

For Big Data

- 기초 통계학 & 데이터 분석 방법론 2부 -

Chapter 10. 가설검정

Chapter 10 가설검정

◎ Intro

- 가설이란?
- 가설을 설정하는 방법은?
- 가설의 타당성 여부는 어떻게 결정되는가?
- 1종오류와 2종오류란?

Chapter 10 가설검정

◎ 가설검정의 개념

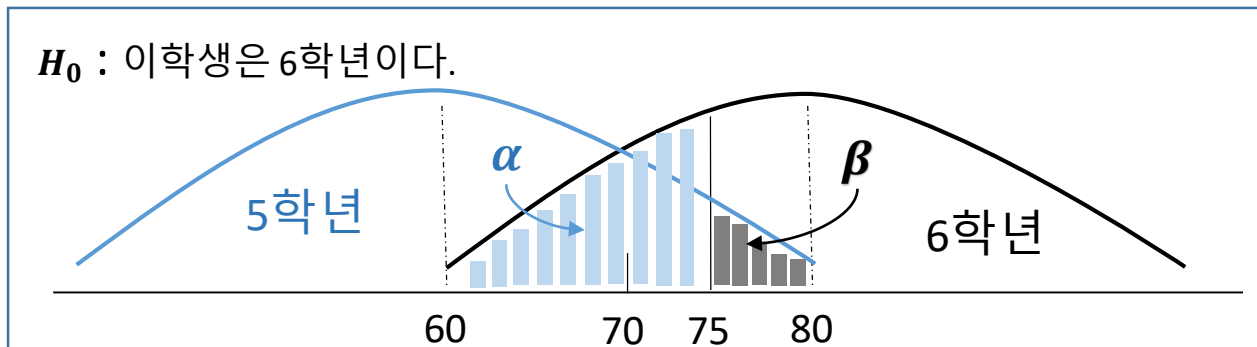
통계적 가설검정

표본에서 얻은 사실을 근거로 하여, 모집단에 대한 가설이 맞는지 틀리는지 통계적으로 검정하는 분석방법을 통계적 가설검정이라고 한다.

◎ 가설검정의 예

예제 10-1

어느 초등학교의 6학년과 5학년 학생들이 같은 문제를 가지고 시험을 본 결과, 성적의 분포가 아래 그림과 같았다. 6학년들의 평균 성적은 80점이었고, 5학년 학생들의 평균 성적은 60점이었으며, 두 분포가 정규분포라 가정하자. 만약 한 학생의 점수가 70점이라면, 해당 학생은 5학년일까? 6학년일까?



Chapter 10 가설검정

◎ 가설검정의 개념

- 예제1은 그 학생이 속해 있는 모집단의 평균을 얼마로 보아야 하는가 하는 문제와 마찬가지로. 만약 교사가 아래와 같은 가설을 세웠다고 하자.

“그 학생은 6학년이다”

- 그 교사는 이 가설이 틀릴 경우에 대비하여 다른 가설을 세웠다고 하면 아래와 같다.

“그 학생은 5학년이다”

- 그 학생이 5학년인지 6학년인지는 그 학생의 점수인 70점으로 판단하여야 함. 70점이 어느 학년에 속하는지를 결정하기 전에, 그 교사는 먼저 몇 점 이상을 6학년으로 보아야 하며 몇 점 이하를 5학년으로 보아야 좋겠는지를 결정하여야 함
- 그가 어떤 이유에서 75점 이상을 6학년으로 간주한다면, 70점을 받은 학생에 대한 의사결정에서 ‘**그 학생은 6학년이다**’라는 가설은 기각되고, 대립적으로 설정한 **“그 학생은 5학년이다”라는 가설이 채택됨**
- 가장 큰 Risk는 어떠한 가설을 설정할 때에 따르는 오류의 위험이다. α (알파)의 부분만큼은 6학년에 속하면서도 5학년이라는 오해를 받게 되고, β (베타)로 표시된 부분이 “그 학생은 6학년이다”라고 가설을 설정할 때 그 가설이 틀렸음에도 불구하고 옳은 것으로 간주할 오류임

Chapter 10 가설검정

◎ 가설검정의 기본용어

- 가설의 설정은 확신에 근거를 두고 이루어지는 것이 아니며, 단지 후에 경험적 또는 논리적으로 검정될 수 있는 조건, 원리 또는 명제(proposition)을 제시하는 것에 불과함
- 귀무가설(null hypothesis) : 검정의 대상이 되는 가설
- 대립가설(alternative hypothesis) : 귀무가설이 받아들여질 수 없을 때 대신 받아들여지는 가설

귀무가설과 대립가설

귀무가설은 직접 검정대상이 되는 가설을 말하며, H_0 로 표시한다. 대립가설은 귀무가설이 기각될 때 받아들여지는 가설로서 H_1 로 표시한다

- 예제 1에서는 아래와 같이 가설을 설정할 수 있음

H_0 : 그 학생은 6학년이다.

H_1 : 그 학생은 5학년이다.

- 예제 1의 가설에서는 귀무가설과 대립가설을 서로 바꿀 수 있다. 즉 "그 학생은 5학년이다"를 귀무가설(H_0)로, "그 학생은 6학년이다"를 대립가설(H_1)로 할수도 있다.
- 하지만 다음페이지에서 보여줄 다른 귀무가설은 이야기가 다르다.

Chapter 10 가설검정

◎ 가설검정

하지만 다음 2)의 사례를 보자

H_0 : 주당 평균 전력소비량이 60kw이다 즉, $\mu = 60\text{kw}$

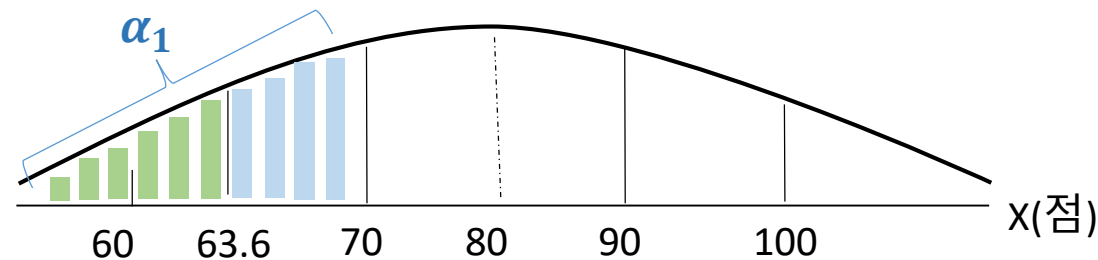
H_1 : 주당 평균 전력소비량이 60kw가 아니다 즉, $\mu \neq 60\text{kw}$

- 2)의 가설은 귀무가설과 대립가설을 서로 바꾸기 어려움. 그 이유는 귀무가설은 실제 검정대상이 되는 가설이며, 대립가설은 검정대상이 되지 않고 귀무가설이 거부될 때 자동적으로 받아들여지는 가설이기 때문이므로 실제로 검정할 수 없거나 검정하기에 곤란한 가설을 귀무가설로 설정하는 것은 바람직하지 않다
- 위의 2)번 사례에서 $\mu \neq 60\text{kw}$ 를 귀무가설로 설정하였을 경우 $\mu \neq 60\text{kw}$ 인 모집단이 무수히 많으므로 이를 현실적으로 검정하는데에는 매우 많은 어려움이 따른다.
- 따라서 $\mu = 60\text{kw}$ 로 귀무가설을 설정하는 것이 검정에 용이하다.

Chapter 10 가설검정

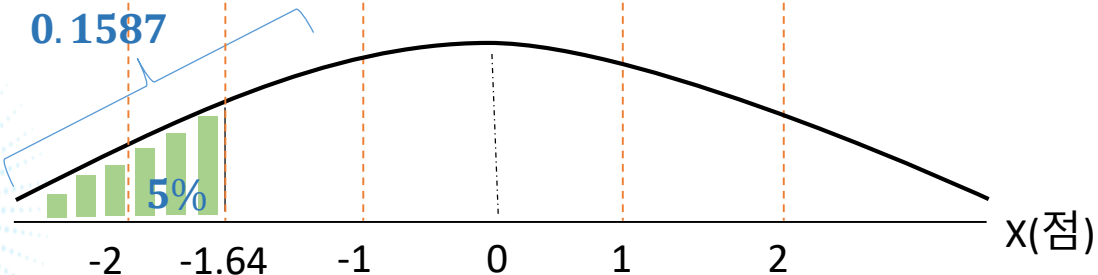
◎ 유의수준과 임계값

- 표본에서 계산된 통계량이 가설로 설정된 모집단의 성격과 현저한(significant) 차이가 있는 경우에는 모집단에 대해 설정한 귀무가설을 기각하게 됨
- 모집단에 대해 설정한 가설을 채택 또는 기각하는 임계값(critical value)이 어디가 되어야 하는지가 그 문제
- 앞 선 예제에서 5학년의 평균이 60점이며, 6학년의 평균이 80점일 때 어떤 점수에서부터 6학년으로 보아야 하는가?
- 아래의 그림을 통해서 70점 이상을 6학년으로 보게 된다면, 6학년 중에서 5학년으로 간주될 학생은 α_1 만큼 되며 이 비율은 0.1587이 됨 그러나 15.87% 라는 오류는 매우 크기에 오류를 줄이기 위해 α 를 5%로 한다면, 이 z값은 -1.64가 되며 이에 대응하는 x는 63.6이 됨



$$z = \frac{X - \mu}{\sigma} = \frac{X - 80}{10} = -1.64$$

$$X = 63.6$$



Chapter 10 가설검정

◎ 유의수준과 임계값

즉, 63.6점을 기준으로 하여 그 점수 이상의 학생을 4학년이라고 한다면, 실제로는 6학년이면서도 5학년으로 분류되는 오류가 발생할 확률은 약 5%가 된다.

