확률과 통계

섹션 - 4

강사 : James 쌤

가설검정의 원리

키포인트

- 가설검정의 원리.
- 가설검정의 순서.
- 가설검정 오류의 유형.

가설검정 개요

- 가설검정은 모집단의 모수에 대해서 기존 통설을 가설로 하고 표본을 추출하여 얻은 "검정 통계량"으로 가설의 진위를 판단하는 것.
- 한 차례 표본 추출 만으로 검정 (test)한다.
- 가설 검정의 대상이 되는 주장:

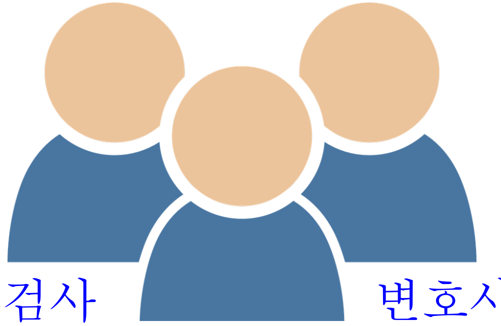
  예). "대한민국 성인 남성의 독서량은 1년에 10권이다"

  예). "대한민국 직장인의 1년 노동시간은 2900시간이다"

  예). "S사의 사원 근속연수는 15년이다"

현대 법정의 원리

증거



검사

피고인

변호사

현대 법정의 원리

- 충분한 증거를 제시하기 전까지는 피고의 무죄를 전제한다.
- 피고의 무죄는 증명할 필요 없다.
- 증명의 대상은 피고의 “유죄” 여부이다.
- 판사는 증거를 바탕으로 판결을 내려야 한다.

가설검정의 요소

- H_0 : 귀무가설. 모집단의 모수에 대한 통설 또는 주장. $\Rightarrow$ “피고의 무죄”
- H_1 : 대립가설. 대립적인 주장. $\Rightarrow$ “피고의 유죄”
- 검정 통계량 : 검정을 위하여 표본으로 계산한 값. $\Rightarrow$ “증거”
- p-값 (유의 확률) : 귀무가설이 옳다는 전제를 하고 계측된 검정 통계량이 발생했을 확률. $\Rightarrow$ “증거”

  $\Rightarrow$ p-값이 크면 귀무가설이 옳다는 전제가 강해진다.

  $\Rightarrow$ p-값이 작으면 귀무가설이 옳다는 전제가 약해지며 대립가설이 강해진다.

가설검정의 요소

- 유의수준 α : 귀무가설이 사실이더라도 기각될 수 있는 최대 확률.

  ⇒ *p*-값에 대한 "임계치"의 역할을 한다. ⇒ "판결 수준"

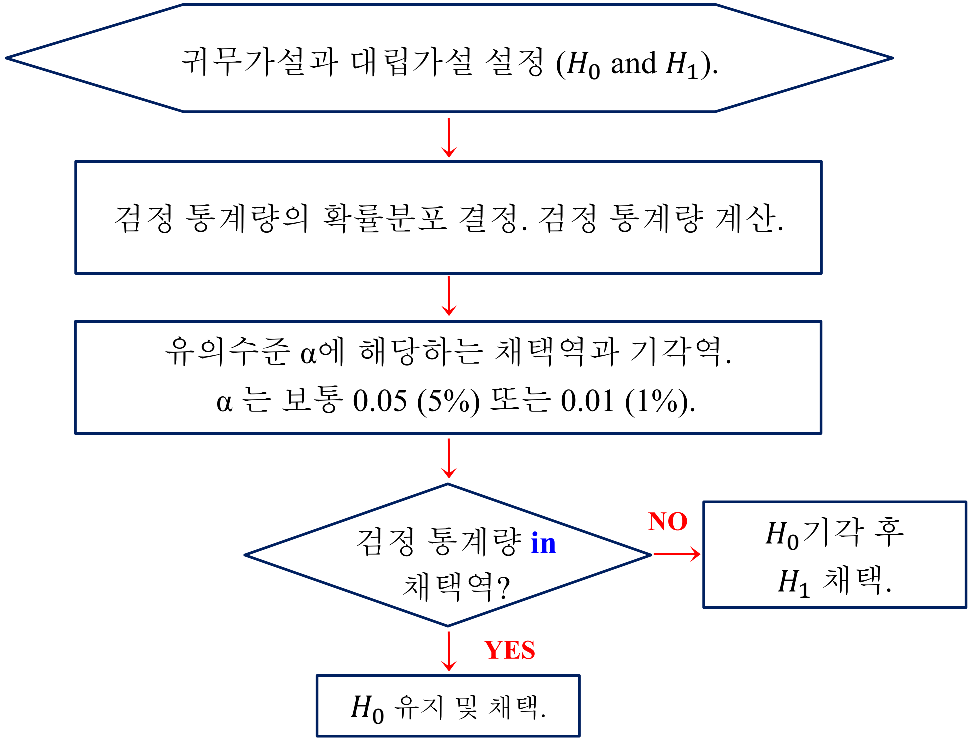
  ⇒ 보통 5% (0.05)를 사용한다.

- 1 – α : 사실인 귀무가설이 채택될 확률.

- 검정력 : 허위인 귀무가설이 기각되고 대립가설이 채택될 확률.

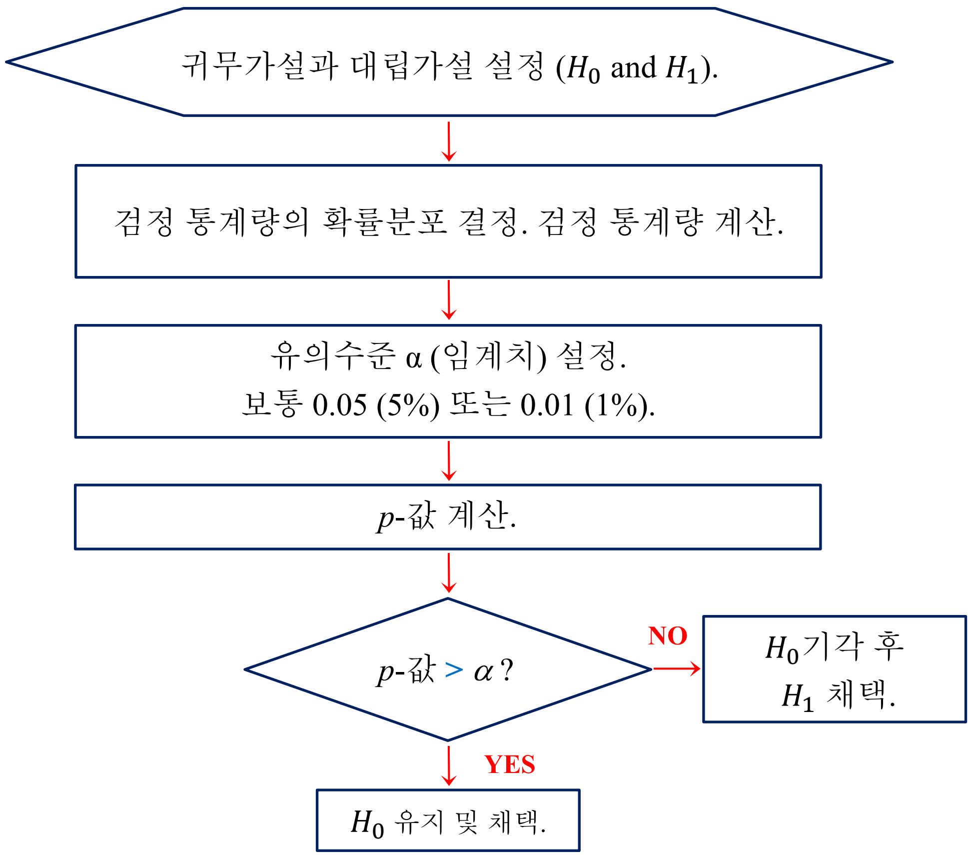
가설검정의 순서

검정 통계량을 사용하는 경우



가설검정의 순서

p-값을 사용하는 경우



가설검정 오류의 유형

실제상황 / 검정결과	귀무가설이 사실임	귀무가설이 허위임
귀무가설 채택 대립가설 무시	옳은 결정 확률: $1-\alpha$	2종 오류 확률: β
귀무가설 기각 대립가설 채택	1종 오류 확률: α → "유의수준"	옳은 결정 확률: $1-\beta$ → "검정력"

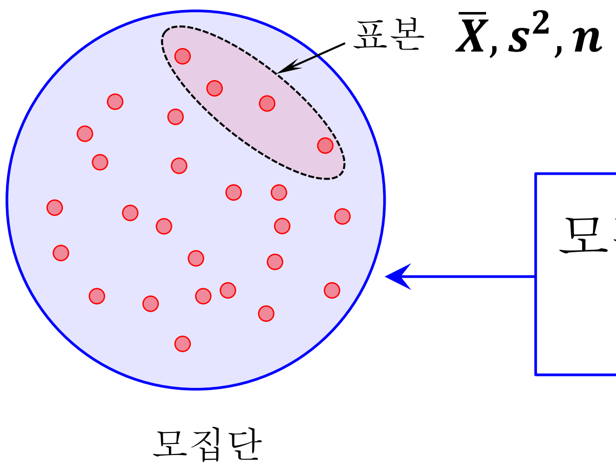
모평균 검정

키포인트

- 모집단이 한 개인 경우 모평균 검정.
- 모집단이 두 개인 경우 모평균 검정.
- 독립표본과 대응표본의 차이.

