6장. 두 모집단의 비교에 관한 추론

박명현

서울대학교 학부대학

두 모집단의 비교에 관한 추론

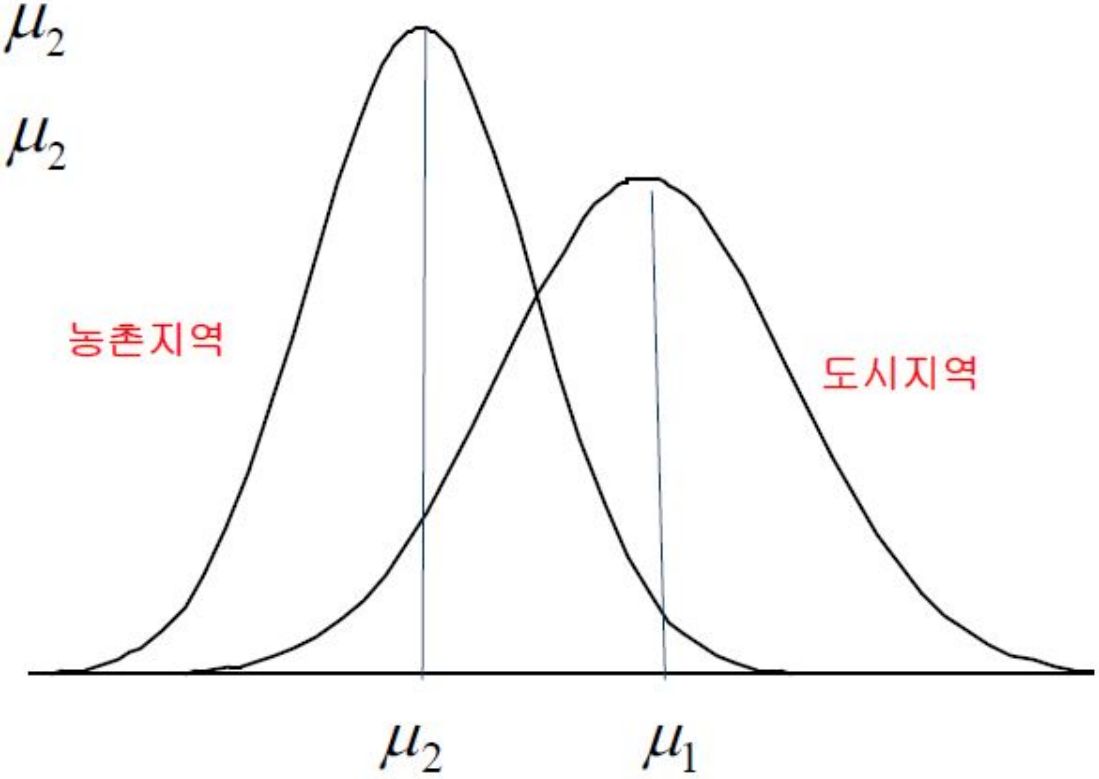
- 두 모집단의 평균에 대한 추론
  - 대응비교에 의한 모평균의 비교
- 두 모집단의 비율에 대한 추론
- 두 모집단의 분산에 대한 추론
  - F분포

두 모집단의 모평균의 비교

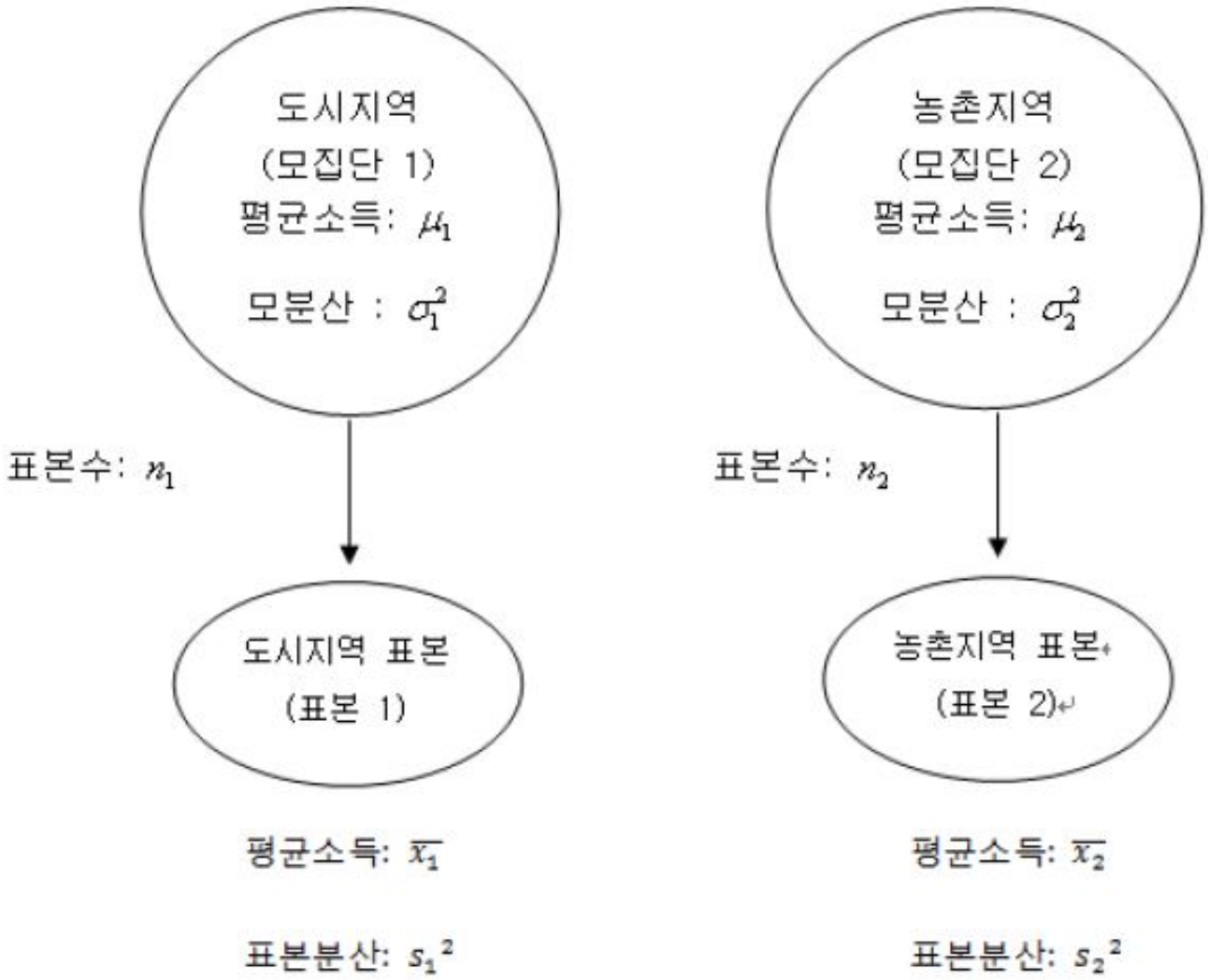
예) 국가에서 국민의 삶의 질을 향상시키기 위한 정책을 내놓기 위해 현재 계층 간, 지역 간 삶의 질의 차이에 대한 관찰을 하고자 함.

- 도시/농촌지역의 소득에 차이가 있는지 평균비교를 할 수 있다.

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$



두 모집단의 모평균의 비교



- 표본1: $\{X_{11}, \cdots, X_{1n_1}\}$, 표본2: $\{X_{21}, \cdots, X_{2n_2}\}$

두 모집단의 모평균의 비교

- 두 모집단의 평균 비교에 필요한 가정

  (1) 각 그룹에서의 관측값들은 각 모집단에서의 랜덤표본이다.

  (2) 서로 다른 그룹에서의 관측값들은 독립적으로 관측된 것이다.

  (3) 두 모집단은 각각 정규분포를 따른다.

  : 정규분포를 따르지 않더라도 표본이 클 경우 CLT 적용 가능.

- 서로 다른 두 모집단에서 각각 랜덤추출된 두 표본 => 위의 가정사항을 만족
- 주어진 실험단위를 대상으로 두 가지의 처리를 적용하여 비교하는 경우 => 전체 실험단위를 랜덤하게 두 그룹으로 나누어 각 그룹에 별로의 처리를 적용해야 위의 가정사항을 만족할 수 있다.
- 랜덤화(randomization): 각 처리를 적용할 실험단위를 랜덤하게 정하는 과정

$\bar{X}_1 - \bar{X}_2$의 표본분포

- $\mu_1 - \mu_2$를 표본평균의 차인 $\bar{X}_1 - \bar{X}_2$로 추정한다.

$$\widehat{\mu_1 - \mu_2} = \bar{x}_1 - \bar{x}_2$$

- $\bar{X}_1 - \bar{X}_2$의 기대값

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

- $\bar{X}_1 - \bar{X}_2$의 분산과 표준오차

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

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

$\bar{X}_1 - \bar{X}_2$의 표본분포

- 정규 모집단이면 $(\bar{X}_1 - \bar{X}_2)$도 정규분포를 따르고,
  비정규 모집단이더라도 표본수가 크면 근사적으로 $(\bar{X}_1 - \bar{X}_2)$는 정규분포에 가까워진다.(by, CLT)
- 따라서 $\bar{X}_1 - \bar{X}_2$의 표본분포는 (근사적으로) 다음과 같고,

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

이를 표준화 하면 다음과 같이 (근사적으로) 표준정규분포를 따르게 된다.