다시 말해 5%의 오류를 감수할 때 4학년의 분포와 현저하게 차이가 볼 수 있다는 기준값(임계치 : Critical Value)는 63.6이 되며 이에 해당하는 Z의 값은 -1.64가 되는 것이다.

이 **임계치**를 기준으로 귀무가설의 **기각영역(rejection area)**와 **채택영역(acceptance area)**가 결정된다.

- 유의수준(significance level) : 위의 예에서의 5%, 15.87%(만약 70점 이상을 6학년으로 볼 때의 오류) 등의 오류 가능성

임계값

임계값이란 주어진 유의수준에서 귀무가설의 채택과 기각에 관련된 의사결정을 할 때, 그 기준이 되는 점이다.

그러면 실제 연구에서 유의수준, 즉 오류를 감수할 확률을 얼마로 결정하여야 하는가?
이에 대한 해답은 연구의 성격, 연구자의 주관 등이 개입하게 되므로 어느 연구에나 적용될 수 있는 보편타당한 기준은 없다.

그러나 보통 연구에서는 **α 수준을 0.01, 0.05, 0.10** 등으로 정하는 경우가 많다
(유의수준은 뒤에 설명될 **α -오류와 동일한 것을 의미**한다.)

Chapter 10 가설검정

◎ 양측검정과 단측검정

- 귀무가설이 기각되면 대립가설이 채택된다. 그런데 우리가 알고 있지 못하는 모집단의 특성, 즉 **모수(parameter)에 대한 가설검정**을 할 때에는 다음과 같이 두 가지로 **귀무가설 :: H_0** 과 **대립가설 :: H_1** 을 나타낼 수 있다. 첫째는 모수, 예를 들면 μ 가 어떤 수와 꼭 같다는 가설이며, 다른 하나는 모수 μ 가 어떤 수보다 크거나 또는 작다고 하는 가설이다.

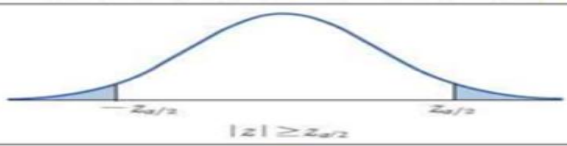
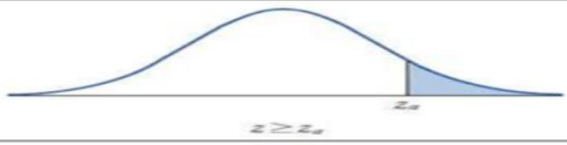
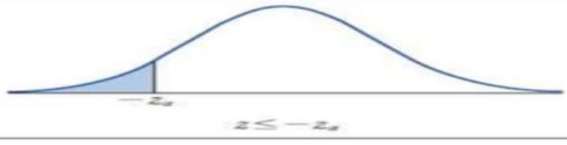
$$(1) \begin{aligned} H_0 &: \mu = q \\ H_1 &: \mu \neq q \end{aligned}$$

$$(2) \begin{aligned} H_0 &: \mu \geq q & \text{또는} & H_0 : \mu \leq q \\ H_1 &: \mu < q & & H_1 : \mu > q \end{aligned}$$

- (1)과 같이 귀무가설이 $\mu = q$ 로, 대립가설이 $\mu \neq q$ 로 설정되어 있는 경우를 생각해본다면, 이때 표본을 뽑아서 그 표본에서 얻은 통계량이 $\mu = q$ 과 매우 근접하여 있으면 귀무가설을 채택할 것이나, 그렇지 않고 검정을 위한 통계량이 q 보다 매우 크거나, 또는 q 보다 현저히 작을 때에는 귀무가설을 채택할 수 없게 됨
- 따라서 귀무가설을 기각하는 영역은 확률분포의 양측에 있게 됨. 가설검정에서 기각영역이 양쪽에 있는 것을 양측검정(two-tailed test)라 함
- 또한 유의수준 α 도 양쪽 극단으로 갈리게 되어 한쪽의 면적이 $\alpha/2$ 가 됨

Chapter 10 가설검정

◎ 양측검정과 단측검정

| 검정의 종류 | 귀무가설과 대립가설 | 기각역 (색칠한 부분의 가로축 좌표) |
|--------|--|---|
| 양측검정 | $H_0: \mu = \mu_0$ 대 $H_1: \mu \neq \mu_0$ |  $ z \geq z_{\alpha/2}$ |
| 단측검정 | $H_0: \mu = \mu_0$ 대 $H_1: \mu > \mu_0$ |  $z \geq z_{\alpha}$ |
| | $H_0: \mu = \mu_0$ 대 $H_1: \mu < \mu_0$ |  $z \leq -z_{\alpha}$ |

$$(2) \begin{array}{ll} H_0: \mu \geq q & \text{또는 } H_0: \mu \leq q \\ H_1: \mu < q & \quad \quad H_1: \mu > q \end{array}$$

- 한편 (2)와 같이 귀무가설을 $\mu \geq q$ 라 하고 대립가설을 $\mu < q$ 라고 설정하여 가설검정할 때에는, 선택된 표본의 통계량이 q 보다 현저히 작지 않으면 귀무가설을 채택하게 됨
- 따라서 확률분포의 오른쪽 극단에는 귀무가설의 기각역이 없음!
- 다만, 통계량이 q 보다 현저히 작으면 귀무가설을 기각하면 됨. 위의 그림은 이러한 사실을 나타내주고 있음
이렇게 가설검정에서 기각영역이 어느 한쪽에만 있게 되는 경우를 단측검정(one-tailed test)이라 함

Chapter 10 가설검정

◎ 양측검정과 단측검정

요약

(1)과 같이 귀무가설(H_0)이 $\mu = q$ 로, 대립가설(H_a)이 : $\mu \neq q$ 로 설정되어 있는 경우를 생각해보자.

이 때 표본을 뽑아서 그 표본에서 얻은 통계량이 귀무가설 $\mu = q$ 과 매우 근접하여 있으면 귀무가설이 채택될 것이나 **검정을 위한 통계량**이 (이하 **검정통계량**) q 보다 매우 크거나, 또는 q 보다 현저히 작을 때에는 **귀무가설**을 채택할 수 없게 된다

따라서 귀무가설을 기각하는 영역은 확률분포의 양측에 있게 되는데, 이처럼 가설검정에서 기각영역이 양쪽에 있는 것을 **양측검정(two-tailed test)**이라 한다

그러므로 유의수준 α 도 양쪽 극단으로 갈리게 되어 한쪽의 면적이 $\alpha/2$ 가 된다.

반면 (2)와 같이 귀무가설을 $\mu \geq q$ 라 하고 대립가설을 $\mu < q$ 라고 설정하여 가설검정할 때에는, 선택된 표본의 통계량이 q 보다 현저히 작으면 귀무가설을 채택하게 된다. 따라서 확률분포의 오른쪽 극단에는 기각역이 없다. 다만 통계량이 q 보다 현저히 작을 때에만 귀무가설을 기각하게 된다. 따라서 α 로 나타내는 기각영역은 나타내는 기각영역은 분포의 왼쪽 극단에만 존재하게 된다.

반대로 귀무가설을 $\mu \leq q$ 라 하고 대립가설을 $\mu > q$ 로 할 때에는, 위에서 설명한 것과 반대의 현상이 나타난다. 즉, 통계량이 q 보다 현저히 클 때에만 귀무가설을 기각하게 되므로, 기각영역은 오른쪽에만 있게 된다. 이렇게 2가지 경우와 같이 기각영역이 어느 한쪽에만 있게 되는 경우를 **단측검정(one-tailed test)**라 한다.

Chapter 10 가설검정

◎ 가설검정의 오류

- 가설검정은 표본에서 뽑은 통계량을 기초로 하여 모집단의 특성을 알아보려고 하는 것이기 때문에 표본이 어떻게 선택되느냐에 따라 잘못된 결론을 내릴 수 있음
- 표집오차(sampling error)는 언제나 발생하기에, 표본에 근거를 둔 가설검정에서도 늘 오류가 뒤따름
- 가설검정에 따르는 오류는 두 가지로 나눌 수 있는데 :: α -오류 며, 다른 하나는 β -오류임

가설검정의 오류

α -오류는 실제로는 귀무가설이 옳은데도 검정 결과 귀무가설을 기각하는 오류를 말한다. α -오류는 제1종오류(type I error)라고도 한다.