모집단 한 개

- 우측검정 (Right-Tail Test).
- 좌측검정 (Left-Tail Test).
- 양측검정 (Two-Tail Test).



모평균 μ 를 특정 수치 μ_0와 비교하는 검정.

모집단 한 개

- 모표준편차를 아는경우의 검정 통계량: “z 통계량” and “z 검정”

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- 모표준편차를 모르는 경우의 검정 통계량: “t 통계량” and “t 검정”

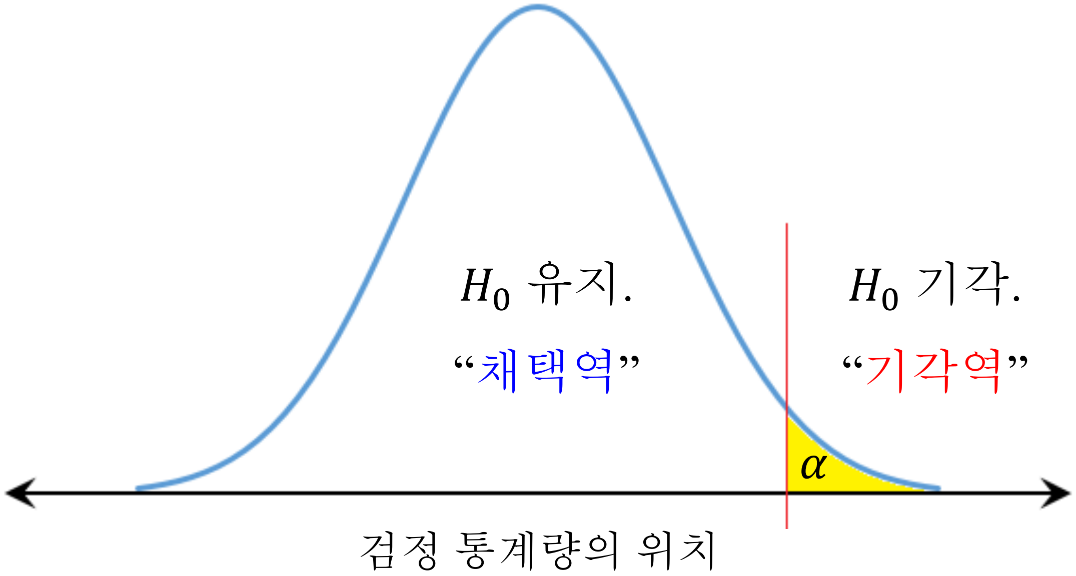
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

모집단 한 개

- 우측검정 (Right-Tail Test):

$H_0 \ : \mu \leq \mu_0$

$H_1 \ : \mu > \mu_0$ "더 큼"



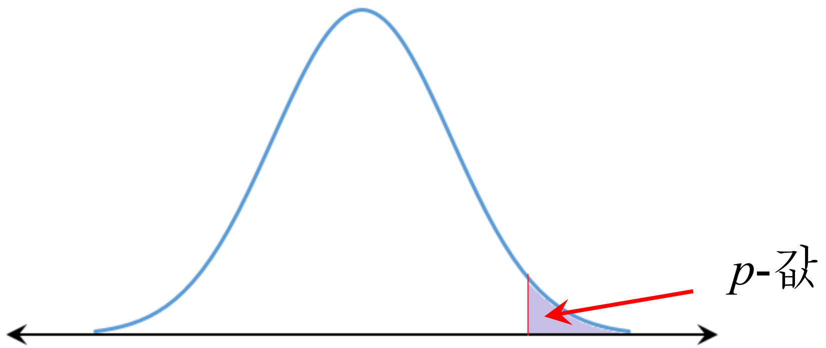
모집단 한 개

- 우측검정 (Right-Tail Test):

  $H_0 \ : \mu \leq \mu_0$

  $H_1 \ : \mu > \mu_0$ "더 큼"

- 검정 통계량을 사용하여 p-값을 구한 후 α와 비교해서 p-값 > α 이면 H_0유지 아니면 H_0기각하고 H_1채택. p-값 $= P(Z > \text{검정 통계량})$.

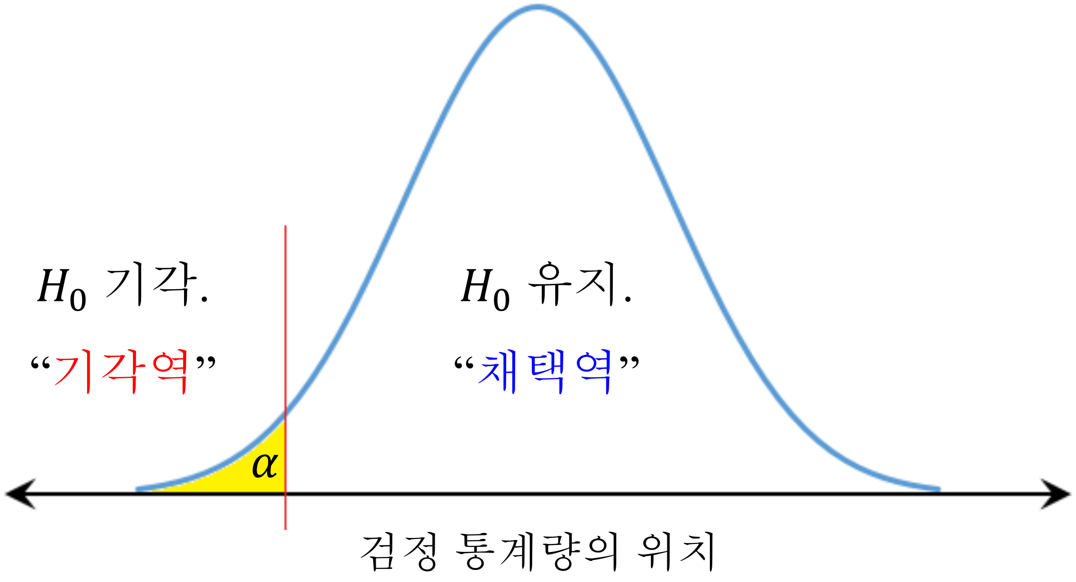


모집단 한 개

- 좌측검정 (Left-Tail Test):

$H_0 \ : \mu \geq \mu_0$

$H_1 \ : \mu < \mu_0$ "더 작음"



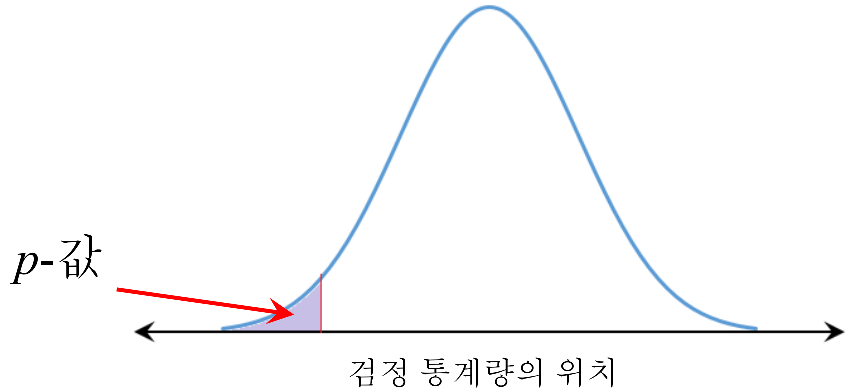
모집단 한 개

- 좌측검정 (Left-Tail Test):

H_0 : $\mu \geq \mu_0$

H_1 : $\mu < \mu_0$ "더 작음"

- 검정 통계량을 사용하여 p-값을 구한 후 α와 비교해서 p-값 > α 이면 H_0유지 아니면 H_0기각하고 H_1채택. p-값 $= P(Z <$ 검정 통계량$)$.