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

두 모평균의 차이에 대한 추론(σ_1^2, σ_2^2를 아는 경우)

I. σ_1^2, σ_2^2를 아는 경우

1. 두 모집단의 모분산 값이 알려진 경우
2. 정규분포 모집단 가정. 아닌 경우에도 대표본 근사 가능한 경우

- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$신뢰구간:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

두 모평균의 차이에 대한 추론(σ_1^2, σ_2^2를 아는 경우)

- 귀무가설: $H_0 : \mu_1 = \mu_2 + a$ or $H_0 : \mu_1 - \mu_2 = a$
- 검정통계량:

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

- 검정통계량의 관측값을 z_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert Z\rvert \geq \lvert z\rvert)$	$\lvert z_0\rvert \geq z_{\alpha/2}$

- - p값에 의한 판정: $P \leq \alpha$이면, H_0기각
  - 기각역에 의한 판정: 위의 기각역 조건이면 H_0기각

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

II. σ_1^2, σ_2^2를 모르는 경우

1. 두 모분산이 같다고 가정할 때: $\sigma_1^2 = \sigma_2^2 = \sigma^2$(등분산 가정)
2. 정규분포 모집단 가정

- 합동표본분산(pooled sample variance): S_p^2

  S_p^2은 공통 모분산 σ^2의 추정량으로, S_1^2, S_2^2의 자유도인 $n_1 - 1$과 $n_2 - 1$의 가중평균의 형태로서 다음과 같이 정의된다.

$$
\begin{aligned}
S_p^2 &= \frac{(n_1 - 1)}{(n_1 - 1) + (n_2 - 1)} S_1^2 + \frac{(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} S_2^2 \\
&= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}
\end{aligned}
$$

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

- $\bar{X}_1 - \bar{X}_2$의 분산 추정

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

$$\widehat{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} = \hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2}) = S_p^2(\frac{1}{n_1} + \frac{1}{n_2})$$

- $\bar{X}_1 - \bar{X}_2$를 추정된 분산으로 표준화한 통계량은 자유도 $(df = n_1 + n_2 - 2)$인 t분포를 따르게 됨(증명)

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

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

- 등분산 정규모집단 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$에서의 t분포 증명

- $[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)] / \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \sim N(0,1)$,
  $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$이며, 이들은 서로 독립이다.(p.72)

- 따라서, t분포의 정의로부터 다음이 성립하고

$$\frac{[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)] / \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}{\sqrt{\frac{(n_1+n_2-2)S_p^2}{\sigma^2} / (n_1 + n_2 - 2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- 자유도가 $n_1 + n_2 - 2$인 다음의 t분포를 따른다고 할 수 있다.

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

두 모평균의 차이에 대한 추론(등분산, $\sigma_1^2 = \sigma_2^2$)

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

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

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

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

- 검정통계량의 관측값을 t_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(T \geq t)$	$t_0 \geq t_\alpha(n_1 + n_2 - 2)$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(T \leq t)$	$t_0 \leq -t_\alpha(n_1 + n_2 - 2)$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert T \rvert \geq \lvert t \rvert)$	$\lvert t_0 \rvert \geq t_{\alpha/2}(n_1 + n_2 - 2)$

두 모평균의 차이에 대한 추론 예제

Example

한 의학 연구에서 비만 치료를 위한 GLP-1(Glucagon-Like Peptide-1) 작용제인 세마글루타이드(semaglutide)의 체중 감소 효과를 다음과 같이 분석했다. 건강 상태가 비슷한 지원자 12명을 무작위로 6명씩 두 그룹으로 나누고 한 그룹에는 세마글루타이드를, 다른 그룹에는 위약을 투여했다. 각 그룹은 12주 동안 주 1회 피하주사를 맞았으며, 식습관 운동 등 생활습관은 비슷하게 유지하도록 통제했다. 12주가 지난 시점에서 각 지원자의 체중 감소량을 측정했고, 결과는 다음 표와 같다.

(단위 : kg)

세마글루타이드군	5.12	4.87	5.64	5.01	5.49	5.25
위약군	3.62	3.98	4.14	3.25	4.42	3.60

두 집단의 평균 체중 감소량에 차이가 있는지를 유의수준 5%에서 검정하고, 평균 체중 감소량 대한 95% 신뢰구간을 구하라. 단, 두 집단의 분산은 같고 정규모집단을 가정한다.

sol) 두 집단의 평균 체중 감소량을 각각 μ_1, μ_2라 하면, 검정하려는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

- 주어진 자료에서 평균과 표준편차를 구하면 다음과 같다.

	세마글루타이드군	위약군
표본크기(n_i)	6	6
평균($\bar{x}_i$)	5.23	3.84
표준편차(si)	0.29	0.42

- 이를 이용하여 합동표본분산을 구하면

$$S_p^2 = \frac{(6-1)0.29^2 + (6-1)0.42^2}{6+6-2} = 0.13025, \quad S_p = 0.361$$

- 따라서 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(5.32 - 3.84) - 0}{0.361\sqrt{\frac{1}{6} + \frac{1}{6}}} = 6.67$$

- t분포표에서 $t_{0.025}(10) = 2.228$
- 기각역으로 판정: 기각역은 $|t| \geq t_{0.025}(10) = 2.228$이고, $|t_0| > 2.228$이므로 유의수준 5%에서 귀무가설을 기각할 수 있다.
- p값으로 판정: $P = P(|t| \geq 6.67)$이고 t분포표에서 $P < 0.001$임을 알 수 있다.
- 지금까지의 내용을 종합하면 세마글루타이드와 위약군 간의 평균 체중 감소량이 다르다는 증거가 충분하다.
- $\mu_1 - \mu_2$에 대한 95%신뢰구간을 구하면

$$
\begin{aligned}
&(\bar{x}_1 - \bar{x}_2) \pm t_{0.025}(10) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= (5.23 - 3.84) \pm t_{0.025}(10) 0.361 \sqrt{\frac{1}{6} + \frac{1}{6}} \\
&= 1.39 \pm 0.46 = (0.93, 1.85)
\end{aligned}
$$

두 모평균의 차이에 대한 추론(등분산, 대표본의 경우)

II. σ_1^2, σ_2^2를 모르는 경우

1. $\sigma_1^2 = \sigma_2^2 = \sigma^2$(등분산 가정)
2. but, 모집단의 분포가 정규분포가 아닌 경우

- 모집단의 분포에 관계없이 대표본의 경우 다음의 통계량은 중심극한정리에 따라 표준정규분포를 따르게 된다.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \dot{\sim} N(0,1)$$

- 대표본의 조건: $n_1 \geq 30, n_2 \geq 30$

두 모평균의 차이에 대한 추론(등분산, 대표본의 경우)

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

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

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

- 검정통계량의 관측값을 z_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert Z \rvert \geq \lvert z \rvert)$	$\lvert z_0 \rvert \geq z_{\alpha/2}$

두 모평균의 차이에 대한 추론(이분산, $\sigma_1^2 \neq \sigma_2^2$)

II. σ_1^2, σ_2^2를 모르는 경우

1. 두 모분산이 다르다고 가정할 때: $\sigma_1^2 \neq \sigma_2^2$(이분산 가정)
2. 정규분포 모집단 가정

- $\bar{X}_1 - \bar{X}_2$의 분산: 이분산의 경우 두 표본평균간의 독립성에 의해

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

- 분산의 추정치:

$$\widehat{Var}(\bar{X}_1 - \bar{X}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

- 모집단 정규분포이고 표본크기 n_1, n_2가 5 이상이면 $\bar{X}_1 - \bar{X}_2$를 추정된 분산으로 표준화한 통계량은 자유도를 적절히 선택함으로써 t분포에 가까운 것이 알려져 있다.

두 모평균의 차이에 대한 추론(이분산, $\sigma_1^2 \neq \sigma_2^2$)

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} t(df), \qquad df = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

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

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(df)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- 검정통계량의 관측값을 t_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(T \geq t)$	$t_0 \geq t_\alpha(df)$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(T \leq t)$	$t_0 \leq -t_\alpha(df)$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t_0\rvert \geq t_{\alpha/2}(df)$

두 모평균의 차이에 대한 추론 예제

Example

최근에는 하루 만에 배송이 가능한 서비스가 활성화되면서, 식료품을 온라인으로 구매하는 가정이 증가하고 있다. 한 경제심리학자는 소비자가 오프라인 매장에서 식료품을 구매할 때의 지출이 온라인 구매할 때의 지출보다 적다고 주장한다. 이를 확인하기 위해 무작위로 선정한 18명과 20명에게 각각 오프라인과 온라인에서 4인 가구의 2주일 치 식량을 쇼핑하도록 하고, 지출액을 조사한 결과는 다음과 같다.