β -오류는 실제로는 귀무가설이 틀렸는데도 검정 결과 귀무가설이 옳은 것으로 받아들이는 오류를 말한다. β -오류는 제2종오류(type II error)라고도 한다

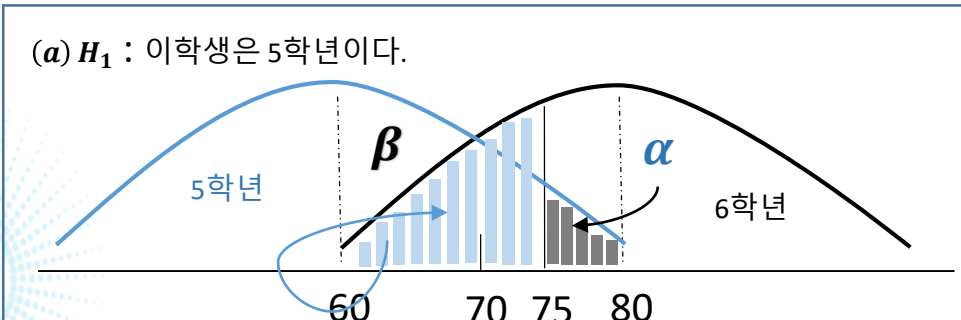
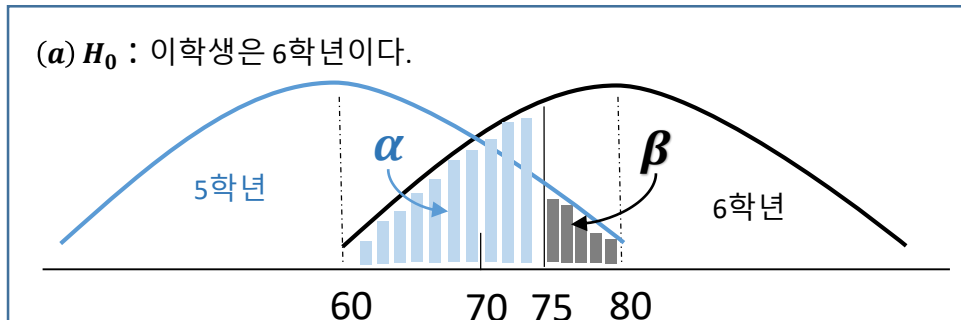
| | H_0 가 맞을 경우 | H_0 가 틀릴 경우 |
|----------|----------------------|---------------------|
| H_0 채택 | $1 - \alpha$ (옳은 결정) | β -오류(2종오류) |
| H_0 기각 | α -오류(1종오류) | $1 - \beta$ (옳은 결정) |

- α 와 β 의 크기는 서로 반대방향으로 변하고 있으므로, $1 - \alpha$ (옳은결정)와 $1 - \beta$ (옳은결정)를 동시에 크게 하기는 현실적으로 불가능
- 두 가지 오류들 중 '1종 오류'가 '2종 오류'에 비해 훨씬 중요하게 여겨짐

Chapter 10 가설검정

◎ 가설검정의 오류

- 앞선 점수를 판단하는 예제에서 어느 선생님이 “이 학생은 6학년이다”라는 귀무가설(H_0)을 세우고 임계값을 **75점으로 결정**했다고 하면, 학생들이 실제로는 6학년임에도 불구하고 5학년으로 잘못 판단할 오류, 즉 **귀무가설이 옳은데도 불구하고 귀무가설(H_0)을 기각할 오류**는 아래 그림에서의 α 의 면적과 같으며, 반대로 5학년을 6학년으로 잘못 판단할 오류는 β 만큼 된다.
- 반대로 “이 학생은 5학년이다”를 귀무가설(H_0)로 설정했을 때에는 α 와 β 의 위치를 서로 바꾸면 된다.



- 전체를 1이라 하면 귀무가설이 맞을 때 귀무가설을 기각하는 오류가 α 이므로 맞는 귀무가설을 올바르게 채택하는 경우는 $1 - \alpha$ 가 됨
- 또한 귀무가설이 틀릴 때 귀무가설을 받아들이는 오류는 β 이므로 틀린 귀무가설을 올바르게 거부하는 경우는 $1 - \beta$ 가 됨
- 결론 :: $1 - \alpha$ 와 $1 - \beta$ 를 크게 할수록 옳은 결정!!!

Chapter 10 가설검정

◎ 가설검정의 오류

가설검정의 오류

α -오류는 실제로는 귀무가설이 옳은데도 검정 결과 귀무가설을 기각하는 오류를 말한다. α -오류는 **제1종오류(type I error)**라고도 한다.

β -오류는 실제로는 귀무가설이 틀렸는데도 검정 결과 귀무가설이 옳은 것으로 받아들이는 오류를 말한다. β -오류는 **제2종오류(type II error)**라고도 한다

- α 와 β 의 크기는 서로 반대방향으로 변하고 있으므로, $1-\alpha$ (옳은 결정)와 $1-\beta$ (옳은 결정)를 동시에 크게 하기는 현실적으로 불가능
- 두 가지 오류들 중 '**1종 오류**'가 '**2종 오류**'에 비해 훨씬 중요하게 여겨짐
- 따라서 1종오류를 미리 1% 또는 5% 정도의 매우 작은 값으로 제한시킬 필요성 존재 → 유의수준 0.01, 0.05 혹은 0.10
- 흔히, 유의수준 α 의 검정법이란 제1종오류를 점할 확률이 α 이하 라는 것을 의미

| | H ₀ 가 맞을 경우 | H ₀ 가 틀릴 경우 |
|-------------------|----------------------------------|------------------------|
| H ₀ 채택 | 중요 1 - α (옳은결정) | β -오류(2종오류) |
| H ₀ 기각 | α -오류(1종오류) | 1 - β (옳은결정) |

Chapter 10 가설검정

◎ 가설검정의 오류

가설검정의 순서

- 1) 귀무가설과 대립가설의 설정
- 2) 유의수준의 결정
- 3) 유의수준을 충족시키는 임계값의 결정
- 4) 통계량의 계산과 임계값과의 비교
- 5) 결과의 해석

예제 10-3

국내 아이돌그룹 멤버들의 평균 키를 알기 위해 16명의 아이돌그룹멤버의 키를 표본조사하였더니 평균 키가 175cm였다. 국내 아이돌 그룹 전체의 평균 키에 대한 표준편차가 5cm라고 하면, 국내 아이돌그룹 멤버의 평균 키가 180cm 이상이라고 할 수 있을까? 유의수준(α)를 5%로 하여 검정하시오.

$$(1) H_0 : \mu \geq 180 \text{ cm} \\ H_1 : \mu < 180 \text{ cm}$$

$$(2) \alpha = 5\%$$

$$(3) \text{채택영역} : Z \geq -1.64 \\ \text{기각영역} : Z < -1.64$$

$$(4) 175\text{cm에 대응하는 } Z\text{값} : z = \frac{X - \mu}{\sigma} = \frac{175 - 180}{5 / \sqrt{16}} = \frac{-5}{1.25} = -4$$

(5) $Z = -4$ 는 -1.64 보다 작아서 기각영역에 속하므로 H_0 를 기각한다.

위의 결과를 토대로 국내 아이돌그룹 멤버들의 평균 키가 180cm 이상이라고 할 수 없다.


Chapter 11. 단일모집단 가설검정

Chapter 11 단일모집단에 관한 가설검정

◎ 가설검정의 예

예제 11-1

우리나라 여성 전체의 평균 키는 160cm이고, 분산은 200이라고 한다. 10,000명을 표본으로 하여 조사한 결과 평균 169cm를 얻었다. 우리나라 여성의 평균 키가 160cm라 할 수 있을까?



위의 예제가 모집단 평균에 관한 가설검정의 가장 기본적인 형태다. 그러나 예문을 자세히 읽어보면 의문점이 생긴다. 즉, “여성 전체의 키의 분산이 $\sigma^2 = 200$ ” 이라는 것이 바로 그것이다. 모집단의 분산 또는 표준편차는 **모집단 평균을 모르고는 계산할 수가 없는데, 해당 문제에서는 모집단의 분산을 이미 아는 것으로 가정하였다**. 모집단 평균을 모르기 때문에 가설검정을 하려는 것인데, 모집단 분산을 알고 있다고 가정하는 것은 모순적이다. 따라서 위의 문제는

아래와 같이 **변화되어야 한다**.

예제 11-2

우리나라 여성의 평균 키는 160cm라고 한다. 이에 대한 가설을 검정하기 위하여 10,000명을 표본으로 하여 조사한 결과, 평균 169cm, 분산 300을 얻었다. 우리나라 여성의 평균 키가 160cm라 할 수 있을까?

분석 시 모집단 분산을 알고있는 경우와 예제 11-2와 같이 모집단 분산을 알지 못하는 경우로 나뉘게 된다.

- 모집단 분산을 모르는 경우 - t분포를 활용
- 모집단 분산을 아는 경우 - z분포를 활용

Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 평균에 대한 가설검정 순서 - 모집단 분산을 아는 경우