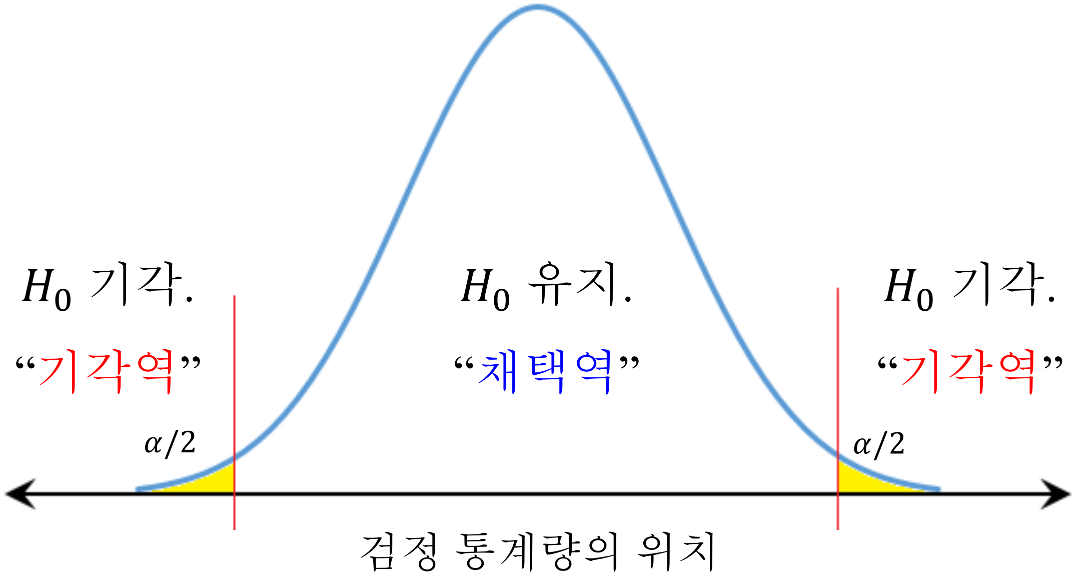


모집단 한 개

- 양측검정 (Two-Tail Test):

$H_0 \ : \ \mu = \mu_0$

$H_1 \ : \ \mu \neq \mu_0$ "다름"



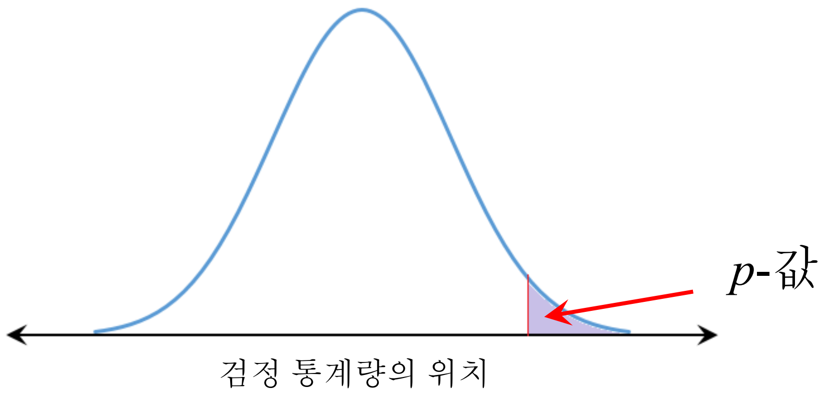
모집단 한 개

- 양측검정 (Two-Tail Test):

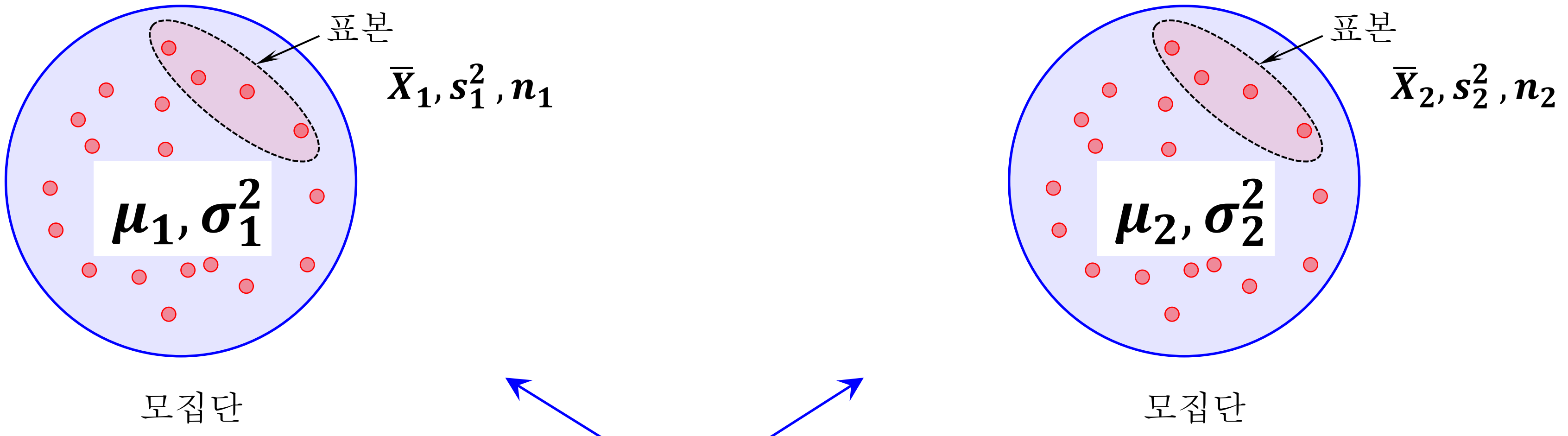
$H_0 \ : \mu = \mu_0$

$H_1 \ : \mu \neq \mu_0$ "다름"

- 검정 통계량의 절대값을 사용하여 p-값을 구한 후 α와 비교해서 ***p*-값 > α** 이면 H_0유지 아니면 H_0기각하고 H_1채택. p-값 $= 2 \times P(Z > |$검정 통계량$|)$.



모집단 두 개: 독립표본



두 개의 모평균 μ_1과 μ_2를

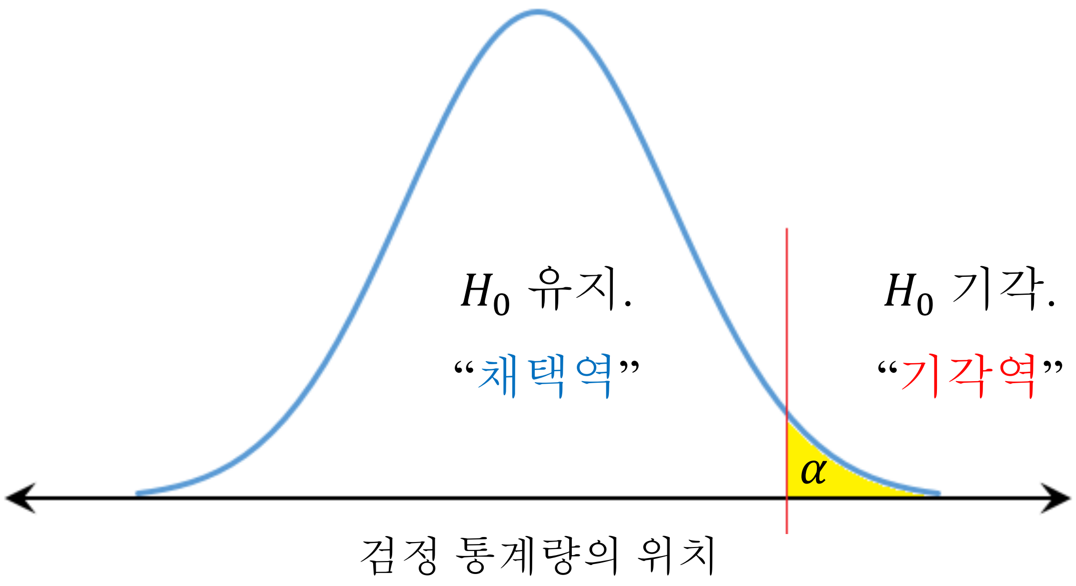
서로 비교하는 검정.

모집단 두 개: 독립표본

- 우측검정 (Right-Tail Test):

$$H_0 \ : \mu_1 - \mu_2 \leq 0 \quad \Leftrightarrow \quad \mu_1 \leq \mu_2$$

$$H_1 \ : \mu_1 - \mu_2 > 0 \quad \Leftrightarrow \quad \mu_1 > \mu_2$$

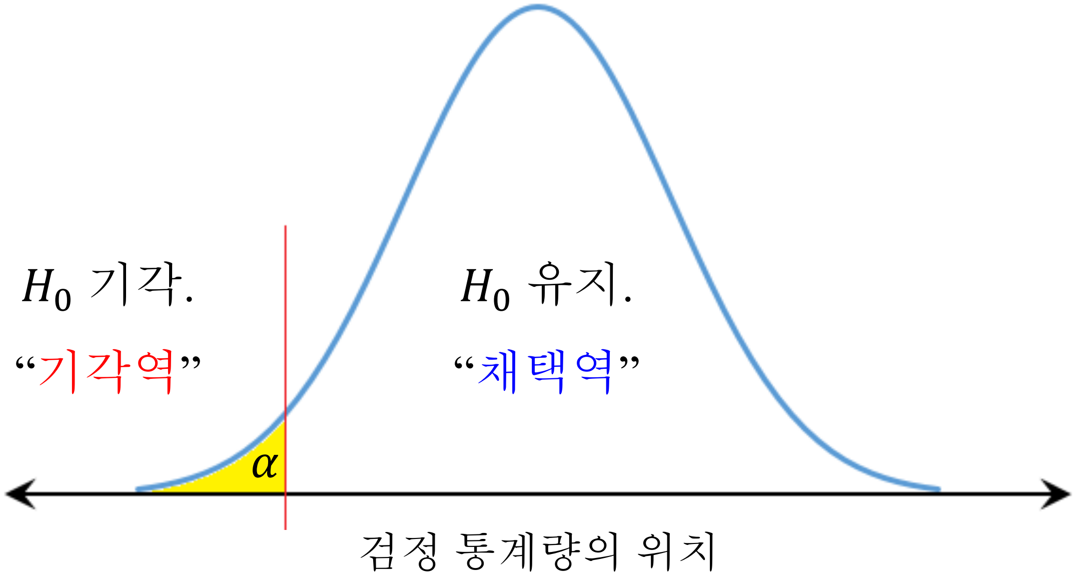


모집단 두 개: 독립표본

- 좌측검정 (Left-Tail Test):