(단위 : 천 원)

오프라인	169	206	257	294	252	283	240	207	230	183
	298	269	256	277	300	126	318	184		
온라인	289	334	278	268	336	438	388	388	394	394
	425	386	356	342	305	365	355	312	209	458

오프라인 쇼핑의 지출액이 온라인 쇼핑보다 적다는 증거가 있는지 유의수준 5%에서 검정하라.

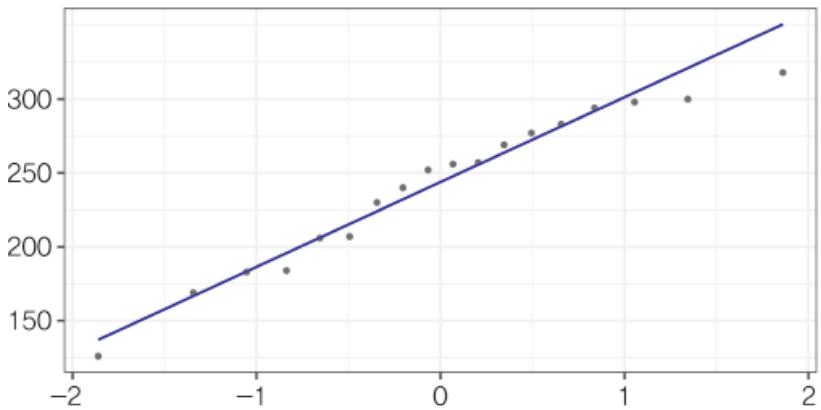
sol) 오프라인 쇼핑의 평균 지출액을 μ_1, 온라인 쇼핑의 평균 지출액을 μ_2라고 하면 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 < \mu_2$$

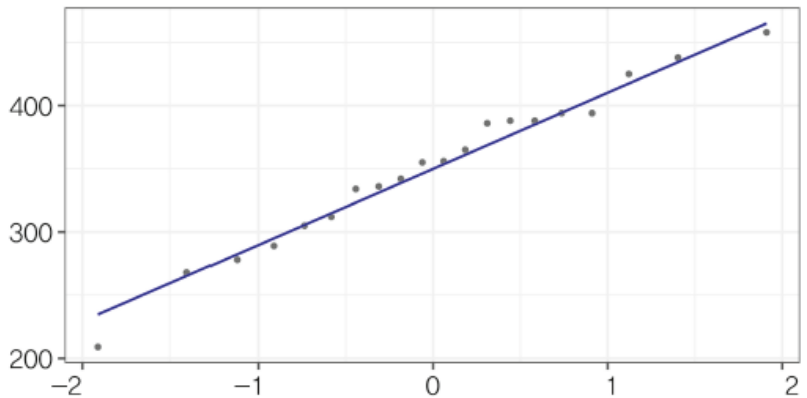
- 주어진 자료에서 평균과 표준편차를 구하면 다음과 같다.

	오프라인	온라인
표본크기(n_i)	18	20
평균($\bar{x}_i$)	241.6	351.0
표준편차(si)	52.82	61.92

- 두 집단의 정규분포 분위수대조도를 그려보면 다음과 같이 전반적으로 정규모집단을 전제로 하는 방법의 적용에 큰 무리가 없는 것으로 판단된다.



(a) 오프라인 쇼핑



(b) 온라인 쇼핑

- 따라서 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(241.6 - 351.0) - 0}{\sqrt{\frac{52.82^2}{18} + \frac{61.92^2}{20}}} = -5.875$$

- 자유도를 계산하면 다음과 같다.

$$df = \frac{(52.82^2/18 + 61.92^2/20)^2}{\frac{(52.82^2/18)^2}{17} + \frac{(61.92^2/20)^2}{19}} = 35.91$$

- $t_{0.05}(35.91)$의 값을 t분포표에서 찾아보면

$$t_{0.05}(30) = 1.697 > t_{0.05}(35.91) > t_{0.05}(40) = 1.684$$

이므로 보간법을 이용하면 다음과 같다.

$$t_{0.05}(35.91) = 1.684 + (1.697 - 1.684) \times 0.409 = 1.689$$

- 실제 문제에서 자유도 df는 t분포표에서 df에 제일 가까운 자연수의 자유도를 이용한다. (위의 경우 $t_{0.05}(40) = 1.684$이용)

- 기각역으로 판정:

  기각역은 $t_0 \leq -t_{0.05}(35.91)$이고, $t_0 = -5.875 < -1.684$이므로 유의수준 5%에서 귀무가설을 기각할 수 있다.

  * 통계 소프트웨어를 이용해서 계산하면 $t_{0.05}(35.91) = 1.68841$

- p값으로 판정:

  $P = P(T \leq -5.875)$이고 t 분포표에서 $P < 0.0005$이므로 귀무가설을 기각할 수 있다.

- 이상을 종합하면, 유의수준 5%에서 오프라인 매장에서 식료품을 구매할 때 온라인 구매보다 평균 지출액이 적다는 증거가 뚜렷하다.

두 모평균의 차이에 대한 추론(이분산, 대표본의 경우)

II. σ_1^2, σ_2^2를 모르는 경우

1. $\sigma_1^2 \neq \sigma_2^2$(이분산 가정)
2. but, 모집단의 분포가 정규분포가 아닌 경우

- 모집단의 분포에 상관없이 n_1, n_2가 충분히 크면 중심극한정리에 의해 대략적으로 표준정규분포를 따르게 된다.

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$$

- 대표본의 조건: $n_1 \geq 30, n_2 \geq 30$

두 모평균의 차이에 대한 추론(이분산, 대표본의 경우)

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

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- 귀무가설이 $H_0 : \mu_1 - \mu_2 = a$인 경우에 검정통계량은

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$$

- 검정통계량의 관측값을 z_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_1 - \mu_2 > a$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : \mu_1 - \mu_2 < a$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : \mu_1 - \mu_2 \neq a$	$P = P(\lvert Z\rvert \geq \lvert z\rvert)$	$\lvert z_0\rvert \geq z_{\alpha/2}$

두 모평균의 차이에 대한 추론 예제

Example

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

	도시 1	도시 2
표본크기(n_i)	90	100
평균($\bar{x}_i$)	76.4	81.2
표준편차(s_i)	10.2	7.6

두 도시의 중학교 1학년 학생 전체의 평균 성적에 차이가 있는가 유의수준 5%에서 검정하고, 유의확률도 구하여라. 또한, 평균성적에 차이에 대한 신뢰수준 95%의 신뢰구간도 구하여라.

sol) 도시 1, 2의 중학교 1학년 학생 전체의 평균성적을 각각 μ_1, μ_2라고 하자.

- 검정하려는 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

- 조사한 표본의 크기가 각각 $n_1 = 90$, $n_2 = 100$으로 충분히 크므로, 이 경우에 검정통계량의 관측값을 계산하면

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(76.4 - 81.2) - 0}{\sqrt{\frac{10.2^2}{90} + \frac{7.6^2}{100}}} = -3.65$$

- $z_{0.025} = 1.96$이므로, $\mu_1 - \mu_2$에 대한 95%신뢰구간은 다음과 같다.

$$(\bar{x}_1 - \bar{x}_2) \pm z_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (76.4 - 81.2) \pm 1.96\sqrt{\frac{10.2^2}{90} + \frac{7.6^2}{100}}$$

$$= (-7.38, -2.22)$$

- 기각역으로 판정:

  기각역은 $|z_0| \geq 1.96$이고, $|z_0| = 3.65 \geq 1.96$이므로 기각역에 속하여 유의수준 5%에서 귀무가설을 기각할 수 있다.

- p값으로 판정:

  유의확률은 $P = P(|Z| \geq)$이고 표준정규분포표에서 그 값을 찾아보면

$$\begin{aligned} P(|Z| \geq 3.65) &= P(Z < -3.65) + P(Z > 3.65) \\ &= 0.00013 + 0.00013 = 0.00026 \end{aligned}$$

- P값은 $\alpha = 0.05$보다 작아 귀무가설을 기각한다.
- 이상을 종합하면, 두 도시의 중학교 1학년 학생의 학력에는 차이가 있다는 뚜렷한 증거가 있다.

대응비교에 의한 모평균의 비교

예) 다이어트하기 전과 다이어트 한 달 후의 몸무게처럼 동일한 개체에 대해 실험 전과 실험 후의 측정값의 차이를 비교하는 경우

- 앞 절에서 다룬 바와 같이 측정 결과를 두 모집단의 표본으로 간주하고 평균 비교를 할 수도 있지만 대응비교(paired comparison)를 이용하면 더욱 정확한 실험 효과를 측정할 수 있다.

- 대응비교: 두 모집단의 평균을 비교할 때 동질적인 비교 대상들로 쌍을 이루어 각 쌍내에서의 차를 이용하여 비교하는 방법

대응비교에 의한 모평균의 비교

예) 두 종류의 진통제가 수면에 미치는 효과를 비교하고자 한다. 실험에 참여하기로 한 환자들 중 6명의 환자를 랜덤추출하고 각 환자에게 두 종류의 진통제를 랜덤한 순서로 각각 1회씩 복용하게 하여 숙면시간에 차이가 있는지 조사하였다.