귀무가설과 대립가설

- 두 귀무가설은 부등식보다는 등식으로 표현되는 것이 더욱 바람직. 그 이유는 쉽게 가설을 검정할 수 있는 형태가 더욱 바람직하기 때문임

$$(1) H_0 : \mu = 160$$

$$H_1 : \mu \neq 160, \text{ 또는 } \mu > 160, \text{ 또는 } \mu < 160$$

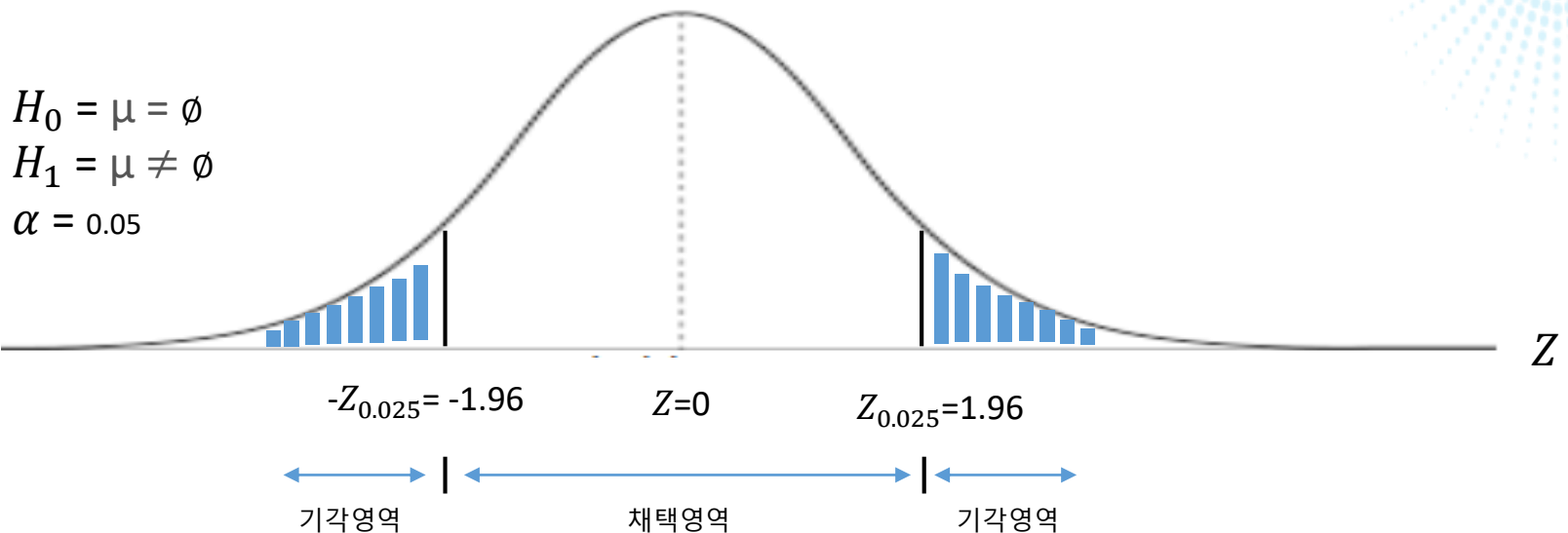
유의수준의 결정

- 유의수준의 결정은 연구자의 연구목적, 주관적인 판단 등에 따라 달라지나 대체로 0.01, 0.05, 0.10 등으로 정하는 경우가 보통임
- 예를 들어 의약품의 성분용량을 검정할 때에는 엄격하게 하여야 하기 때문에, 귀무가설이 맞는데도 불구하고 기각할 확률인 α -오류(1종오류), **즉 유의수준을 크게 하여 검정함**으로써 **불량품이 발생할 위험을 줄여야 한다**

채택영역과 기각영역 :: 임계값

- 이미 결정된 유의수준(significance level)을 충족시키는 임계값은 \bar{x} 로 표시할 수 있으며, 또한 \bar{x} 에 대응하는 Z값으로도 표시가 가능
- 즉, '표본의 평균 \bar{x} 가 170이상 또는 150이하일때 모집단 평균이 160이라는 귀무가설을 기각한다'는 기준 또는 '표본의 평균 \bar{x} 에 대응하는 Z값이 +1.96 이상 또는 -1.96 이하이면 모집단 평균에 대한 귀무가설을 기각한다' 등의 기준도 세울 수 있음
- Z값을 이용하는 것이 편리하기에 대부분 Z값을 임계값으로 사용하여 귀무가설의 채택영역과 기각영역을 정함

Chapter 11 단일모집단에 관한 가설검정

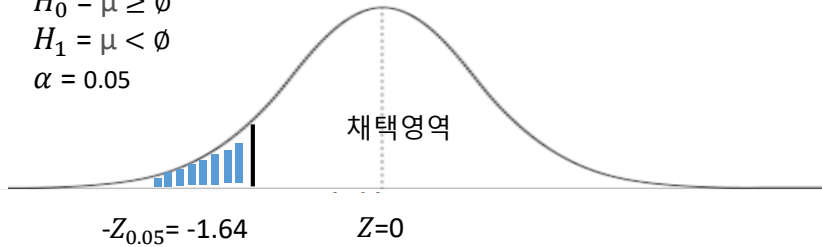


양측검정에서의 임계값

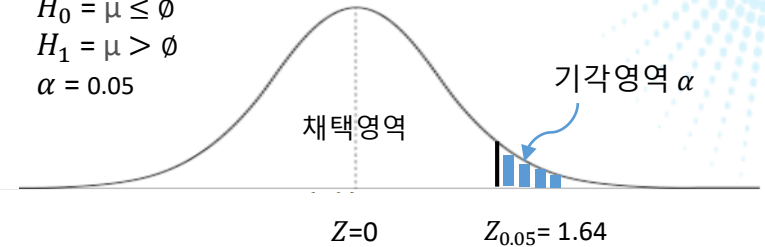
- 먼저 양측검정에서 임계값이 어떻게 결정되는지 살펴보기
- 유의수준이 α 일 때, 양측검정에서는 오른쪽의 임계값은 $z_{\alpha/2}$ 가 되며, 왼쪽의 임계값은 $-z_{\alpha/2}$
- 유의수준 α 가 0.05인 예라면 $-z_{0.025}$ 는 -1.96이며, $z_{0.025}$ 는 1.96임

Chapter 11 단일모집단에 관한 가설검정

$$\begin{aligned} H_0 &= \mu \geq \emptyset \\ H_1 &= \mu < \emptyset \\ \alpha &= 0.05 \end{aligned}$$



$$\begin{aligned} H_0 &= \mu \leq \emptyset \\ H_1 &= \mu > \emptyset \\ \alpha &= 0.05 \end{aligned}$$



단측검정에서의 임계값

- 단측검정에서 임계값을 z 로 표시할 때, 유의수준을 α 로 한다면, 임계값은 $-Z_\alpha$ 나 Z_α 가 됨
- 즉, 대립가설이 $\mu < \emptyset$ 면 $-Z_\alpha$ 가 임계값이 되며, $\mu > \emptyset$ 가 대립가설이 될 때에는 Z_α 가 임계값이 됨
- 오른쪽단측검정 또는 왼쪽단측검정인지에 따라 $-Z_{0.05} = -1.64$ 나 $Z_{0.05} = 1.64$ 가 됨

| α | 양측검정 | | 단측검정 | |
|----------|-----------------|----------------|-------------|------------|
| | $-Z_{\alpha/2}$ | $Z_{\alpha/2}$ | $-Z_\alpha$ | Z_α |
| 0.01 | -2.57 | 2.57 | -2.33 | 2.33 |
| 0.05 | -1.96 | 1.96 | -1.64 | 1.64 |
| 0.10 | -1.64 | 1.64 | -1.28 | 1.28 |

Chapter 11 단일모집단에 관한 가설검정

통계량의 계산과 임계값과의 비교

- 임계값은 \bar{x} 를 사용해 표시할 수도 있으나 z 값으로 나타내는 것이 더욱 편리
- 그러므로 표본에서 얻은 \bar{x} 도 z 값으로 바꾸어야 비교가 가능
- 표본을 기초로 계산된 임계값을 비교해서, z 값이 기각영역 안에 있으면 H_0 를 기각, 채택영역 안에 있으면 H_0 를 채택

z -통계량

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

결과의 해석

- 최종적으로 귀무가설을 기각하거나 채택하는 것이 무엇을 의미하는가를 해석하는 것이 중요
- 예를 들어 우리나라 여성의 평균 키가 160cm라는 가설 검정을 할 때, "귀무가설을 기각한다."라는 표현보다는 "귀무가설을 기각하므로 우리나라 여성의 평균키가 160cm라고 할 수 없다"라고 해석을 해야만 연구자의 임무를 다하는 것

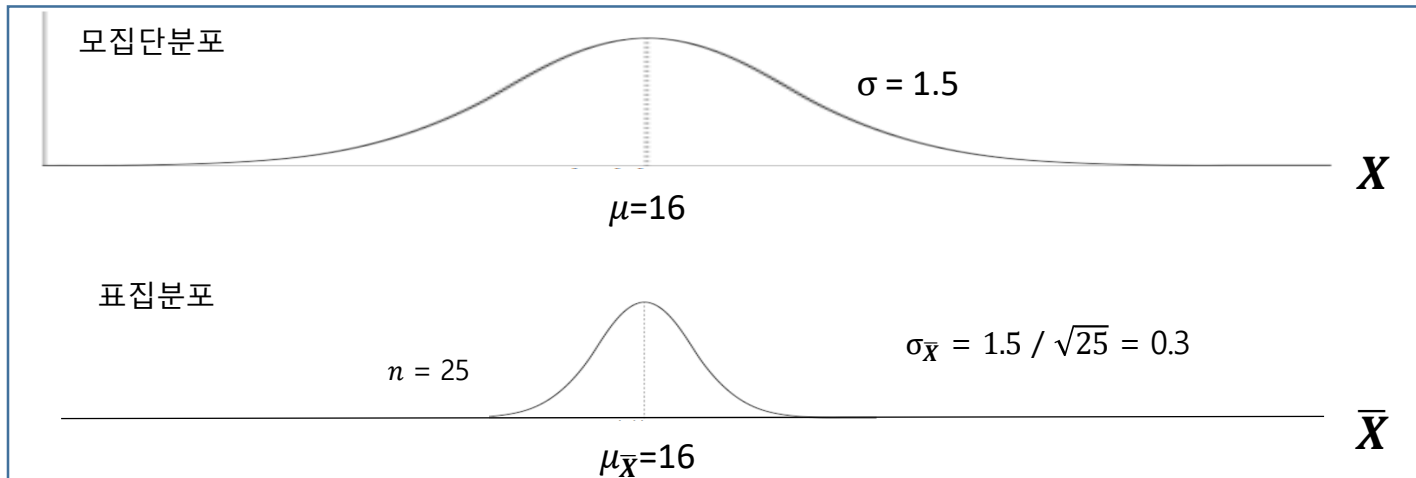
Chapter 11 단일모집단에 관한 가설검정

◎ 모집단의 분산을 알고 있을 때

예제 11-3

통조림회사에서 수출용 참치 통조림을 생산하는데, 그 통조림의 무게가 16온스며, 무게의 분포가 정규분포라 한다. 그러나 해외에서는 통조림 무게가 16온스가 아니라는 불평이 들려온다고 하자. 회사측에서는 이를 확인하기 위해 25개의 통조림을 표본으로 뽑아 평균을 조사하여 본 결과, $\bar{X} = 15.5$ 온스였다. 모집단의 표준편차는 1.5온스라는 것을 과거의 경험으로 알고 있다고 하자. $\alpha = 0.05$ 로 하면, 위의 결과로부터 이 회사의 통조림 무게가 16온스라고 말할 수 있을까?