$$H_0 \; : \mu_1 - \mu_2 \geq 0 \quad \Leftrightarrow \quad \mu_1 \geq \mu_2$$

$$H_1 \; : \mu_1 - \mu_2 < 0 \quad \Leftrightarrow \quad \mu_1 < \mu_2$$

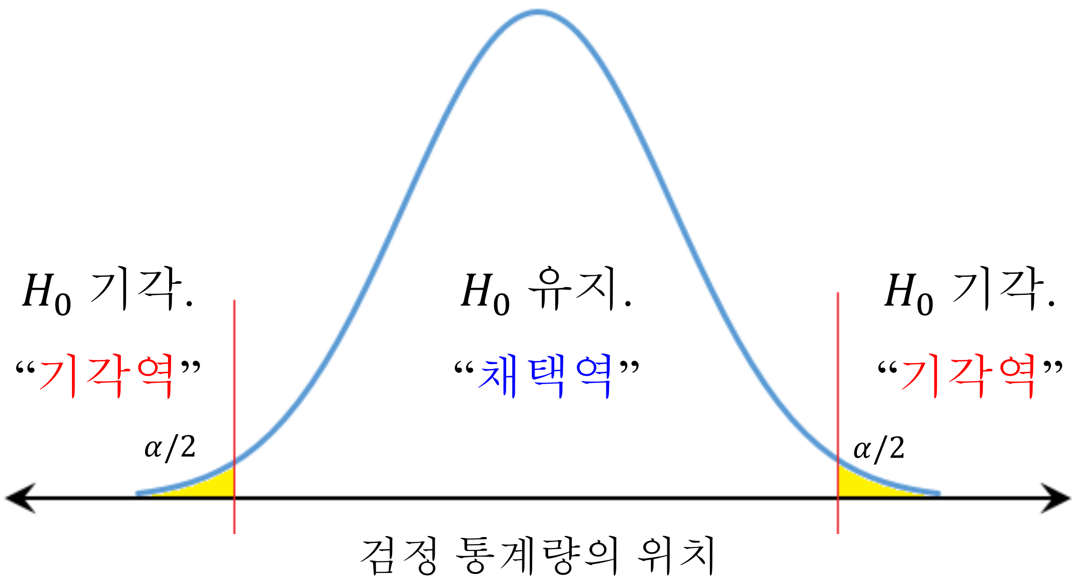


모집단 두 개: 독립표본

- 양측검정 (Two-Tail Test):

$$H_0 \ : \mu_1 - \mu_2 = 0 \quad \Leftrightarrow \quad \mu_1 = \mu_2$$

$$H_1 \ : \mu_1 - \mu_2 \neq 0 \quad \Leftrightarrow \quad \mu_1 \neq \mu_2$$



모집단 두 개: 독립표본

- 표본의 크기가 크면 (대략 $n > 30$), z 통계량을 사용해서 z-검정을 실행할 수 있다. 또한 모분산을 아는지 여부와는 상관 없다.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- 또한 다음 슬라이드에서 알아보게 되는 "합동분산" vs "이분산" 고려가 필요 없다.

모집단 두 개: 독립표본

- 표본의 크기가 작고 ($n < 30$), 두 모집단의 모분산은 모르지만 서로 동일하다는 가정 하에 t-검정을 실시할 수 있다. ⇒ “합동분산”
- 표본의 크기가 작고 ($n < 30$), 두 모집단의 모분산은 모르지만 서로 다르다는 가정하에 t-검정을 실시할 수 있다. ⇒ “이분산”
- 모분산이 서로 동일한지 여부는 다음 슬라이드에서 다루게 되는 “분산비 검정”을 활용해서 알아볼 수 있다.

모집단 두 개: 분산비 검정

- 합동분산 또는 이분산은 분산비 검정을 통해서 밝혀낼 수 있다.

  H_0 : 모분산 사이에는 차이가 없다.

  H_1 : 모분산 사이에는 차이가 있다.

- F 분포함수를 사용하여 검정하며 검정 통계량은 다음과 같이 계산한다.

$$F = \frac{S_1^2}{S_2^2}$$

⇒ 확률분포 $F(n_1 - 1, n_2 - 1)$를 따른다. n_1과 n_2는 표본의 크기.

모집단 두 개: 대응표본

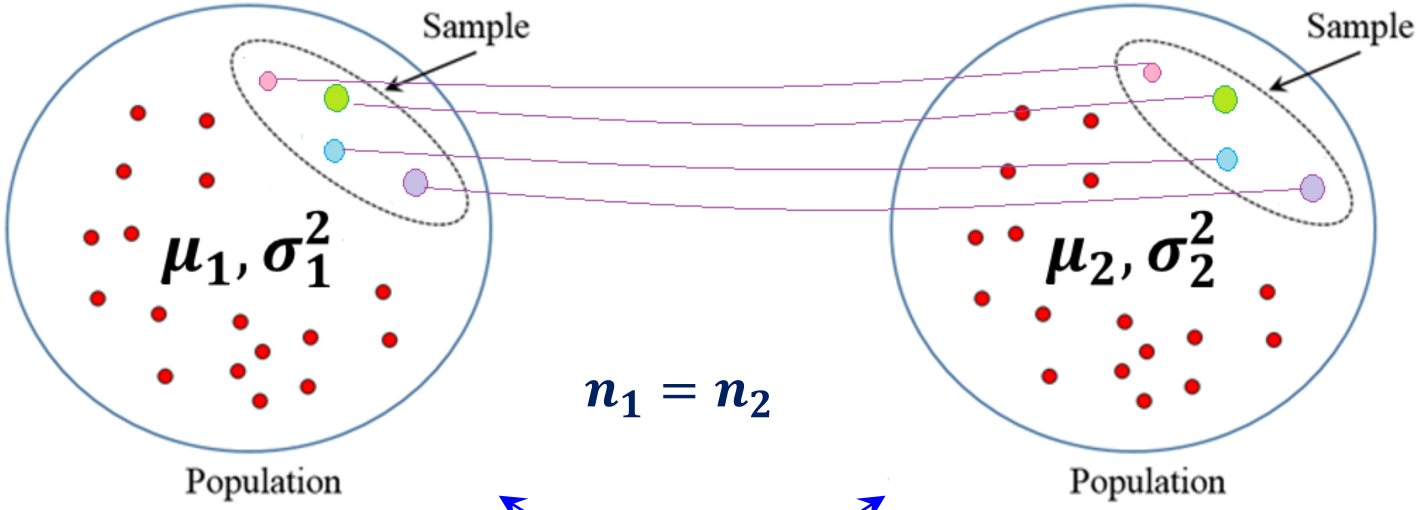
- 한 모집단에서 어떤 표본이 추출되고 그것과 대응 또는 쌍(pair)을 이루는 표본이 다른 모집단에서 추출되는 경우이다.
- 다음은 대응표본 검정을 적용할 수 있는 경우들이다.

  예). 새로 개발한 당료병 치료약을 복용한 후에는 혈당치에 차이가 있나?

  예). SNS 광고에 노출된 사람들과 노출되지 않은 사람들의 구매율에 차이가 있는가?

  예). Pepsi와 Coke 시음평가: 어느 것이 더 맛있나?

모집단 두 개: 대응표본



$n_1 = n_2$

두 개의 표본이 연결됨.

모집단 개수와 검정 유형

- 모집단이 한 개인 경우. ⇒ t 검정 or z 검정.
- 모집단이 두 개이며 독립적인 경우. ⇒ 독립표본 t 검정 or z 검정.
- 모집단이 두 개이며 서로 1대1 연결되어 있다. ⇒ 대응표본 t 검정.
- 모집단이 여러개 있는 경우. ⇒ ANOVA 검정. 차후 자세히 알아본다.

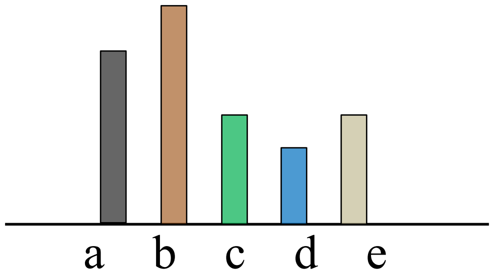
카이제곱 검정

키포인트

- 카이제곱 검정.
- 도수분포표와 적합도검정.
- 분할표와 독립성검정.

도수분포표

- 도수분포표 (frequency table)은 한개의 명목형(범주형) 변수가 있을 때 유형의 도수(거듭된 횟수)를 집계한 표이다.

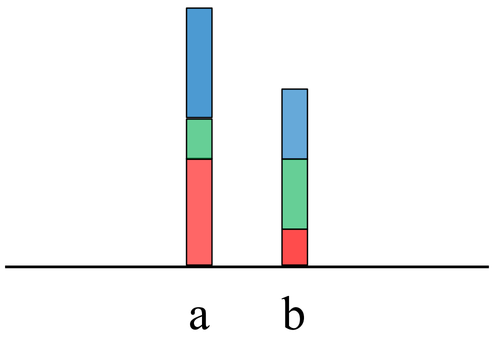


- 상대도수분포표는 도수의 상대적인 비율을 나타내어 준다.
- 연속형 변수를 사용하는 경우에는 구간을 설정해서 명목형 변수화 할 수 있다.