환자 1: (진통제 A, 진통제 B)

환자 2: (진통제 A, 진통제 B)

... ...

환자 6: (진통제 A, 진통제 B)

- 실험에 참여하기로 한 환자들의 건강상태에 상당한 차이가 있다.

  이런경우 두 집단이 독립이라는 가정을 만족 못하므로, 환자들의 이질성을 감안하여 대응비교를 하고자 한다.

  => 두 집단의 측정값의 차이를 이용하여 추론하면 된다.

대응비교에 의한 모평균의 비교

- 주어진 자료: $(X_1, Y_1), \cdots, (X_n, Y_n)$
- 처리효과를 나타내는 모평균 $\mu_1 = E(X_i)$, $\mu_2 = E(Y_i)$를 비교하려면, 자료에서 얻은 측정값의 차이인 d_i를 구하여 그 차이가 0인가에 대한 문제로 바꾸어 추론한다.

$$d_i = X_i - Y_i \ (i = 1, 2, \cdots, n)$$

- 가정사항
  1. $d_i = X_i - Y_i$들은 각 모집단에서의 랜덤표본이라 가정.
  2. 모집단은 정규분포를 따른다고 가정.

     (표본크기 n이 충분히 클 때에는 근사적으로 성립함)
  3. 실험단위를 랜덤하게 추출하고, 처리의 적용순서도 랜덤하게 설정.

대응비교에 의한 모평균의 비교

- d_i의 기대값: $E(d_i) = \mu_1 - \mu_2 = \mu_D$

- 모평균 μ_D의 추정량: 표본평균 $\bar{D} = \sum_{i=1}^{n} \frac{d_i}{n}$
  모분산의 추정량: 표본분산 S_D^2

$$S_D^2 = \frac{1}{n-1} \sum (d_i - \bar{d})^2$$

- $\bar{D}$의 기대값: $E(\bar{D}) = \mu_D$
  $\bar{D}$의 분산:

$$Var(\bar{D}) = \frac{s_D^2}{n}$$

대응비교에 의한 모평균의 비교

- $H_0 : \mu_1 - \mu_2 = 0$ $(\mu_D = 0)$ 에 대한 검정통계량:

$$T = \frac{\bar{D} - \mu_D}{\sqrt{\frac{s_D^2}{n}}} \sim t_{n-1}, \quad df = n - 1$$

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

$$\bar{d} \pm t_{\alpha/2}(n-1)\sqrt{\frac{s_D^2}{n}}$$

- 검정통계량의 관측값을 t_0라고 할 때

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \mu_D > 0$	$P = P(T \geq t)$	$t_0 \geq t_\alpha(n-1)$
(b) $H_1 : \mu_D < 0$	$P = P(T \leq t)$	$t_0 \leq -t_\alpha(n-1)$
(c) $H_1 : \mu_D \neq 0$	$P = P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t_0\rvert \geq t_{\alpha/2}(n-1)$

대응비교 예제

Example

두 종류의 진통제가 수면에 미치는 효과를 비교하기 위해 실험을 진행하려 한다. 그러나 실험에 참여한 환자들의 건강상태가 서로 간에 크게 달라, 이를 고려한 실험설계가 필요하다. 따라서 환자 6명을 랜덤추출하고, 각 환자에게 진통제 두 종류를 각각 1회씩 복용시킨 후 수면 시간의 차이를 측정하여 효과를 비교하기로 했다. 이때 복용 순서에 따라 효과가 달라질 수 있으므로 환자마다 어떤 진통제를 먼저 복용할지를 동전을 던져 무작위로 결정했다. 두 진통제 A, B의 수면시간을 각각 μ_A, μ_B라 하자. 아래 표에 제시된 실제 조사 결과를 바탕으로 수면 시간의 차 $\mu_A - \mu_B$에 대한 95% 신뢰구간을 구하여라.

표 6-1 진통제에 따른 수면 시간의 측정값

(단위 : 시간)

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

sol) 대응비교에 의한 자료이므로 그 차이 d_i를 구하면 다음과 같다.

Table: 진통제에 따른 숙면시간의 측정값(시간)

환자	1	2	3	4	5	6
진통제A	4.8	4.0	5.8	4.9	5.3	7.4
진통제B	4.0	4.2	5.2	4.9	5.6	7.1
차이	0.8	-0.2	0.6	0.0	-0.3	0.3

- 차이에 대한 평균과 표준편차를 구하면 각각 다음과 같다.

$$\bar{d} = 0.20, \quad s_D = 0.443$$

- 자유도는 $n - 1 = 5$이므로 $t_{0.025}(5) = 2.571$이다.
- 따라서, $\mu_A - \mu_B$의 95% 신뢰구간은 다음과 같다.

$$(\bar{d} \pm t_{0.025}(5)\sqrt{\frac{s_D^2}{n}} = 0.20 \pm 2.571 \times \frac{0.443}{\sqrt{6}} = (-0.26, 0.66)$$

대응비교 예제

Example

재활치료를 위해 개발된 보행 보조 웨어러블 로봇이 실제로 운동 효과를 증가시키는지 확인하고자 한다. 이를 위해 한 재활 센터에서 무작위로 선정한 15명을 대상으로 실험을 진행했다. 모든 환자가 웨어러블 로봇을 착용하지 않은 상태와 착용한 상태에서 각각 10분간 가벼운 걷기 운동을 수행했을 때의 소모 칼로리를 비교한 결과는 아래 표와 같다. 재활치료 시 웨어러블 로봇을 착용하면 칼로리 소모가 더 커진다는 증거가 있는지 유의수준 1%에서 검정하라.

표 6-2 보행 보조 웨어러블 로봇 착용 전과 후의 소모 칼로리

(단위 : kcal)

로봇 착용 여부	1	2	3	4	5	6	7	8
미착용	23.9	23.9	20.0	19.2	22.2	11.8	21.7	27.7
착용	25.2	28.2	24.9	20.2	19.7	16.9	22.6	20.8
로봇 착용 여부	9	10	11	12	13	14	15	
미착용	15.6	22.5	23.2	19.3	15.2	10.0	11.2	
착용	23.8	18.3	23.7	22.5	16.2	12.1	22.5	

sol) 웨어러블 로봇 착용 전 평균 소모 칼로리를 μ_1, 착용 후 평균 소모 칼로리를 μ_2라 하면,

- 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \ (\text{or } \mu_D = 0) \quad H_1 : \mu_1 < \mu_2 \ (\text{or } \mu_D < 0)$$

- 관측값의 차이 d_i를 구하면 다음과 같다.

$$-1.3 \quad -4.3 \quad -4.9 \quad -1.0 \quad 2.5 \quad -5.1 \quad -0.9 \quad 6.9$$
$$-8.2 \quad 4.2 \quad -0.5 \quad -3.2 \quad -1.0 \quad -2.1 \quad -11.3$$

- 이들의 평균과 표준편차를 구하면 각각 다음과 같고

$$\bar{d} = -2.01, \quad s_D = 4.58$$

- 이로부터 t검정통계량의 관측값을 계산하면

$$t_0 = \frac{\bar{d} - 0}{s_D/\sqrt{n}} = \frac{-2.01 - 0}{4.58/\sqrt{15}} = -1.70$$

- 자유도는 $n-1=14$이므로 $t_{0.01}(14)=2.624$이다.
- 기각역으로 판정:

  기각역은 $t_0 \le -2.624$이고, $t_0 = -1.7 \ge -2.624$이므로 유의수준 1%에서 귀무가설을 기각할 수 없다.

- 유의확률은 다음과 같다.

$$P = P\{T \le -1.7\}, \quad T \sim t(14)$$

  t분포표로부터 $0.05 < P < 0.1$이다. 통계 소프트웨어를 이용하면 실제 유의확률은 $P=0.056$으로 계산된다.
- 종합하면 1% 유의수준에서는 재활치료 웨어러블 로봇이 소모 칼로리를 증가시킨다는 뚜렷한 증거가 없다.

등분산 이표본 비교 VS 대응비교

- 주어진 자료: $(X_1, \cdots, X_n), (Y_1, \cdots, Y_n)$
- 모평균 $\mu_X = E(X_i)$, $\mu_Y = E(Y_i)$를 비교하고자 한다.
  총 자료의 갯수: n or $2n$, σ_X^2, σ_Y^2의 값은 모른다고 가정.

- 가설: $H_0 : \mu_X = \mu_Y$ $(or\ \mu_D = 0)$

- 각각의 검정 통계량과 표본 분포는 다음과 같다.

