}{\sigma_2^2} \Rightarrow \frac{S_1^2}{S_2^2}$
- ▶ 추정량의 표집분포

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

두 모분산에 대한 추론

- ▶ 두 모분산의 비에 대한 $(1-\alpha)100\%$ 신뢰구간 : 정규모집단의 경우,

$$\frac{s_1^2}{s_2^2} \frac{1}{F_{\alpha/2}(n_1-1, n_2-1)} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} F_{\alpha/2}(n_2-1, n_1-1)$$

- ▶ $H_0: \sigma_1^2 / \sigma_2^2 = 1$ 의 검정통계량

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

두 모분산에 대한 추론

- ▶ 예제 6.10 : 콘크리트에 균열이 있는 경우, 이의 보수를 위하여 흔히 중합물질인 폴리머를 주입하게 된다. 두 폴리머의 주입 압축률의 산포가 다르다는 증거가 있는가를 유의수준 5%에서 검정하여라. 단, 모집단의 분포는 모두 정규분포를 따른다고 가정한다.

Epoxy	1.75	2.12	2.05	1.97
MMA Prepolymer	1.77	1.59	1.70	1.69