⇒ 이 경우 변수의 구간을 계급이라고 부른다. 구간의 레이블 ~ 명목형 변수의 값.

분할표

- 분할표 (contingency table)은 두 개의 명목형(범주형) 변수의 도수를 정리한 표이다.
- 분할표에서 도수분포표를 쉽게 구할 수 있다 (margin).



단변량 카이제곱검정

- 단변량 카이제곱검정 (Pearson's chi squared test)는 표본이 대표하는 모집단의 도수분포와 전제된 모형 사이의 차이 여부를 밝히기 위한 검정이다.

  ⇒ 귀무가설 $\boldsymbol{H_0}$: 모집단의 도수분포가 전제된 모형과 다르지 않다.

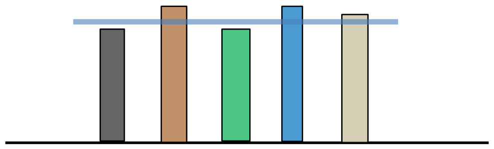
  ⇒ 대립가설 $\boldsymbol{H_1}$: 모집단의 도수분포가 전제된 모형과 다르다.

- 적합도 검정 (goodness of fit test)라고도 부른다.
- 두개의 유형만이 있는 경우에는 "비율검정" (proportion test)으로 대체될 수 있다.

단변량 카이제곱검정

- 예를 들어서 “Eye color (눈색) 변수”의 도수분포를 모형과 비교 검정해 본다.

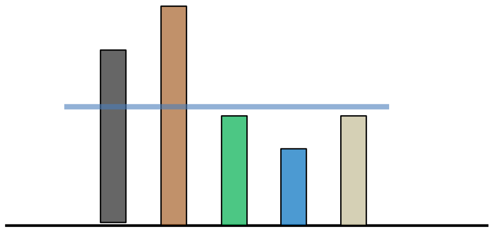
H_0 **Null Hypothesis (귀무가설)**



모형과 유의하게 다르지 않다.

vs.

H_1 **Alternative Hypothesis (대립가설)**



모형과 유의한 차이가 있다.

Eye color의 도수분포는 고르지 않다.

단변량 카이제곱검정

- 다음과 같은 검정 통계량을 사용한다. k =유형의 수이다.

$$\text{검정 통계량} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- $E_i = np_i$는 **모형이 제시하는 도수**이고 실제 측정된 값은 O_i이다.

- 위 검정 통계량은 자유도가 $k-1$인 카이제곱 분포를 따른다.

  $\Rightarrow$ O_i의 합은 n과 같다는 제약 조건이 적용된다.

  $\Rightarrow$ 제약 조건 때문에 자유도가 1 감소한다: k–1.

분할표 카이제곱검정

- 분할표를 사용해서 두 명목형 변수 사이의 독립성을 검정하고자 한다.

  ⇒ 귀무가설 $\boldsymbol{H_0}$: 두 명목형 변수는 서로 독립적이다. “연관성 없다”

  ⇒ 대립가설 $\boldsymbol{H_1}$: 두 명목형 변수는 서로 독립적이지 않다. “연관성 있다”

- 독립성 검정 (independence test)라고도 부른다.

NOTE: 귀무가설, 대립가설은 항상 모집단에 대한 주장이다.

분할표 카이제곱검정

- 예를 들어서 타이태닉호의 "생존 여부" 변수와 "성인 여부" 변수의 독립성 검정.

H_0

Null Hypothesis
(귀무가설)

연관성 없음
(독립)

	생존	생존하지 못함
성인	**55%**	45%
미성년	**55%**	45%

성인 여부는 생존율과 독립적이다.

vs.

H_1

Alternative Hypothesis
(대립가설)

연관성 있음
(독립 아님)

	생존	생존하지 못함
성인	38%	**62%**
미성년	**85%**	15%

성인여부와 생존율은 독립이 아니다.

미성년자의 생존율이 높다.

분할표 카이제곱검정

- 다음과 같은 검정 통계량을 사용한다. $r =$ 행의 수, $c =$열의 수이다.

$$\text{검정 통계량} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

- E_{ij}는 **독립 모형이 제시하는 도수**이고 실제 측정된 값은 O_{ij}이다.

- 위 검정 통계량은 자유도가 $(r-1)\times(c-1)$인 카이제곱 분포를 따른다.

  $\Rightarrow$ 제약 조건 때문에 자유도가 감소한다.

분할표 카이제곱검정 – 독립 모형

	남 (M)	여 (F)	계
A 반	30	70	100
B 반	50	40	90
계	80	110	190

- 반의 확률만 독립적으로 생각해 보면 다음과 같다:

$$P(\text{반} = A) = \frac{100}{190} \quad , \quad P(\text{반} = B) = \frac{90}{190}$$

- 성별의 확률만 독립적으로 생각해 보면 다음과 같다:

$$P(\text{성별=남}) = \frac{80}{190} \quad , \quad P(\text{성별=여}) = \frac{110}{190}$$

분할표 카이제곱검정 – 독립 모형

	남 (M)	여 (F)	계
A 반	30	70	100
B 반	50	40	90
계	80	110	190

- 만약에 반과 성별이 서로 독립적이라면 다음과 같은 인수분해가 가능해 진다:

$$P(\text{반}, \text{성별}) = P(\text{반})\times P(\text{성별})$$

$$\Rightarrow N(\text{반}, \text{성별}) = P(\text{반})\times P(\text{성별})\times N_{tot}$$

$$= E_{ij}$$ “변수들이 서로 독립인 경우 기대할 수 있는 도수”

- 그러므로, E_{ij}를 실제 측정된 값 O_{ij}와 비교하여 독립성 검정을 실행한다.

분산검정과 분산비검정

키포인트

- 분산검정 (카이제곱 확률분포 사용).
- 분산비검정 (F 확률분포 사용).

분산검정

- 한개의 모집단이 있고 이것의 분산을 특정수치 $\sigma_0{}^2$와 비교 검정한다.
- 양측검정:

$$H_0 \;: \sigma^2 = \sigma_0{}^2$$

$$H_1 \;: \sigma^2 \neq \sigma_0{}^2$$

- 좌측검정과 우측검정 방법도 있다.

분산검정

- 다음 검정 통계량은 카이제곱 확률분포를 따른다:

$$\text{검정 통계량} = \frac{(n-1)S^2}{\sigma_0^2}$$

⇒ 자유도 $= n - 1$인 카이제곱 확률분포를 따른다. n은 표본의 크기.

분산검정

- 다음과 같은 방법으로 95% 신뢰구간을 만들 수 있다:

$$P\left(qchisq(0.025) < \frac{(n-1)S^2}{\sigma^2} < qchisq(0.975)\right) = 0.95$$

$$\Downarrow$$

$$P\left(\frac{1}{qchisq(0.975)} \le \frac{\sigma^2}{(n-1)S^2} \le \frac{1}{qchisq(0.025)}\right) = 0.95$$

$$\Downarrow$$

$$P\left(\frac{(n-1)S^2}{qchisq(0.975)} \le \sigma^2 \le \frac{(n-1)S^2}{qchisq(0.025)}\right) = 0.95$$

분산검정

- 다음과 같은 방법으로 95% 신뢰구간을 만들 수 있다:

$$\frac{(n-1)S^2}{qchisq(0.975)} \qquad\qquad \frac{(n-1)S^2}{qchisq(0.025)}$$

[← <신뢰구간> →]

분산비 검정 (F 검정)

- 정규분포를 따르는 모집단이 두개 있고 이들의 분산을 비교 검정하고자 한다.
- 양측검정:

$\Rightarrow H_0 \ : \sigma_1{}^2 = \sigma_2{}^2$

$\Rightarrow H_1 \ : \sigma_1{}^2 \neq \sigma_2{}^2$

- 좌측검정과 우측검정 방법도 있다.

분산비 검정 (F 검정)

- 다음 검정 통계량은 F 확률분포를 따른다:

$$\text{검정 통계량} = \frac{S_1^2}{S_2^2}$$

$\Rightarrow F(n_1 - 1, n_2 - 1)$를 따른다. n_1과 n_2는 표본의 크기.

요점정리

검정	확률분포	R의 함수
단변량 카이제곱 검정 “적합도 검정”	카이제곱	chisq.test()
이변량 카이제곱 검정 “독립성검정”	카이제곱	chisq.test()
분산검정	카이제곱	EnvStats 패키지의 varTest()
분산비검정	F	var.test()

분산분석 (ANOVA)

키포인트

- 분산분석 (ANOVA).
- 일원 분산분석.
- 이원 분산분석.

분산분석

- 다수의 모집단 (with 표본)이 있는 경우 **모평균**을 비교 검정하고자 한다.
- 표본들을 짝지어서 t 검정을 하는 것은 번거롭고 오류가 누적될 수 있다.

$\Rightarrow$ 단 한번에 다음과 같은 가설을 검정하고자 한다.