[등분산 이표본의 비교]

$$T = \frac{(\bar{X} - \bar{Y}) - 0}{S_p\sqrt{\frac{1}{n} + \frac{1}{n}}} \sim t(2n-2)$$

[대응비교]

$$T = \frac{\bar{D} - 0}{\sqrt{\frac{S_D^2}{n}}} \sim t(n-1)$$

등분산 이표본 비교 VS 대응비교

1. 검정 통계량의 분자는 동일하다.

$$\bar{D} = \bar{X} - \bar{Y}$$

2. $(\bar{X} - \bar{Y})$의 분산을 생각해 보자.
   - If $\bar{X}$ and $\bar{Y}$ are independent:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$$

   - If $\bar{X}$ and $\bar{Y}$ are dependent:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) - 2Cov(\bar{X}, \bar{Y})$$

$\bar{X}$와 $\bar{Y}$가 양의 상관관계를 보인다면$(Cov(\bar{X}, \bar{Y}) > 0)$ 대응비교의 검정통계량 값이 더 커져서 H_0를 기각하기 쉬워진다.

$$S_p\sqrt{\frac{1}{n} + \frac{1}{n}} > \sqrt{\frac{S_D^2}{n}}$$

등분산 이표본 비교 VS 대응비교

3. t분포의 자유도를 생각해보자.

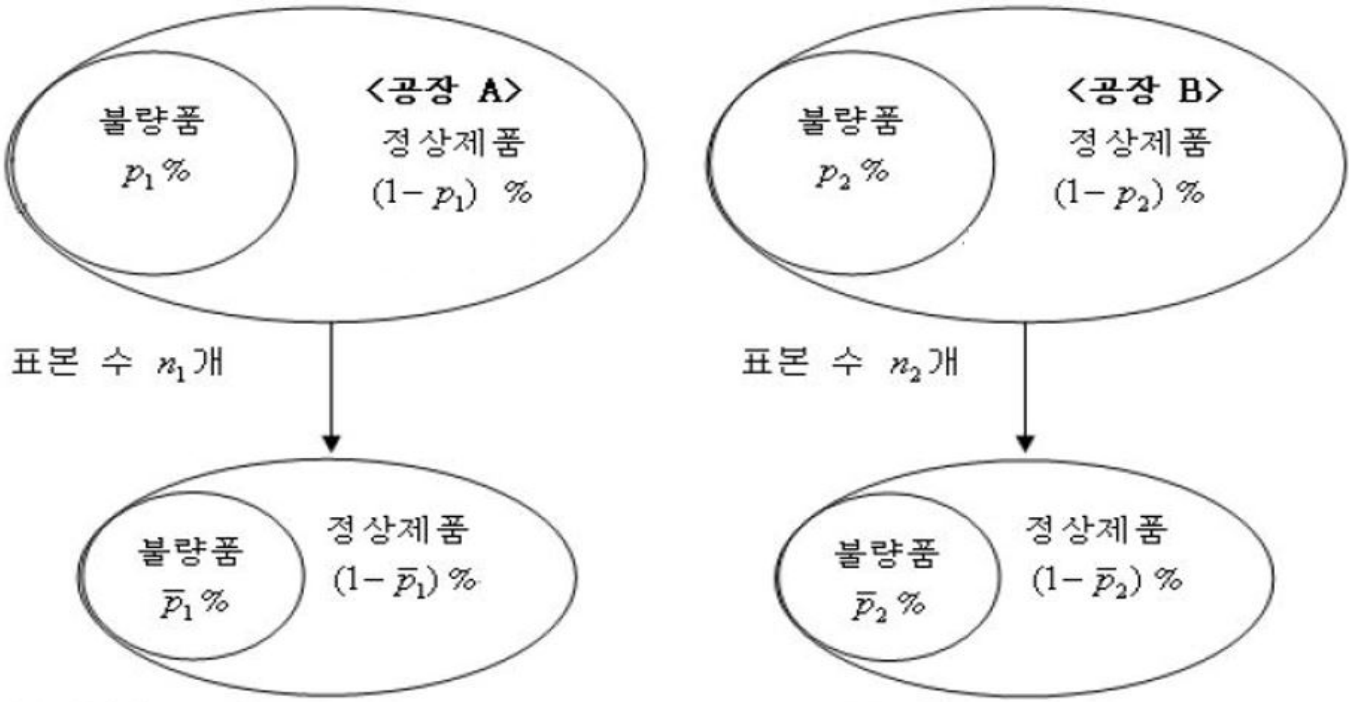
- 등분산 이표본의 비교에선 자유도가 $2n-2$이고 대응비교에서는 $n-1$이다. 따라서 같은 조건일 때 대응비교의 귀무가설 기각이 더 힘들어 진다. (대응비교는 자유도에서 penalty가 있음)

- 따라서 실험단위들이 이질적이고 각 쌍 내에서 값들이 양의 상관관계를 가진다면 대응비교가 적합하고,
- 위의 조건이 아니라면, 자유도의 penalty때문에 대응비교를 사용하는 장점이 사라지게 된다.

두 모비율의 차이에 대한 추론

예) 서로 다른 두 치료방법에 의한 치유율의 비교,
두 제조 과정에서의 불량률의 비교

- 두 모집단에서 각각 n_1, n_2개의 랜덤표본을 서로 독립적으로 추출한 경우에 관심의 대상인 속성이 두 표본에서 나타나는 도수를 각각 X_1, X_2라고 하자.



두 모비율의 차이에 대한 추론

- X_1과 X_2는 각각 이항분포 $B(n_1, p_1)$, $B(n_2, p_2)$를 따르며, 서로 독립이고, p_1과 p_2의 추정량은 $\hat{p}_1 = X_1/n_1$, $\hat{p}_2 = X_2/n_2$이라고 하자.
- 두 모비율의 차인 $p_1 - p_2$의 추정량:

$$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = X_1/n_1 - X_2/n_2$$

- $\hat{p}_1 - \hat{p}_2$의 기대값:

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

- $\hat{p}_1 - \hat{p}_2$의 분산:

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- $\hat{p}_1 - \hat{p}_2$ 표준오차 추정량:

$$\widehat{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

두 모비율의 차이에 대한 추론

- 표본크기 n_1, n_2가 충분히 큰 경우

$$\hat{p}_1 \dot{\sim} N(p_1, p_1(1-p_1)/n_1), \quad \hat{p}_2 \dot{\sim} N(p_2, p_2(1-p_2)/n_2)$$

- $\hat{p}_1$과 $\hat{p}_2$은 서로 독립이므로

$$(\hat{p}_1 - \hat{p}_2) \dot{\sim} N(p_1 - p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2)$$

- 따라서 n_1, n_2가 충분히 큰 경우 다음이 성립하고

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

- 모비율을 표본비율로 대체하면 표본분포는 다음과 같다.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \dot{\sim} N(0,1) \tag{2}$$

두 모비율의 차이에 대한 신뢰구간

- 다음을 이용하여 $p_1 - p_2$에 대한 신뢰구간을 구할 수 있다.

$$P\left\{-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \leq z_{\alpha/2}\right\} \doteq 1 - \alpha$$

- $p_1 - p_2$에 대한 $100(1-\alpha)\%$신뢰구간:

$$((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\widehat{SE}(\hat{p}_1 - \hat{p}_2),\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\widehat{SE}(\hat{p}_1 - \hat{p}_2))$$

- 표본크기가 크다는 것의 기준:
  $n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2,\ n_2(1-\hat{p}_2)$가 모두 5이상

두 모비율의 차이에 대한 신뢰구간 예제

Example (7.5)

현재 누구나 그 효과를 인정하는 소아마비 백신이지만, 1954년 미국에서는 그 효능을 검증하기 위한 대규모 임상시험이 시행되었다. 백신의 효과를 올바르게 평가하기 위해, 백신이 아닌 위약을 접종받는 어린이들로 구성된 대조집단과 백신을 접종받는 처리집단을 대상으로 이중 눈가림 실험이 실시되었다. 접종 후 마비 증세를 보인 어린이들의 수를 관측한 결과는 다음과 같았다.

표 7-1 1954년 소아마비 백신의 실험 결과

(단위 : 명)

	표본 크기	마비 증세가 나타난 어린이
대조집단	201,229	115
처리집단	200,745	33

대조집단과 처리집단에서 마비 증세를 보인 비율은 p_1, p_2라 할 때, $p_1 - p_2$의 추정값과 95%신뢰구간을 구하여라.

sol) 모비율 p_1과 p_2의 추정값이 각각

$$\hat{p}_1 = 115/201,229 = 5.71 \times 10^{-4}, \quad \hat{p}_2 = 33/200,745 = 1.64 \times 10^{-4}$$

이므로, $p_1 - p_2$의 추정값은 $\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2 = 4.07 \times 10^{-4}$이다.