- A. 무게가(=모집단) 정규분포이므로 $n = 25$ 일 때의 표집분포도 정규분포이며, $\mu_{\bar{X}} = \mu$ 이고 $\sigma_{\bar{X}} = \sigma / \sqrt{n}$
즉, $\mu_{\bar{X}} = 16$ 온스이고 $\sigma_{\bar{X}} = 1.5 / \sqrt{25} = 0.3$



Chapter 11 단일모집단에 관한 가설검정

◎ 모집단의 분산을 알고 있을 때

예제 11-3 - 양측검정

① $H_0 : \mu = 16$
 $H_1 : \mu \neq 16$

② $\alpha = 0.05$

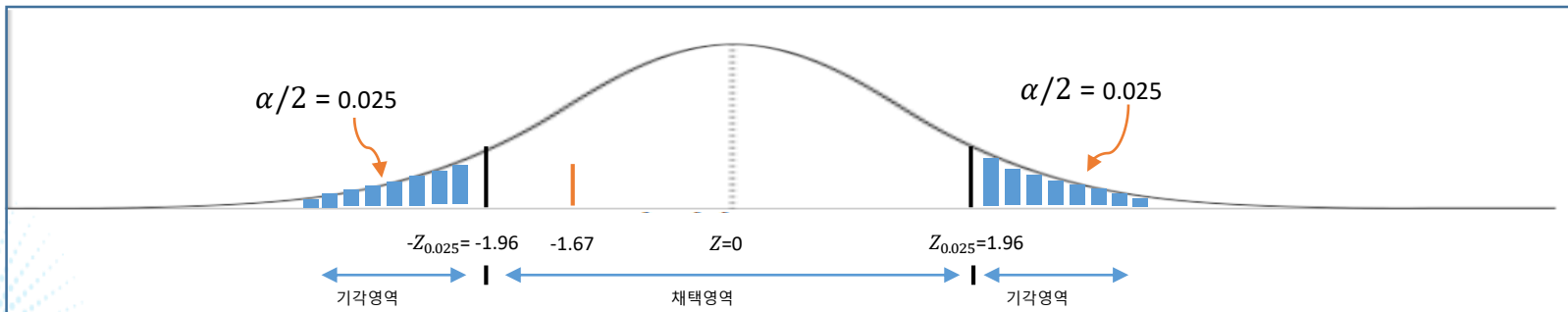
③ 채택영역 : $-1.96 \leq Z \leq 1.96$
기각영역 : $Z > 1.96$ 또는 $Z < -1.96$

④ $\bar{X} = 15.5$ 에 해당하는 z값을 계산하면 아래와 같다.

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{15.5 - 16.0}{0.3} = -1.67$$

$Z = -1.67$ 은 **채택영역**안에 있으므로 H_0 를 기각할 수 없음

⑤ $\alpha = 0.05$ 수준에서 통조림의 무게가 16온스라는 기존의 주장을 기각할 수 없음



Chapter 11 단일모집단에 관한 가설검정

◎ 모집단의 분산을 알고 있을 때

예제 11-3 - 단측검정

① $H_0 : \mu = 16$
 $H_1 : \mu < 16$

② $\alpha = 0.05$

③ 채택영역 : $Z \geq -1.64$
기각영역 : $Z < -1.64$

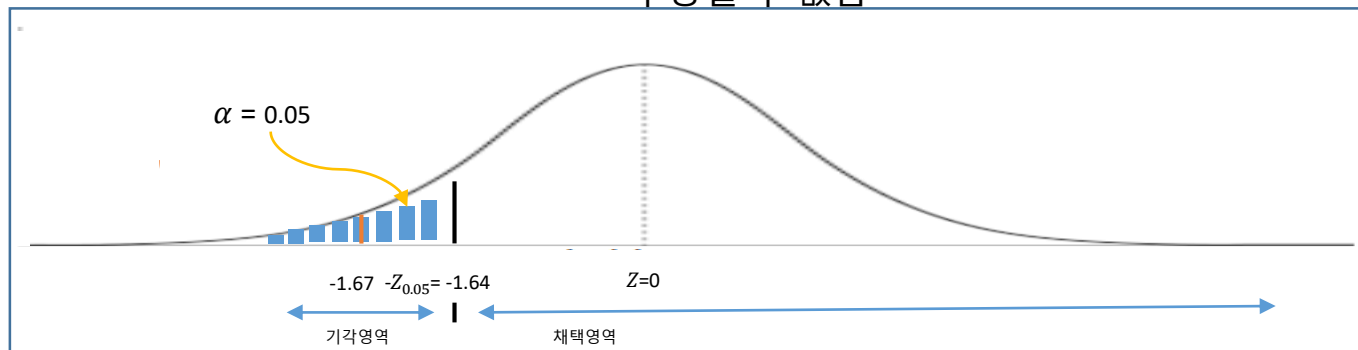
④ $\bar{X} = 15.5$ 에 해당하는 z값을 계산하면 아래와 같다.

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{15.5 - 16.0}{0.3} = -1.67$$

$Z = -1.67$ 은 기각영역안에 있으므로 H_0 는 기각된다

⑤ 양측검정의 결과와는 달리 내용량이 16온스라고 고객들에게

주장할 수 없음



Chapter 11 단일모집단에 관한 가설검정

◎ 모집단의 분산을 모를 때

t -분포와 임계값

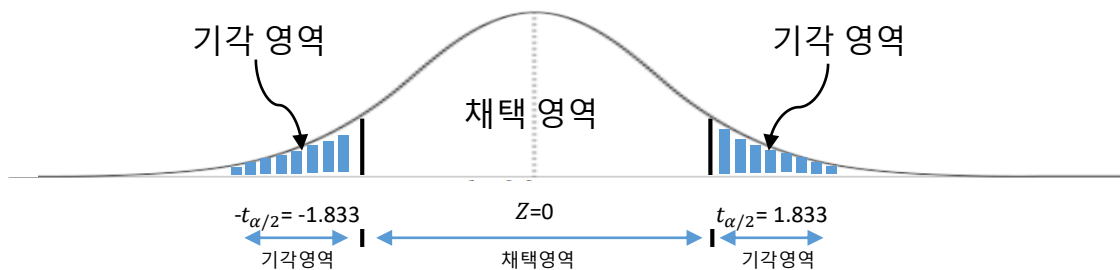
- 모집단의 분산을 모를 때 표본에서 구한 불편 추정량인 s^2 또는 s 를 모집단의 σ^2 의 분산 또는 σ 를 대신해 사용
- 모집단의 분산을 모를 때, t -분포를 활용
- 통계량 t -분포는 자유도 $n-1$ 의 t -분포를 이룸 (chapter 9를 참고해보자!)
- 즉, 왼쪽 임계값은 $-t_{\alpha/2, n-1}$, 오른쪽은 $t_{\alpha/2, n-1}$ 이 됨
- $\alpha = 0.1$ 이고 자유도(df)가 9라면, 양측검정에서의 임계치는 $t = \pm 1.833$ 임
- 단측검정에 있어서는 대립가설이 $\mu < q$ 인지, $\mu > q$ 인지에 따라 임계치가 달라짐
 $\alpha = 0.1$ 이고 자유도(df)가 9라면, 단측검정에서의 임계치는 $-t_{\alpha, n-1}$ 또는 $t_{\alpha, n-1}$
 $\alpha = 0.1$ 이고 자유도(df)가 9라면, 단측검정에서의 임계치는 $t = \pm 1.383$ 임