“귀무가설” $\boldsymbol{H_0}: \mu_1 = \mu_2 = \mu_3 = \cdots$

“대립가설” $\boldsymbol{H_1}: \mu_i \neq \mu_j$와 같은 경우가 있다.

- 분산을 비교하여 검정을 한다. $\Rightarrow$ F 확률분포 사용.

분산분석

- **일원** 분산분석 : 한 개의 수치형 변수 Y와 한 개의 명목형 변수 X.

  $\Rightarrow Y$는 종속변수이고 X는 독립변수이다.

  예). 여러 반 X의 성적 Y를 비교 검정한다.

- **이원** 분산분석 : 한 개의 수치형 변수 Y와 두 개의 명목형 변수 X_A, X_B.

  $\Rightarrow Y$는 종속변수이고 X_A 와 X_B는 독립변수이다.

  예). 반 X_A과 성별 X_B로 그룹을 나누어서 성적 Y를 비교 검정한다.

분산분석

- 다음과 같은 조건을 전제한다:

  $\Rightarrow$ 종속변수 Y는 정규분포를 따른다.

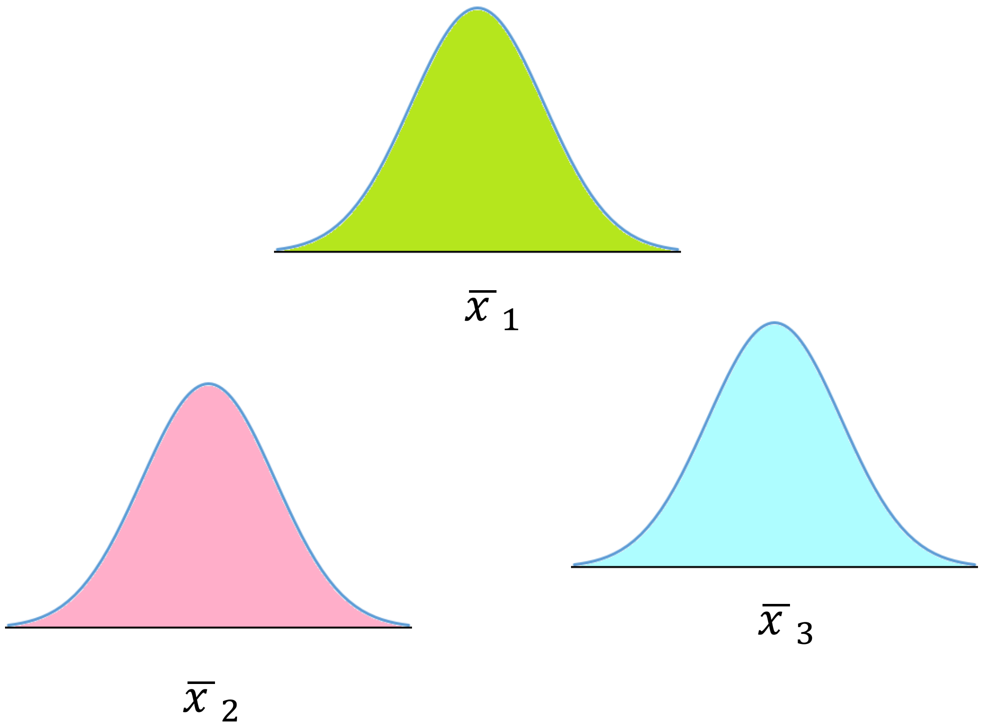
  $\Rightarrow$ 표본이 대표하는 모집단의 분산은 동일하다.

  $\Rightarrow$ 표본은 독립적으로 표집되었다.

일원 분산분석

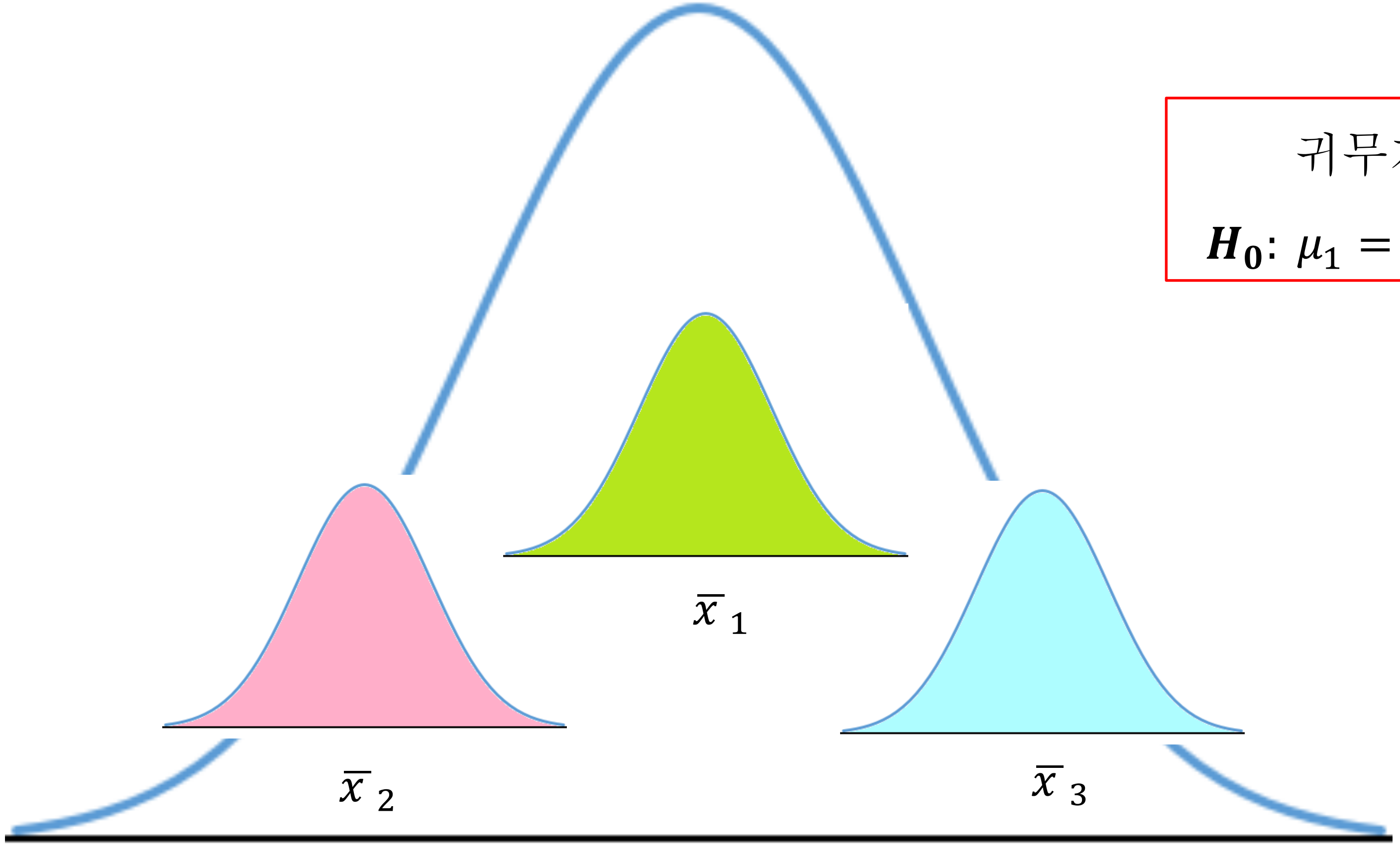
- 다음과 같이 여러 개의 표본 (처리)가 있을 때 이들의 표본평균을 비교해 본다.

  ⇒ 모집단의 평균은 서로 같은가?



일원 분산분석

- 전체적인 데이터 분포와 비교했을 때 표본평균들이 아래와 같이 모여 있을 수 있다.