- 이 추정값의 표준오차는 다음과 같으므로

$$\widehat{SE}(\widehat{p_1 - p_2}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{201,229} + \frac{\hat{p}_2(1-\hat{p}_2)}{200,745}} = 0.60 \times 10^{-4}$$

- $p_1 - p_2$의 95%신뢰구간은 다음과 같다.

$$(4.07 \times 10^{-4} - 1.96 \times 0.60 \times 10^{-4},\ 4.07 \times 10^{-4} + 1.96 \times 0.60 \times 10^{-4})$$

- 즉 신뢰수준 95%에서 $p_1 - p_2$ 값의 범위를 다음과 같이 추정할 수 있다.

$$2.89 \times 10^{-4} < p_1 - p_2 < 5.25 \times 10^{-4}$$

두 모비율의 비교를 위한 검정

- 귀무가설($H_0 : p_1 = p_2$)이 사실일 때,
  두 모집단에서의 공통인 모비율 p를 다음과 같이 생각해보자.

$$p_1 = p_2 = p$$

- p의 추정량: $\hat{p}$ 합동표본비율(pooled sample proportion)

  -두 표본을 합하여 모두 $n_1 + n_2$개의 표본에서 $X_1 + X_2$개가 대응하는 속성을 가진 것으로 생각하자.

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

두 모비율의 비교를 위한 검정

- $\hat{p}_1 - \hat{p}_2$의 분산:

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

- $\hat{p}_1 - \hat{p}_2$의 분산의 추정량:

$$\widehat{Var}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

- $(\hat{p}_1 - \hat{p}_2)$의 표본분포는 귀무가설이 $H_0 : p_1 = p_2 (= p)$이 참일 때 다음과 같다.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p(1-p)(n_1^{-1} + n_2^{-1})}}$$

두 모비율의 비교를 위한 검정

- 귀무가설: $H_0 : p_1 = p_2$
- 검정통계량은 공통 모비율이 $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$ 일 때 다음과 같다.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} \dot{\sim} N(0,1)$$

검정통계량의 관측값을 z_0라고 할 때, 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : p_1 > p_2$	$P = P(Z \geq z)$	$z_0 \geq z_\alpha$
(b) $H_1 : p_1 < p_2$	$P = P(Z \leq z)$	$z_0 \leq -z_\alpha$
(c) $H_1 : p_1 \neq p_2$	$P = P(\lvert Z \rvert \geq \lvert z \rvert)$	$\lvert z_0 \rvert \geq z_{\alpha/2}$

두 모비율의 비교를 위한 검정 예제

Example

Example 7.5의 자료를 이용하여, 소아마비 백신이 효과가 있다고 할 수 있는가를 유의수준 1%에서 검정 하여라.

sol) p_1과 p_2는 각각 가짜약과 백신 접종 후의 마비율을 나타낸다고 하자.

- 검증하려는 가설은 다음과 같다.

$$H_0 : p_1 = p_2, \qquad H_1 : p_1 > p_2$$

- 귀무가설 하에서 공통 모비율의 추정값을 구하면

$$\hat{p} = \frac{115 + 33}{201229 + 200745} = 3.68 \times 10^{-4}$$

- 검정통계량의 관측값은

$$z_0 = \frac{5.71 \times 10^{-4} - 1.64 \times 10^{-4}}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} = 6.727$$

- 오른쪽 단측인 대립가설인 경우의 유의확률을 구하면 다음과 같이 매우 작은 값이 도출됨을 알 수 있고, 따라서 귀무가설을 기각 할 수 있다.

$$P = P(Z \geq 6.727) = 8.66 \times 10^{-12} < 0.01$$

- 즉 임상실험 결과는 소아마비 백신이 예방 효과가 뚜렷함을 시사하고 있다.

두 모비율의 비교를 위한 검정 예제

Example (7.7)

[표 7-2]는 한 범죄학 학술지에 실린 연구 결과로, 충동적 살인범과 계획적 살인법의 교정 효과에 차이가 있는지 조사하기 위한 것이다. 일정 기간 복역 후 가석방된 충동적 살인범과 계획적 살인범 중 각각 42명과 40명을 랜덤추출하여 가석방이 성공인 경우(재범 없음)와 실패인 경우(재범 발생)를 관측한 결과가 다음과 같다. 살인범의 유형에 따라 가석방 성공률, 즉 재범하지 않을 확률에 차이가 있는지를 유의수준 5%에서 검정하라.

표 7-2 살인범의 유형에 따른 가석방 성공 여부

(단위 : 명)

	성공	실패	표본 크기
충동적 살인범	13	29	$n_1 = 42$
계획적 살인범	22	18	$n_2 = 40$
합계	35	47	$n = 82$

sol) 충동적 살인범과 계획적 살인범에 대한 가석방의 성공률을 각각 p_1, p_2라고 하면 검정하려는 가설은 다음과 같다.

$$H_0 : p_1 = p_2, \qquad H_1 : p_1 \neq p_2$$

- p_1, p_2와 귀무가설 하에서의 공통 모비율의 추정값은

$$\hat{p}_1 = 13/42 = 0.31, \qquad \hat{p}_2 = 22/40 = 0.55$$

$$\hat{p} = \frac{13 + 22}{42 + 40} = 0.427$$

- 이로부터 검정통계량의 관측값을 구하면

$$z_0 = \frac{0.31 - 0.55}{\sqrt{\hat{p}(1 - \hat{p})(n_1^{-1} + n_2^{-1})}} = -2.20$$

- 유의수준 5%에서 검정통계량의 관측값이 기각역 $|z_0| \geq z_{0.025} = 1.96$에 포함되므로, 귀무가설을 기각 할 수 있다.

두 모집단의 분산에 대한 추론

- 두 모집단의 분산 비교의 예
  - 서로 다른 투자 포트폴리오에 관련된 위험 비교
  - 두 모평균의 비교 시 등분산 가정이 적합한지에 대한 분석

- 두 모집단의 표준편차 σ_1, σ_2를 비교

$$\sigma_1^2 = \sigma_2^2 \iff \sigma_1^2/\sigma_2^2 = 1$$

$$\sigma_1^2 > \sigma_2^2 \iff \sigma_1^2/\sigma_2^2 > 1$$

$$\sigma_1^2 < \sigma_2^2 \iff \sigma_1^2/\sigma_2^2 < 1$$

두 모집단의 분산에 대한 추론

- 추정량
  - 정규모집단 1의 랜덤표본: $X_{11}, \cdots, X_{1n_1}$
  - 정규모집단 2의 랜덤표본: $X_{21}, \cdots, X_{2n_2}$
  - 모분산 σ_1^2, σ_2^2의 점추정량: 표본분산 s_1^2, s_2^2
  - 모분산 비율 $\frac{\sigma_1^2}{\sigma_2^2}$의 점추정량: $\frac{s_1^2}{s_2^2}$

- 모분산 비율에 대한 추론시 유용한 분포:

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

이를 이용하여 σ_1^2/σ_2^2에 대한 추정이나 유의성검증을 할 수 있음

F 분포

Definition (F 분포:F distribution)

V_1과 V_2가 서로 독립이고 각각 자유도 k_1, k_2인 카이제곱분포를 따를 때,

$$F = \frac{V_1/k_1}{V_2/k_2}$$

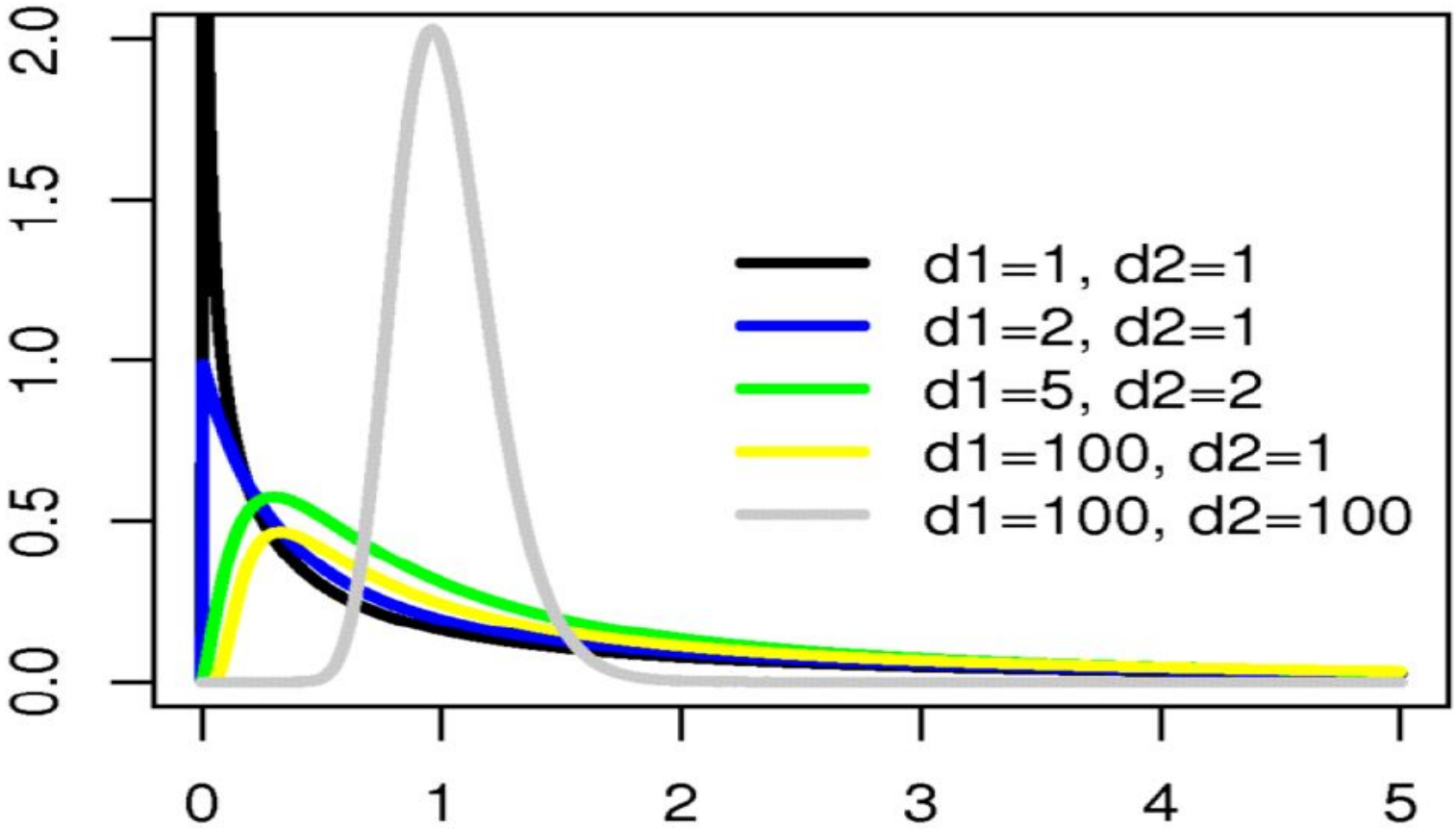
의 분포를 자유도 (k_1, k_2)인 F 분포라고 한다. 이 때 기호로는 다음과 같이 나타낸다.