Chapter 11 단일모집단에 관한 가설검정

◎ 모집단의 분산을 모를 때

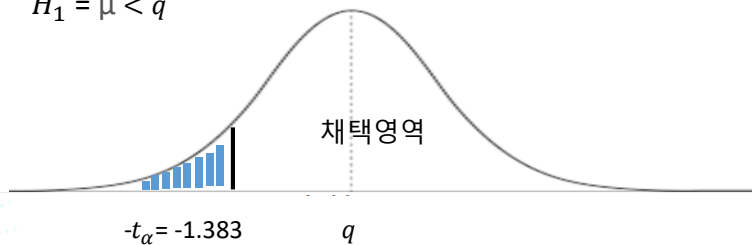
t -분포와
임계값

t 의 임계값 : 양측검정

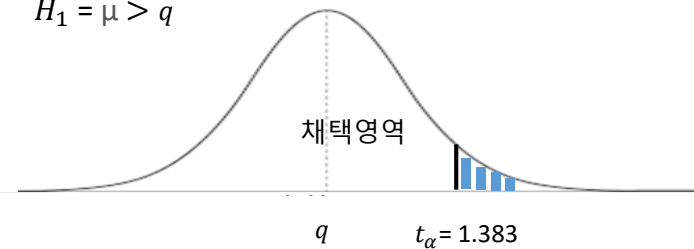


t 의 임계값 : 단측검정

$$H_1 = \mu < q$$



$$H_1 = \mu > q$$



Chapter 11 단일모집단에 관한 가설검정

◎ t -분포표

| α df | 0.4 | 0.25 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0025 | 0.001 | 0.0005 |
|----------------|-------|-------|--------------|--------------|--------|--------|--------|--------|--------|--------|
| 1 | 0.325 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 127.32 | 318.31 | 636.62 |
| 2 | 0.289 | 0.816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 14.089 | 22.327 | 31.599 |
| 3 | 0.277 | 0.765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 7.453 | 10.215 | 12.924 |
| 4 | 0.271 | 0.741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | 0.267 | 0.727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | 0.265 | 0.718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | 0.263 | 0.711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | 0.262 | 0.706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | 0.261 | 0.703 | <u>1.383</u> | <u>1.833</u> | 2.262 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | 0.260 | 0.700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | 0.260 | 0.697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | 0.259 | 0.695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | 0.259 | 0.694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | 0.258 | 0.692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | 0.258 | 0.691 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | 0.258 | 0.690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | 0.257 | 0.689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | 0.257 | 0.688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.197 | 3.610 | 3.922 |
| 19 | 0.257 | 0.688 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | 0.257 | 0.687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| 21 | 0.257 | 0.686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| 22 | 0.256 | 0.686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |
| 23 | 0.256 | 0.685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.104 | 3.485 | 3.768 |
| 24 | 0.256 | 0.685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.091 | 3.467 | 3.745 |
| 25 | 0.256 | 0.684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.078 | 3.450 | 3.725 |
| 26 | 0.256 | 0.684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.067 | 3.435 | 3.707 |
| 27 | 0.256 | 0.684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.057 | 3.421 | 3.690 |
| 28 | 0.256 | 0.683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.047 | 3.408 | 3.674 |
| 29 | 0.256 | 0.683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.038 | 3.396 | 3.659 |
| 30 | 0.256 | 0.683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.030 | 3.385 | 3.646 |
| 40 | 0.255 | 0.681 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 2.971 | 3.307 | 3.551 |
| 60 | 0.254 | 0.679 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 2.915 | 3.232 | 3.460 |
| 120 | 0.254 | 0.677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 2.860 | 3.160 | 3.373 |
| ∞ | 0.253 | 0.674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 2.807 | 3.090 | 3.291 |

Chapter 11 단일모집단에 관한 가설검정

t -검정의 계산

- t -통계량을 계산하는 방법은 제 9장에서 이미 언급되었지만, 다시금 복기해보자.
- 유의수준에 따라 임계치가 결정되면 표본에서 계산한 통계값들이 채택역에 들어가는지, 기각영역에 들어가는지를 결정해야 함
- 임계값을 t 값으로 결정하였기에 표본에서 얻은 통계값도 임계값과 쉽게 비교할 수 있도록 t 값으로 나타내어야 함

t -통계량

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

t -검정의 예

어느 도시에서는 유치원 교사의 이직률이 높아 **평균** 재직기간이 $\mu = 20$ 개월이라 하자. 실제로 그러한 지를 알아보기 위해 **유치원 교사 10명을 뽑아** 평균 재직기간을 조사해 보았더니 아래의 표와 같았다. $\mu = 20$ 개월이라는 가설을 $\alpha = 0.05$ 수준에서 **양측검정과 단측검정**을 수행하시오. (단, 유치원 교사의 재직기간은 정규분포를 이룬다고 가정)

Chapter 11 단일모집단에 관한 가설검정

| 교사 | 재직기간(X_i) | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ |
|----|---------------|-----------------|---------------------|
| A | 16 | -7 | 49 |
| B | 28 | 5 | 25 |
| C | 20 | -3 | 9 |
| D | 34 | 11 | 121 |
| E | 22 | -1 | 1 |
| F | 18 | -5 | 25 |
| G | 30 | 7 | 49 |
| H | 22 | -1 | 1 |
| I | 25 | 2 | 4 |
| J | 15 | -8 | 64 |
| 합계 | 230(개월) | | 348 |

$$\bar{X} = \frac{\sum X_i}{n} = \frac{230}{10} = 23$$

$$S = \sqrt{S^2} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{348}{9}} = 6.2$$

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{6.2}{\sqrt{10}} = 1.96$$

Chapter 11 단일모집단에 관한 가설검정

양측검정

① $H_0 : \mu = 20$
 $H_1 : \mu \neq 20$

② $\alpha = 0.05$

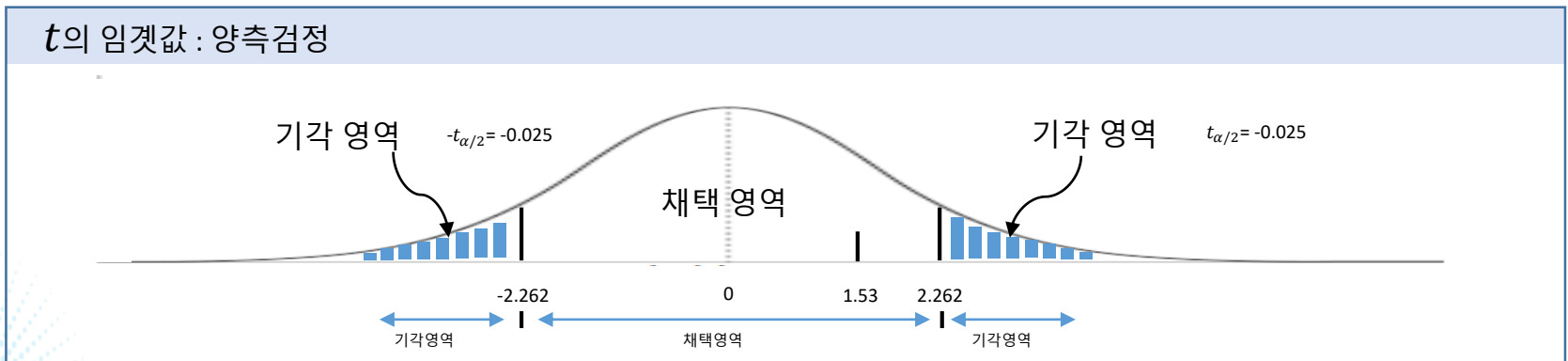
- ③ $\alpha/2 = 0.025$ 이며 자유도는 $n-1 = 9$ 일 때의 t -분포표,
채택영역 : $-2.262 \leq t \leq 2.262$
기각영역 : $t > 2.262$ 또는 $t < -2.262$

- ④ $\bar{X} = 23$ 에 해당되는 t 값을 계산하면,

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{23 - 20}{1.96} = 1.53$$

$t = 1.53$ 은 채택영역 안에 있으므로 $\alpha = 0.05$ 수준에서 귀무가설을 기각할 수 없다.

- ⑤ 귀무가설을 기각할 수 없으므로, 유치원 교사의 평균 재직기간이 이 표본의 결과로 본다면 $\mu = 20$ 개월로 볼 수 있다



Chapter 11 단일모집단에 관한 가설검정

단측검정

① $H_0 : \mu = 20$
 $H_1 : \mu > 20$

② $\alpha = 0.05$

③ $\alpha/2 = 0.025$ 이며 자유도는 $n-1 = 9$ 일 때의 t -분포표,

채택영역 : $t \leq 1.833$

기각영역 : $t > 1.833$

④ $\bar{X} = 23$ 에 해당되는 t 값을 계산하면,

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{23 - 20}{1.96} = 1.53$$

$t = 1.53$ 은 채택영역 안에 있으므로 $\alpha = 0.05$ 수준에서 귀무가설을 기각할 수 없다.

⑤ 귀무가설을 기각할 수 없으므로, 유치원 교사의 평균 재직기간이 이 표본의 결과로 본다면 $\mu = 20$ 개월로 볼 수 있다

t 의 임계값 : 단측검정



Chapter 11 단일모집단에 관한 가설검정

◎ t -분포와 Z -분포와의 관계

- 모집단이 정규분포를 이루며, 모집단의 분산을 알고 있을 때에는 **Z -분포를 활용**
모집단이 정규분포를 이루며, 모집단의 분산을 모를 때에는 **t -분포를 활용**
- 그러나 표본의 크기가 매우 크면, '**모집단 분산**'과 '**표본분산**' 간 차이가 적으므로 ' **t -통계량을 사용**'하거나 ' **Z -통계량을 사용**'하거나 **별 차이**가 없음
- ' **t 분포**'와 ' **Z 분포**'를 비교해보면 $\alpha = 0.05$ 이고 양측검정일 경우
표본의 수가 121때의 ' **Z 값**' = **1.96**
표본의 수가 120때의 ' **t 값**' = **1.98**

통상적으로 표본의 크기 $n=30$ 을 넘는 경우 ' **t 분포**'와 ' **Z 분포**' 중 어느 것을 활용해도 무방

| 모집단의 분산을 알고 있을 때 | 표본이 클 때 | 표본이 작을 때 |
|------------------|---------|----------|
| 모집단이 정규분포 | Z -분포 | Z -분포 |
| 모집단이 비정규분포 | Z -분포 | - |
| 모집단의 분산을 모를 때 | 표본이 클 때 | 표본이 작을 때 |
| 모집단이 정규분포 | Z -분포 | t -분포 |
| 모집단이 비정규분포 | Z -분포 | - |