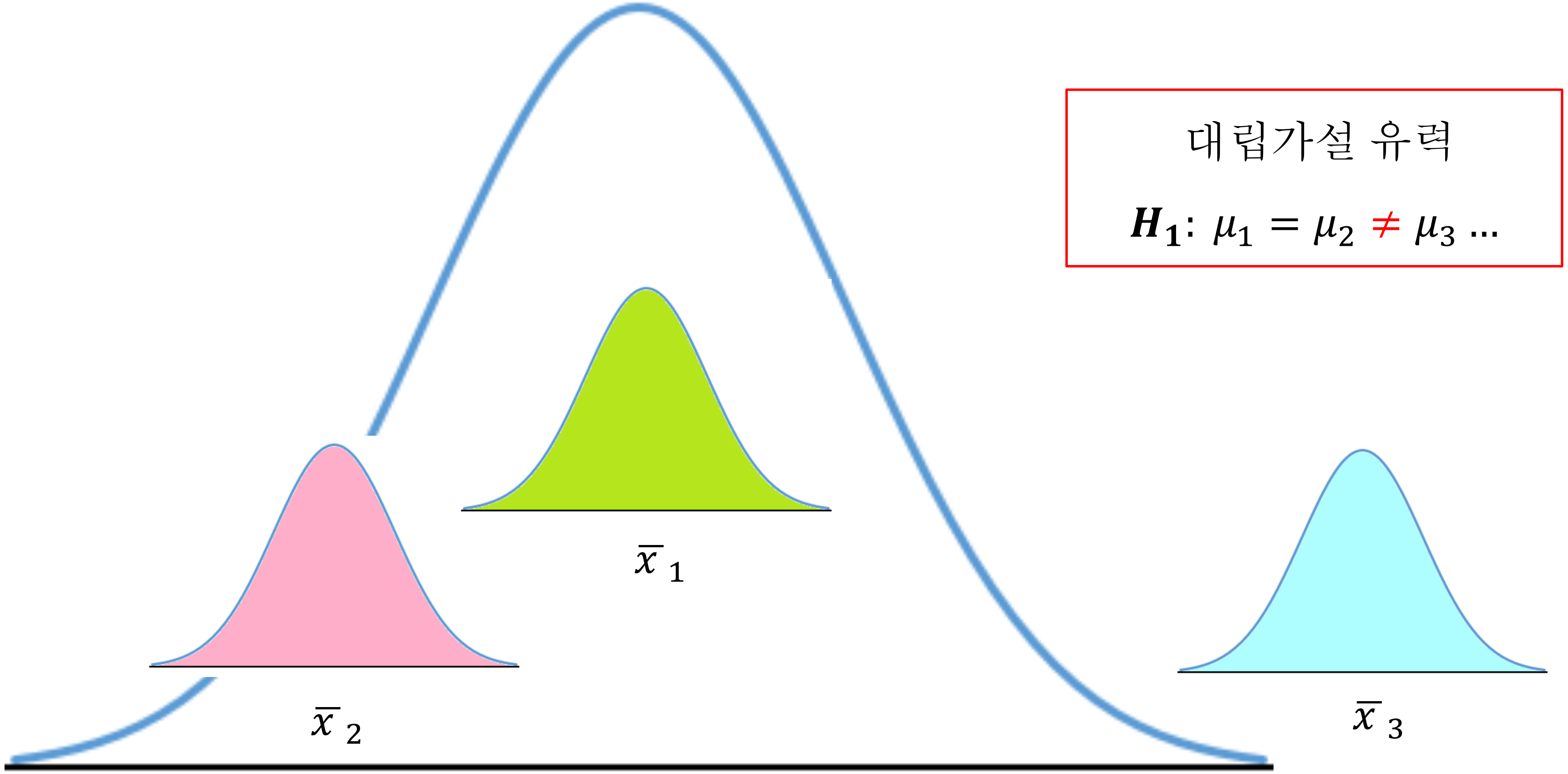


귀무가설 유력

$\boldsymbol{H_0}$: $\mu_1 = \mu_2 = \mu_3 = \cdots$

일원 분산분석

- 아니면 몇 개의 표본평균이 나머지와 많이 떨어져 있을 수도 있다.

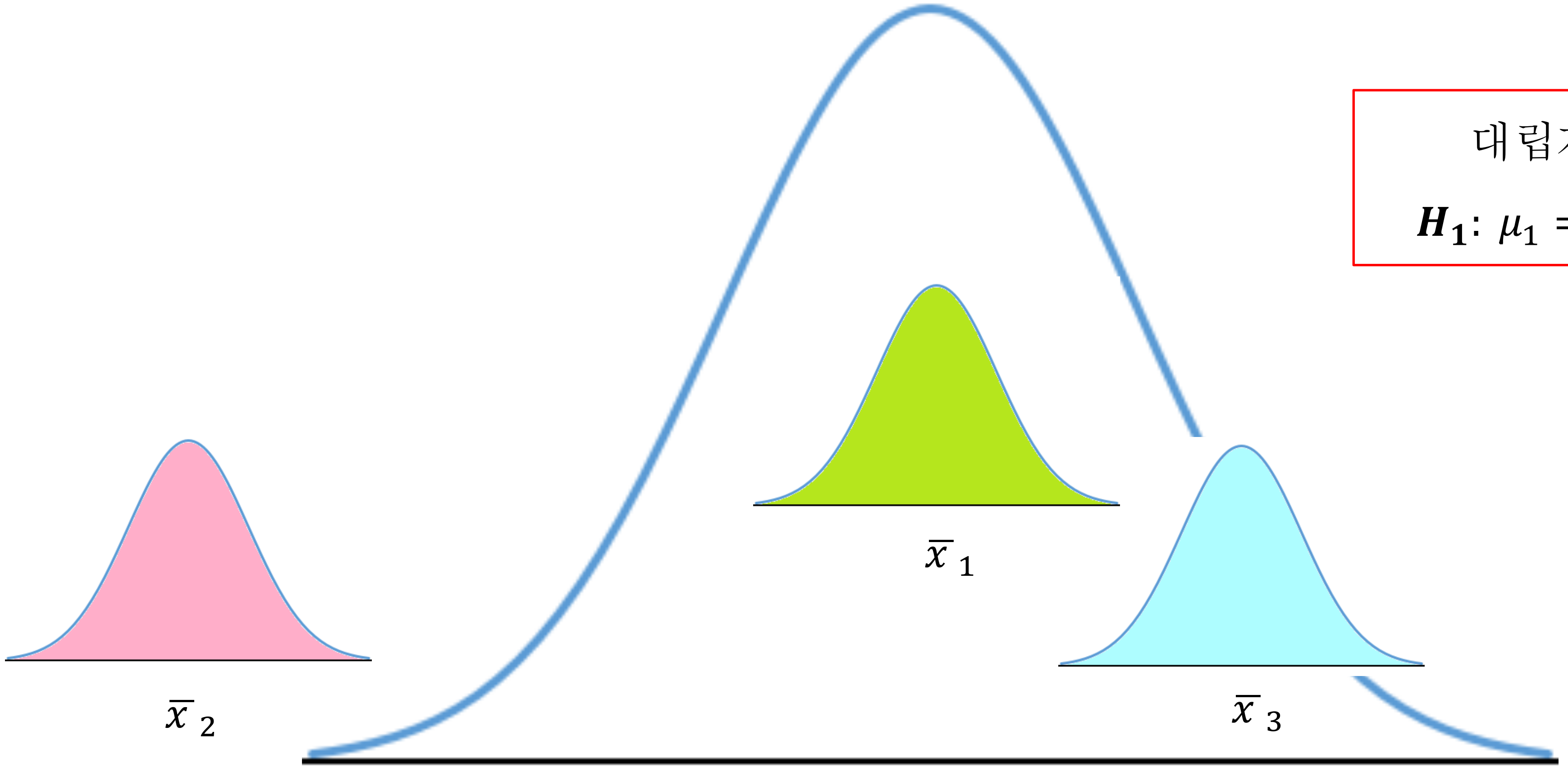


대립가설 유력

$\boldsymbol{H_1}$: $\mu_1 = \mu_2 \neq \mu_3 \ldots$

일원 분산분석

- 아니면 몇 개의 표본평균이 나머지와 많이 떨어져 있을 수도 있다.



대립가설 유력

$\boldsymbol{H_1}: \mu_1 = \mu_3 \neq \mu_2 \ldots$

일원 분산분석

- 다음과 같이 검정통계량을 만들어 본다:

$$\text{검정통계량} = \frac{\text{표본평균 사이의 분산 } (between)}{\text{표본내의 분산 } (within)} = \frac{MStr}{MSE}$$

$\Rightarrow$ $\text{검정통계량} = \dfrac{\text{큼}}{\text{작음}}$ $\Rightarrow$ $\boldsymbol{H_0}$ 기각 후 $\boldsymbol{H_1}$채택.

$\Rightarrow$ $\text{검정통계량} = \dfrac{\text{비슷}}{\text{비슷}}$ $\Rightarrow$ $\boldsymbol{H_0}$ 유지. 조금은 애매한 경우.

$\Rightarrow$ $\text{검정통계량} = \dfrac{\text{작음}}{\text{큼}}$ $\Rightarrow$ $\boldsymbol{H_0}$ 유지. 모든 표본평균이 가까이 모여있음.

일원 분산분석

- 다음과 같이 검정통계량을 만들어 본다:

$$\text{검정통계량} = \frac{\text{표본평균 사이의 분산 } (between)}{\text{표본내의 분산 } (within)} = \frac{MStr}{MSE}$$

⇒ 분산의 비율이므로 F확률분포를 따른다 ⇐ 우측검정.

⇒ 분자의 자유도 = 표본(처리)의 개수 - 1.

⇒ 분모의 자유도 = 전체 데이터의 개수 – 표본(처리)의 개수.

이원 분산분석

- **이원** 분산분석 : 한 개의 수치형 변수 Y와 두 개의 명목형 변수 X_A, X_B.

  예). 반 X_A과 성별 X_B로 그룹을 나누어서 성적 Y를 비교 검정한다.

- X_A에는 p개의 처리가 있고 X_B에는 q개의 처리가 있다고 가정한다.

이원 분산분석

- 귀무가설과 대립가설은 각각 A와 B로 구분한다.

$\Rightarrow$ "귀무가설" $\boldsymbol{H_0}$: $\mu(A_1) = \mu(A_2) = \cdots = \mu(A_p)$

"대립가설" $\boldsymbol{H_1}$: 적어도 하나의 모평균은 나머지와 다르다.