$$F \sim F(k_1, k_2)$$

- 분자 자유도 k_1과 분모 자유도 k_2에 따라 그 모양이 달라진다.
- 두 정규모집단의 분산 비교시 분산비에 관한 추론에 이용됨

F 분포

Figure: 자유도의 변화에 따른 여러가지 F분포의 모양



F 분포

- $F \sim F(k_1, k_2)$일 때 다음의 식에서 $F_\alpha(k_1, k_2)$를 자유도가 (k_1, k_2)인 F 분포의 $(1-\alpha)$분위수라고 한다.

$$P\{F \geq F_\alpha(k_1, k_2)\} = \alpha$$

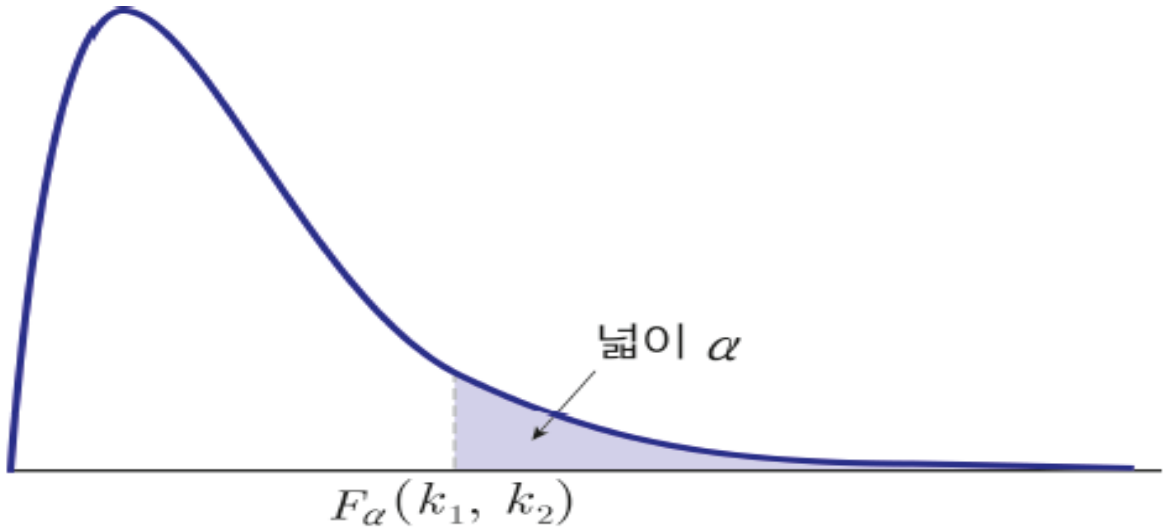
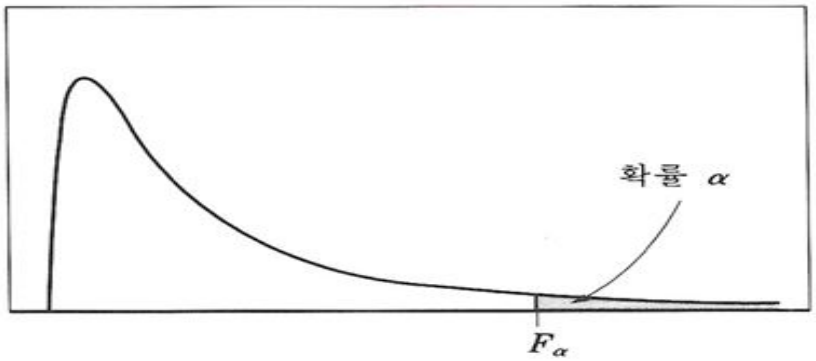


그림 4–10 F 분포의 형태

F 분포표

- F 분포표에서 자유도별로 여러 가지 분위수를 구할 수 있다.
  예) $F_{0.10}(4,2) = 9.24$, $F_{0.05}(4,2) = 19.25$



표의 숫자는 주어진 α 값에 대한 F_α의 값을 나타낸다.

		분자의 자유도								
분모의 자유도	α	1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	5000000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39

F 분포의 성질

Theorem (F 분포의 성질)

확률변수 F가 자유도 (k_1, k_2)인 F 분포를 따를 때, 다음이 성립한다.

$$\frac{1}{F} \sim F(k_2, k_1)$$

- F 분포의 성질을 이용하면 다음과 같은 사실을 알 수 있다.

$$F_{1-\alpha}(k_2, k_1) = \frac{1}{F_{\alpha}(k_1, k_2)}$$

예) $F_{0.05}(3, 4) = 6.59$이므로 $F_{0.95}(4, 3) = 1/6.59 \approx 0.15$

F분포와 t분포와의 관계

Theorem (F분포와 t분포와의 관계)

확률변수 T가 자유도 k인 t분포를 따를 때 다음이 성립한다.

$$T^2 \sim F(1, k)$$

pf) 확률변수 T가 자유도 k인 t분포를 따를 때 다음과 같이 가정하자.

$T = \frac{Z}{\sqrt{V/k}}$ (단, $Z \sim N(0,1)$, $V \sim \chi^2(k)$이며 서로 독립)