Chapter 11 단일모집단에 관한 가설검정

◎ t -분포표

| α df | 0.4 | 0.25 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0025 | 0.001 | 0.0005 |
|----------------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|
| 1 | 0.325 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 127.32 | 318.31 | 636.62 |
| 2 | 0.289 | 0.816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 14.089 | 22.327 | 31.599 |
| 3 | 0.277 | 0.765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 7.453 | 10.215 | 12.924 |
| 4 | 0.271 | 0.741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | 0.267 | 0.727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | 0.265 | 0.718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | 0.263 | 0.711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | 0.262 | 0.706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | 0.261 | 0.703 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | 0.260 | 0.700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | 0.260 | 0.697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | 0.259 | 0.695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | 0.259 | 0.694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | 0.258 | 0.692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | 0.258 | 0.691 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | 0.258 | 0.690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | 0.257 | 0.689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | 0.257 | 0.688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.197 | 3.610 | 3.922 |
| 19 | 0.257 | 0.688 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | 0.257 | 0.687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| 21 | 0.257 | 0.686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| 22 | 0.256 | 0.686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |
| 23 | 0.256 | 0.685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.104 | 3.485 | 3.768 |
| 24 | 0.256 | 0.685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.091 | 3.467 | 3.745 |
| 25 | 0.256 | 0.684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.078 | 3.450 | 3.725 |
| 26 | 0.256 | 0.684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.067 | 3.435 | 3.707 |
| 27 | 0.256 | 0.684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.057 | 3.421 | 3.690 |
| 28 | 0.256 | 0.683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.047 | 3.408 | 3.674 |
| 29 | 0.256 | 0.683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.038 | 3.396 | 3.659 |
| 30 | 0.256 | 0.683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.030 | 3.385 | 3.646 |
| 40 | 0.255 | 0.681 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 2.971 | 3.307 | 3.551 |
| 60 | 0.254 | 0.679 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 2.915 | 3.232 | 3.460 |
| 120 | 0.254 | 0.677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 2.860 | 3.160 | 3.373 |
| ∞ | 0.253 | 0.674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 2.807 | 3.090 | 3.291 |

Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 비율에 관한 가설검정

- 모집단의 비율에 대한 가설 검정은 모정당의 지지율이 총투표수의 40%일 것이라 가정
- 차기 서울시장으로 어떤 정치인이 유력하다든지 하는 것이 바로 모집단의 비율과 관계된 가설
- 비율 π 의 표집분포의 표준편차는 다음과 같음

p 의 표집분포의 표준편차

$$\sigma_p = \sqrt{\pi(1 - \pi)/n}$$

- 비율의 표집분포 $n\pi > 5$, 그리고 $n(1 - \pi) \geq 5$ 일 때 정규분포와 근사한 것으로 볼 수 있음
- 따라서 이러한 조건을 만족하는 경우, 표본비율 p 에 대응하는 계산된 z 값은 아래와 같음

p 에 대응하는 z 값

$$Z = \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}$$

Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 비율에 관한 가설검정

예제 11-4

동전의 확률이 정상적인지를 확인 및 검정하고자 100번을 던져 보니, 앞면이 40번 나왔으며, 뒷면이 60번 나왔다. 이 동전의 확률이 동일한지를 $\alpha = 0.05$ 하에서 검정하자.

정상적인 동전을 던진다면, 앞면이 나올 확률은 50%여야 함

따라서 $\pi = 0.5$ 이고 표본에서의 성공비율 $p = 0.4$

① $H_0 : \pi = 0.5$
 $H_1 : \pi \neq 0.5$

② $\alpha = 0.05$

③ 채택영역 : $-1.96 \leq Z \leq +1.96$

기각영역 : $Z > 1.96$ 또는 $Z < -1.96$

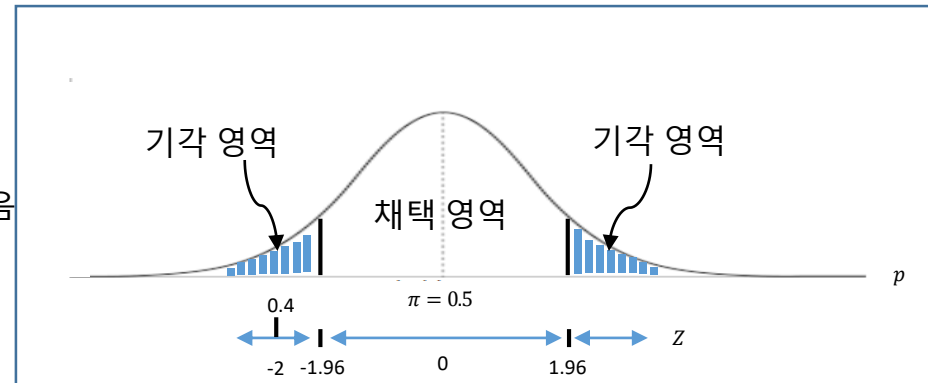
④ 표본의 결과인 $p = 0.4$ 를 Z 값으로 계산하면 다음과 같음

$$\sigma_p = \sqrt{\frac{0.5 \times 0.5}{100}} = 0.05$$

$$Z = \frac{p - \pi}{\sigma_p} = \frac{0.4 - 0.5}{0.05} = -2$$

$Z = -2$ 는 기각영역에 있으므로 귀무가설을 기각한다

⑤ 따라서 동전이 정상이라고 할 수 없다



Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 비율에 관한 가설검정

예제 11-5

조미료 A회사와 B회사가 경쟁이 치열한 상황에서 A 회사는 B 회사 제품 사용자들 중 40% 이하만 계속 B제품을 유지하며, 60%이상이 자기 회사 제품(A회사)로 바꾼다 주장한다. 이러한 A회사의 주장을 조사하기 위해 B회사는 B제품을 경험했던 500명을 표본조사한 결과, 215명이 계속 이용하고 있었다. 이 결과로 A회사의 주장을 반박할 수 있을까? 유의수준 $\alpha = 0.01$ 에서 A회사의 주장을 가설검정 해보자

따라서 $\pi = 0.4$ 이며 $n = 500$ 으로 비율의 표집분포 $n\pi > 5$, 그리고 $n(1 - \pi) \geq 5$ 를 만족하므로, Z 를 이용하여 가설

① $H_0 : \pi \leq 0.4$

$H_1 : \pi > 0.4$

② $\alpha = 0.01$

③ 채택영역 : $Z \leq 2.32$

기각영역 : $Z > 2.32$

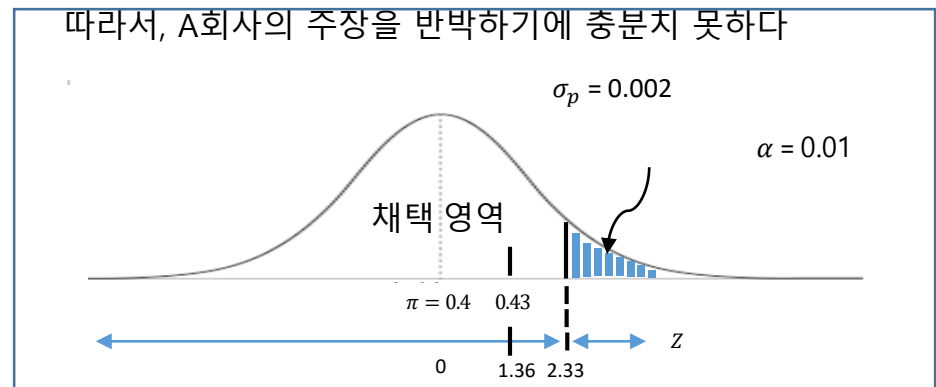
④ 표본의 성공비율 $p = 0.43 = \frac{215}{500}$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4 * (1-0.4)}{500}} = 0.022$$

$$Z = \frac{p - \pi}{\sigma_p} = \frac{0.43 - 0.40}{0.022} = 1.36$$

⑤ $Z = 1.36$ 은 채택영역 안에 있으므로 귀무가설을 기각할 수 없다.

따라서, A회사의 주장을 반박하기에 충분치 못하다



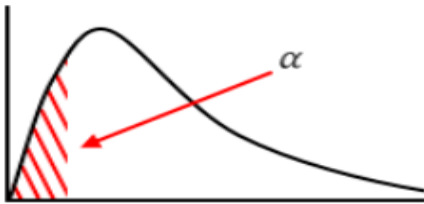
Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 분산에 관한 가설검정

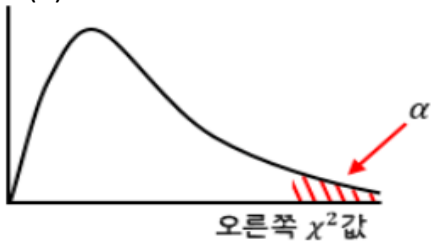
χ^2 분포와 임계값

χ^2 분포표에서의 임계값

(a) 왼쪽단측검정



(b) 오른쪽단측검정



| α df | 왼쪽 임계값 | | | | 오른쪽 임계값 | | | |
|------------------|--------|-------------|-------------|------|--------------|--------------|--------------|-------|
| | 0.01 | 0.025 | 0.05 | 0.10 | 0.90 | 0.95 | 0.975 | 0.99 |
| 1 | .00016 | .00098 | .004 | .016 | 2.71 | 3.84 | 5.02 | 6.63 |
| 2 | .02 | .05 | .10 | .21 | 4.61 | 5.99 | 7.38 | 9.21 |
| 3 | .11 | .22 | .35 | .58 | 6.25 | 7.81 | 9.35 | 11.34 |
| 4 | .30 | .48 | .71 | 1.06 | 7.78 | 9.49 | 11.14 | 13.28 |
| 5 | .55 | .83 | 1.15 | 1.61 | 9.24 | 11.07 | 12.83 | 15.09 |
| 6 | .87 | 1.24 | 1.64 | 2.20 | 10.64 | 12.59 | 14.45 | 16.81 |
| 7 | 1.24 | 1.69 | 2.17 | 2.83 | 12.02 | 14.07 | 16.01 | 18.48 |
| 8 | 1.65 | 2.18 | 2.73 | 3.49 | 13.36 | 15.51 | 17.53 | 20.09 |
| 9 | 2.09 | 2.70 | 3.33 | 4.17 | 14.68 | 16.92 | 19.02 | 21.67 |
| 10 | 2.56 | 3.25 | 3.94 | 4.87 | 15.99 | 18.31 | 20.48 | 23.21 |

Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 분산에 관한 가설검정

- **분산에 대한 가설검정**도 평균에 대한 가설검정과 마찬가지로 계산된 통계량을 토대로 표집분포에서 채택역인지 기각역인지를 판단
- 만일 계산된 χ^2 값이 χ^2 -분포의 양 끝에 있다면, 예외로 볼 수 있음
- 예외로 볼 수 있다면 해당 표본이 귀무가설에서 설정한 모집단에서 뽑혔다고 볼 수 없음
- 따라서 이는 귀무가설을 기각할 수 있다는 주장이 가능해짐

예제 11-6

어떤 모집단의 분산은 0.002다. 이 모집단으로부터 $n = 11$ 인 표본을 뽑았을 때, 이 표본의 분산 s^2 이 0.0032보다 클 확률은?