$\Rightarrow$ "귀무가설" $\boldsymbol{H_0}$: $\mu(B_1) = \mu(B_2) = \cdots = \mu(B_q)$

"대립가설" $\boldsymbol{H_1}$: 적어도 하나의 모평균은 나머지와 다르다.

이원 분산분석

- 마찬가지로 검정통계량도 각각 A와 B로 구분되며 F분포를 따른다.

$$검정통계량_{\boldsymbol{A}} = \frac{MStr_{\boldsymbol{A}}}{MSE} \sim F(p-1, (p-1)\cdot(q-1))$$

$$검정통계량_{\boldsymbol{B}} = \frac{MStr_{\boldsymbol{B}}}{MSE} \sim F(q-1, (p-1)\cdot(q-1))$$

이원 분산분석

- 이원 분산분석은 without replication (반복 없음)과 with replication(반복 있음)으로 세분화 할 수 있다:

$\Rightarrow$ 반복 없음: 유형별 단 한개의 자료값이 있음.

aov()에서 $Y \sim X_A + X_B$와 같은 수식을 사용.

$\Rightarrow$ 반복 있음: 유형별 여러개의 자료값이 있음.

aov()에서 $Y \sim X_A * X_B$와 같은 수식을 사용하여 상호작용 포함.

확률분포의 용도

키포인트

- 연속확률분포의 유형별 용도.
- 지금까지 배운 내용 정리.

연속확률분포의 유형별 용도

명칭	용도
연속균등분포	시뮬레이션. 모델링.
정규분포 (표준정규분포)	대표본 구간 추정. 대표본 평균 검정 (가설검정).
스튜던트 t 분포	대/소표본 구간 추정. 소표본 평균 검정 (가설검정). *선형회귀 계수 검정 (가설검정).* → *Later*
카이제곱 분포	도수표 검정 (가설검정). 분할표 검정 (가설검정). 분산검정 (가설 검정).
F 분포	분산비 검정 (가설검정). 분산분석, ANOVA (가설검정). *선형회귀식의 설명력 검정 (가설검정).* → *Later*

비모수검정

키포인트

- 부호검정 (한 개의 모집단 or 두 개의 모집단).
- Mann-Whitney U 검정.
- Wilcoxon 검정.
- Shapiro-Wilk 검정.

비모수 검정

- 다음과 같은 경우 비모수 검정에 해당한다.

  $\Rightarrow$ 모집단의 확률분포에 대한 전제가 없다.

  $\Rightarrow$ 표본의 크기가 매우 작다.

부호검정: 모집단 한 개

- 부호검정 (sign test)는 '+'라는 사건이 '-'라는 사건보다 더 많은지를 검정한다.

  예). 어느 회사의 월간 수익률이 '+'인 경우가 '–'인 경우보다 많은지 검정.

- 부호는 '+' 또는 '–' 이므로 이분법적인 상황이다. ⇒ 이항확률분포

- 그러므로 이항분포를 사용하여 다음 가설검정을 실행한다 (우측검정).

  귀무가설 $\boldsymbol{H_0} : p \leq 0.5$

  대립가설 $\boldsymbol{H_1} : p > 0.5$

  ⇒ 성공확률 ('+'의 확률) $p = 0.5$인 이항확률분포 ~*Binom(n,p)*를 사용한다.

부호검정: 모집단 한 개

문제: 시행의 횟수가 6인 경우 유의수준이 5%라면 표본에서 '+'의 개수가 몇 개 이상 이어야 부호 검정의 귀무가설을 기각할 수 있는가? (임계치)

부호검정: 모집단 한 개

문제: 시행의 횟수가 6인 경우 유의수준이 5%라면 표본에서 ‘+’의 개수가 몇 개 이상 이어야 부호 검정의 귀무가설을 기각할 수 있는가? (임계치)

$p = 0.5$를 전제하고 이항확률을 계산해 본다. ‘+’의 횟수를 Q로 표기한다.

$P(Q \geq 6) = P(Q = 6) = 0.0156$

$P(Q \geq 5) = P(Q = 5) + P(Q = 6) = 0.09375 + \ 0.0156 = 0.109375$

유의수준 5%에 해당하는 임계치는 5와 6 사이에 있다. 그러므로 ‘+’가 5개 까지는 귀무가설을 유지하고 ‘+’가 6개이면 귀무가설을 기각한다.

부호검정: 모집단 두 개

- 짝 지어진 두개의 관측치 사이에 양의 편향성이 있는지 검정하고자 한다.

  예). 동일한 환자 대상으로 약 투여 이후와 이전 사이의 차이에 대한 검정.

  예). 동일한 환자 대상으로 치료약 X와 Y 사이의 차이에 대한 검정.

- 두개의 확률변수 X와 Y가 있다고 가정하고, $p = P(X > Y)$라고 정의하고, 다음 가설을 검정한다 (우측검정).

  귀무가설 $\boldsymbol{H_0}$: $p \le 0.5$

  대립가설 $\boldsymbol{H_1}$: $p > 0.5$

부호검정: 모집단 두 개

문제: 신약 X와 기존의 약 Y를 비교하여 효과 지속시간에 대한 검정을 하려고 한다. 효과 지속시간의 차이 $X-Y$의 부호 검정을 환자 6명 대상 유의수준 5% 로 적용하고자 한다. 귀무가설의 유지 또는 기각 조건을 구하여라.

부호검정: 모집단 두 개

문제: 신약 X와 기존의 약 Y를 비교하여 효과 지속시간에 대한 검정을 하려고 한다. 효과 지속시간의 차이 $X - Y$의 부호 검정을 환자 6명 대상 유의수준 5% 로 적용하고자 한다. 귀무가설의 유지 또는 기각 조건을 구하여라.

$p = 0.5$를 전제하고 이항확률을 계산해 본다. 이제 Q는 $x_i - y_i$의 부호가 '+'인 횟수를 나타낸다. 이전 문제에서 구했듯이 $P(Q \geq 6) = 0.0156$이고 $P(Q \geq 5) = 0.109375$이다. 그러므로 '+'가 5개 까지는 귀무가설을 유지하고 '+'가 6개이면 귀무가설을 기각한다. 모집단 한 개인 경우와는 '+' 부호의 정의에서 차이가 있으니 주의한다.

Mann Whitney U 검정

- 두 집단 (A 와 B) 사이의 차이 여부를 검정하고자 한다.
- 다음의 경우 적용한다:

  ⇒ 자료가 서열척도로 표현되어 있다 (평균, 분산 등을 직접 계산 못함).

  ⇒ 표본의 크기가 30미만이어서 정규성을 충족시키지 못하는 경우.
- 귀무가설과 대립가설은 다음과 같다.

  귀무가설 $\boldsymbol{H_0}$: A의 순위 평균은 B의 것과 같다.

  대립가설 $\boldsymbol{H_1}$: A의 순위 평균은 B의 것과 다르다.

Mann Whitney U 검정

- 순위는 두 개의 표본을 통합하여 구한다.

예). 남학생의 시험점수: 100, 80

여학생의 시험점수: 95, 75, 68

$\Downarrow$

통합된 시험점수 {100, 95, 80, 75, 68}를 순위로 변환 {1, 2, 3, 4, 5}

$\Downarrow$

남학생의 시험점수 (순위척도): 1, 3

여학생의 시험점수 (순위척도): 2, 4, 5

남/여학생의 시험점수의 순위에서 차이가 있는가? $\Rightarrow$ Mann Whitney U 검정

Wilcoxon 검정

- Wilcoxon 검정을 부호화된 순위 검정 (signed rank test)라고도 부른다.
- 순위와 '+' or '-' 부호를 동시에 고려한다.
- 짝이 지어진 두 집단 (A 와 B) 간의 차이 여부를 검정하고자 한다.

귀무가설 $\boldsymbol{H_0}$: A와 B사이에는 차이가 없다.

대립가설 $\boldsymbol{H_1}$: A와 B사이에는 차이가 있다.

Wilcoxon 검정

- 부호화된 순위를 계산하는 방법을 다음 예를 통해서 알아본다.

예).

i	a_i	b_i	$a_i - b_i$	부호	$\lvert a_i - b_i \rvert$	$\lvert a_i - b_i \rvert$의 순위	부호화된 순위
#1	5.5	4.5	1	+	1	1	+1
#2	1	17	-16	-	16	6	-6
#3	10	6	4	+	4	5	+5
#4	16	14	2	+	2	2.5	+2.5
#5	13	11	2	+	2	2.5	+2.5
#6	8	5	3	+	3	4	+4
#7	9	9	0	없음	0	제외	제외

⇒ 이 경우 부호화 된 순위의 합은 9이다. 이것이 0과 의미 있는 차이가 있는지 검정한다.

Shapiro-Wilk 검정

- “정상성”을 검정하고자 한다. “정상성”은 정규분포를 따르는 성질을 의미.
- QQ plot과 같은 시각화 방법보다 선호된다.
- Shapiro-Wilk 검정에서 귀무가설과 대립가설은 다음과 같다.

  귀무가설 $\boldsymbol{H_0}$: 표본은 정규분포를 따르는 모집단에서 추출되었다.

  대립가설 $\boldsymbol{H_1}$: 표본은 정규분포를 따르지 않는 모집단에서 추출되었다.

끝