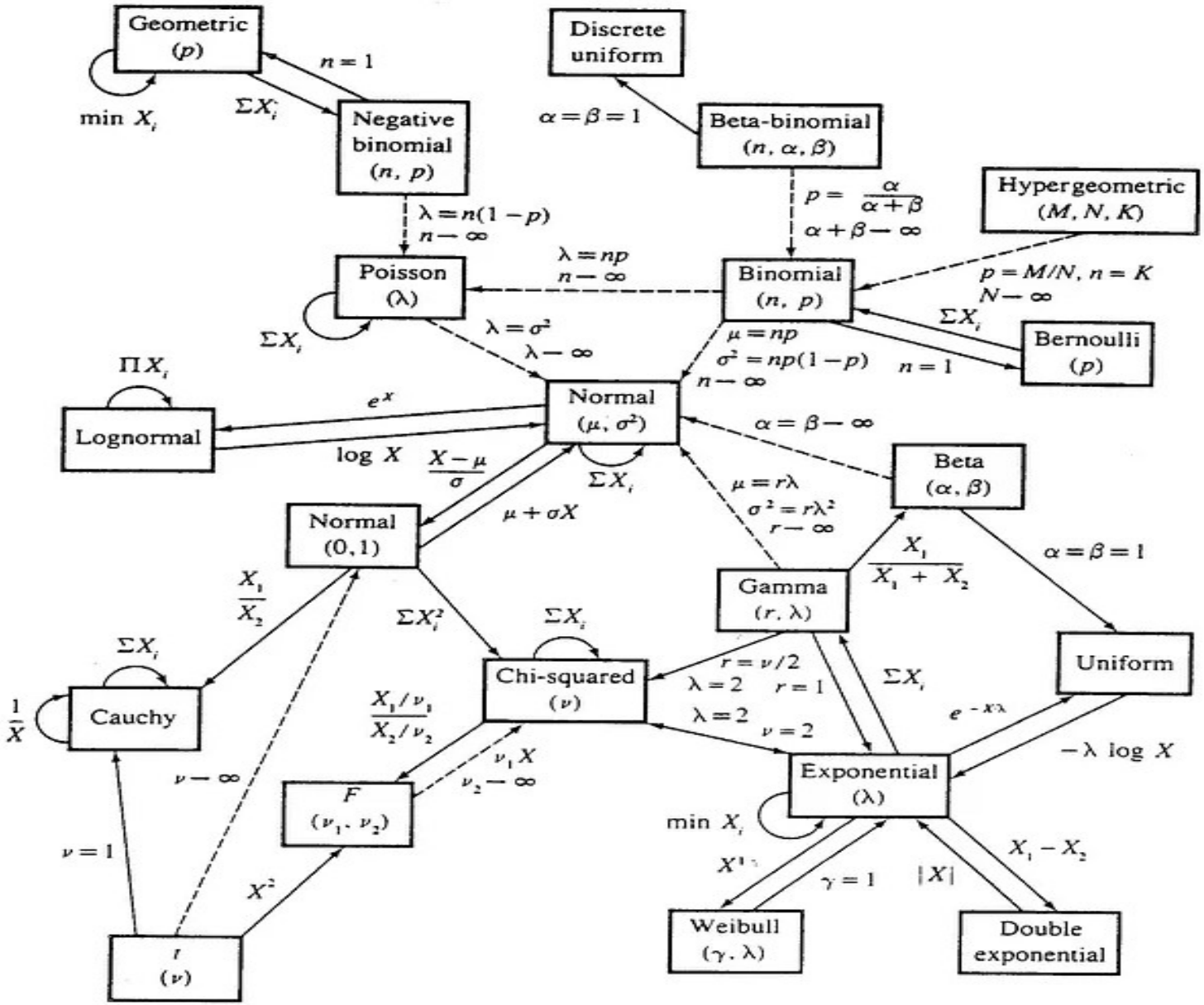
양변을 제곱하면 다음과 같고

$$T^2 = \frac{Z^2/1}{V/k}$$

여기에서 $Z^2 \sim \chi^2(1)$이고, Z와 V는 서로 독립이므로 다음의 사실이 성립한다.

$$T^2 \sim F(1, k)$$

Table of Common Distributions



표본분산의 비에 관한 분포

Theorem

$(X_{11}, \cdots, X_{1n_1})$, $(X_{21}, \cdots, X_{2n_2})$가 서로 독립이고 각각 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$에서의 랜덤표본이라 하자. 이 때, S_1^2과 S_2^2을 각각 두 랜덤표본의 표본분산이라고 하면 다음이 성립하고, 이들은 서로 독립이다.

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

따라서 F분포의 정의에 따라 이들을 각각 자신의 자유도로 나눈 후의 비는 다음과 같이 자유도 (n_1-1, n_2-1)인 F분포를 따르게 된다.

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

- 모분산 비교시 가정:
  1. 두 모집단이 정규분포를 따름
  2. 표본추출시 독립적인 단순 임의추출을 해야 한다.

모분산의 비에 대한 신뢰구간(정규모집단의 경우)

- 다음을 이용하여 σ_1^2/σ_2^2에 대한 신뢰구간을 구할 수 있다.

$$P\left(F_{1-\alpha/2}(n_1-1,n_2-1) \le \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \le F_{\alpha/2}(n_1-1,n_2-1)\right) = 1-\alpha$$

$$P\left(\frac{S_1^2}{S_2^2}/F_{\alpha/2,n_1-1,n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}/F_{1-\alpha/2,n_1-1,n_2-1}\right) = 1-\alpha$$

한편, F분포의 성질: $F_{1-\alpha/2}(n_1-1,n_2-1) = 1/F_{\alpha/2}(n_2-1,n_1-1)$

- 모분산의 비 σ_1^2/σ_2^2에 대한 $100(1-\alpha)\%$ 신뢰구간 :

$$\left(\frac{s_1^2}{s_2^2}/F_{\alpha/2}(n_1-1,n_2-1),\ \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2}(n_2-1,n_1-1)\right)$$

모분산의 비에 대한 유의성검증(정규모집단의경우)

- 귀무가설 $H_0 : \sigma_1^2 = \sigma_2^2$인 경우에 검정통계량은

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$$

이고, 이의 관측값을 f라고 할 때 다음이 성립한다.

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \sigma_1^2 > \sigma_2^2$	$P = P(F \geq f)$	$f \geq F_\alpha(n_1 - 1, n_2 - 1)$
(b) $H_1 : \sigma_1^2 < \sigma_2^2$	$P = P(F \leq f)$	$f \leq 1/F_\alpha(n_2 - 1, n_1 - 1)$
(c) $H_1 : \sigma_1^2 \neq \sigma_2^2$	$P = 2P(F \geq f)$ or $P = 2P(F \leq f)$	$f \geq F_{\alpha/2}(n_1 - 1, n_2 - 1)$ or $f \leq 1/F_{\alpha/2}(n_2 - 1, n_1 - 1)$

  - 양측가설인 경우 $2P(F \geq f)$ or $2P(F \leq f)$ 중 1보다 작은 값을 유의확률로 계산

	유의확률 P	유의수준 α의 기각역
(a) $H_1 : \sigma_1^2 > \sigma_2^2$	$P = P(F \geq f)$	$f \geq F_\alpha(n_1 - 1, n_2 - 1)$
(b) $H_1 : \sigma_1^2 < \sigma_2^2$	$P = P(F \leq f)$	$f \leq \frac{1}{F_\alpha(n_2-1,n_1-1)}$
(c) $H_1 : \sigma_1^2 \neq \sigma_2^2$	$P = 2P(F \geq f)$ or $P = 2P(F \leq f)$	$f \geq F_{\alpha/2}(n_1 - 1, n_2 - 1)$ or $f \leq \frac{1}{F_{\alpha/2}(n_2-1,n_1-1)}$

두 모집단의 분산에 대한 추론 예제

Example

콘크리트에 균열이 발생하면 이를 보수하기 위해 주로 폴리머(polymer)를 주입한다. 다음은 한 연구보고서에서 폴리머 두 종류를 사용했을 때 측정한 주입 압축률에 대한 자료이다.

(단위 : 10^6psi)

에폭시(epoxy)	1.75	2.12	2.05	1.97
MMA 프리폴리머(prepolymer)	1.77	1.59	1.70	1.69

폴리머 두 종류의 주입 압축률 산포가 다르다는 증거가 있는지를 유의수준 5%에서 검정하고, 이들의 표준편차 비에 대한 95% 신뢰구간을 구하여라. (정규모집단이라고 가정)

sol) 두 폴리머의 주입 압축률의 표준편차를 각각 σ_1, σ_2라고 하자.

- 검정하고자 하는 가설은 다음과 같다.

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

- 주어진 자료에서 평균과 표준편차를 구하면 각각 다음과 같다.

	Epoxy	MMA Prepolymer
표본크기(n_i)	4	4
평균($\bar{x}_i$)	1.97	1.69
표준편차(s_i)	0.160	0.074

- F검정통계량의 관측값을 계산하면

$$f = S_1^2/S_2^2 = 0.160^2/0.074^2 = 4.675$$

- 기각역으로 판정:

  기각역은 $f \leq 1/F_{0.025}(3,3) = 1/15.44$ 또는 $f \geq F_{0.025}(3,3) = 15.44$

  검정통계량의 관측값 $f = 4.675$이 기각역에 속하지 않아 유의수준 5%에서 귀무가설을 기각할 수 없다.

- 모분산의 비인 σ_1^2/σ_2^2에 대한 95% 신뢰구간은 다음과 같고

$$\left(\frac{0.160^2}{0.074^2}/15.44,\ \frac{0.160^2}{0.074^2} \cdot 15.44 \right) = (0.301, 72.182)$$

- 모표준편차의 비인 σ_1/σ_2에 대한 95% 신뢰구간은 다음과 같다.

$$(\sqrt{0.301}, \sqrt{72.182}) = (0.549, 8.496)$$

두 정규모집단에서의 표본분산의 분포(등분산 가정)

- $X_1, \cdots, X_{n_1}$과 $Y_1, \cdots, Y_{n_2}$은 서로 독립이고 각각 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$에서의 랜덤표본이라 가정하자.
- 각 표본의 표본분산은 다음과 같고,

$$S_1^2 = \sum_{i=1}^{n_1}(X_i - \bar{X})^2/(n_1 - 1), \quad S_2^2 = \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2/(n_2 - 1)$$

- 각각의 분포는 서로 독립이고 다음과 같다.

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1), \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

- 따라서 카이제곱분포의 가법성에 의하여 다음이 성립한다.

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) \tag{3}$$

두 정규모집단에서의 표본분산의 분포(등분산 가정)

- 이때, S_p^2을 다음과 같이 정의하면

$$S_p^2 = \frac{(n_1-1)}{(n_1-1)+(n_2-1)} S_1^2 + \frac{(n_2-1)}{(n_1-1)+(n_2-1)} S_2^2$$
$$= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

- (3)의 식을 다음과 같이 나타낼 수 있다.

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

- S_p^2은 공통분산 σ^2의 추정량으로서 모분산이 σ^2인 두 랜덤표본을 모두 이용하여 만든 것이다. 이를 두 랜덤표본의 합동표본분산(pooled sample variance)이라고 한다.