오른쪽 단측검정을 하기 위하여 0.0032에 해당하는 χ^2 값을 구하여 보면 됨 n 이 11 이므로 자유도는 $11-1 = 10$ 이다

$$\chi_{10}^2 = \frac{(n-1) S^2}{\sigma^2} = \frac{10 \times 0.0032}{0.002} = 16$$

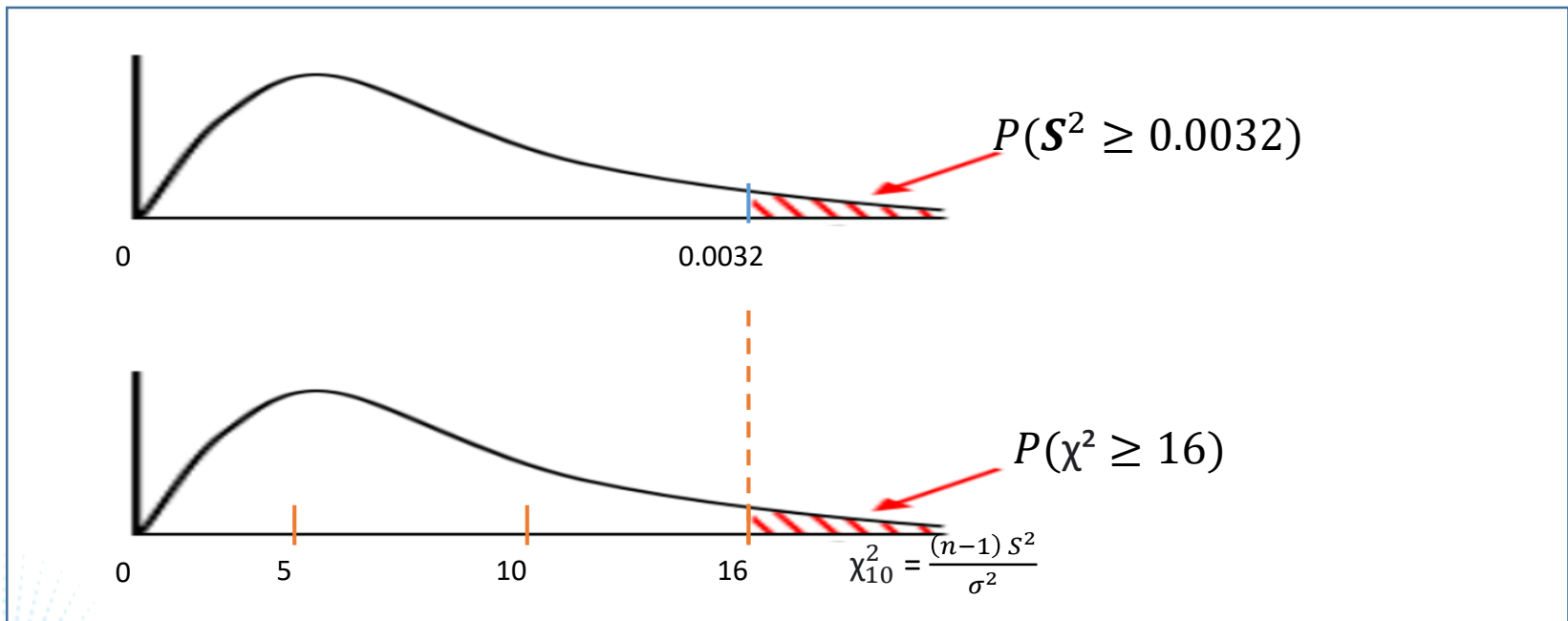
그러므로 예제에서 표본분산이 0.0032보다 클 확률은 χ_{10}^2 이 16보다 클 확률과 같음

$$P(S^2 \geq 0.0032) = P(\chi_{10}^2 \geq 16)$$

Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 분산에 관한 가설검정

예제 11-6 - 계속



Chapter 11 단일모집단에 관한 가설검정

◎ 모집단 분산에 관한 가설검정

양측검정

- 분산에 대한 가설검정에 있어서도 이미 설명한 모집단 평균에 대한 가설검정에서와 마찬가지로 양측검정(two-tailed test)과 단측검정(one-tailed test)을 할 수 있으며 그 절차도 동일
- 먼저 양측검정을 설명하면, 양측검정에 있어서 귀무가설과 대립가설은 다음과 같음 (단, q 는 임의의 수치)

$$\begin{aligned}H_0 : \sigma^2 &= q \\ H_1 : \sigma^2 &\neq q\end{aligned}$$

- 귀무가설을 채택할 수 있는 의사결정기준은 유의수준 α 에 따라 다르나, 일정한 α 수준에 따른 채택영역과 기각영역안 다음 페이지와 같이 표시가 가능

Chapter 11 단일모집단에 관한 가설검정

◎ 분산 가설검정 - 양측검정

양측검정의 경우

채택영역 $\chi^2_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}$

기각영역 $\chi^2_{\alpha/2} > \frac{(n-1)S^2}{\sigma^2}$

또는

$$\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2}$$

| α df | 왼쪽 임계값 | | | | 오른쪽 임계값 | | | |
|------------------|--------|--------------|--------------|-------|--------------|--------------|--------------|-------|
| | 0.01 | 0.025 | 0.05 | 0.10 | 0.90 | 0.95 | 0.975 | 0.99 |
| 25 | 11.52 | 13.11 | 14.61 | 16.47 | 34.38 | 37.65 | 40.64 | 44.31 |
| 26 | 12.19 | 13.84 | 15.37 | 17.29 | 35.56 | 38.88 | 41.92 | 45.64 |
| 27 | 12.87 | 14.57 | 16.15 | 18.11 | 36.74 | 40.11 | 43.19 | 46.96 |
| 28 | 13.56 | 15.30 | 16.92 | 18.93 | 37.91 | 41.33 | 44.46 | 48.27 |
| 29 | 14.25 | 16.04 | 17.70 | 19.76 | 39.08 | 42.55 | 45.72 | 49.58 |
| 30 | 14.95 | 16.79 | 18.49 | 20.59 | 40.25 | 43.77 | 46.97 | 50.89 |
| 40 | 22.16 | 24.43 | 26.50 | 29.05 | 51.80 | 55.75 | 59.34 | 63.69 |
| 50 | 29.70 | 32.35 | 34.76 | 37.68 | 63.16 | 67.50 | 71.42 | 76.15 |
| 60 | 37.48 | 40.48 | 43.18 | 46.45 | 74.39 | 79.08 | 83.29 | 88.37 |

Chapter 11 단일모집단에 관한 가설검정

◎ 분산 가설검정 - 양측검정

예제 11-7

어느 한 실험의 분산은 10이라고 한다. 그런데 최근 들어서 실험의 결과가 이상하게 나오는 것이, 아무래도 실험의 분산은 10이 아니라는 의견이 나왔다. 실제로 어떠한지를 알아보기 위해 표본 9개를 뽑았더니 표본 분산은 3이 나왔다고 한다. 이때 분산은 10이라고 할 수 있는지 유의수준 10%에서 검정하시오.

풀이) $H_0: \sigma^2=10$

$H_1: \sigma^2 \neq 10$

$$= \frac{(n-1)s^2}{\sigma^2}$$

$$= \frac{(9-1) \times 3}{10}$$

$$= 2.4$$

| α | 0.995 | 0.99 | 0.975 | 0.95 | 0.9 | 0.5 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
|----------|---------|--------|-------|-------|------|------|-------|-------|-------|-------|-------|
| 1 | 0.00004 | 0.0002 | 0.001 | 0.004 | 0.02 | 0.45 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | 0.01 | 0.02 | 0.05 | 0.10 | 0.21 | 1.39 | 4.61 | 5.99 | 7.38 | 9.21 | 10.60 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 2.37 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 |
| 4 | 0.21 | 0.30 | 0.48 | 0.71 | 1.06 | 3.36 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |
| 5 | 0.41 | 0.55 | 0.83 | 1.15 | 1.61 | 4.35 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 |
| 6 | 0.68 | 0.87 | 1.24 | 1.64 | 2.20 | 5.35 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 |
| 7 | 0.99 | 1.24 | 1.69 | 2.17 | 2.83 | 6.35 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 3.49 | 7.34 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 4.17 | 8.34 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 |



검정통계량이 기각역 안에 위치하므로 귀무가설 기각!
따라서 제품의 분산은 10이라고 할 수 없다.

Chapter 11 단일모집단에 관한 가설검정

◎ 분산 가설검정 - 단측검정

예제 11-8

A회사는 LED 전구를 생산하는데, 이 제품의 분산은 25라고 알려져 있다. 그런데 최근 품질관리팀의 분석에 의하면 제품의 불량률이 높아져서 분산이 25보다 커진 것 같다는 의견이 나왔다. 이에 실상을 파악하기 위해 표본 10개를 뽑아 조사하였더니, 표본분산은 29가 나왔다고 한다. 이때 분산이 25보다 크다고 할 수 있는지 유의수준 5%에서 검정하시오.

풀이) $H_0: \sigma^2 \leq 25$

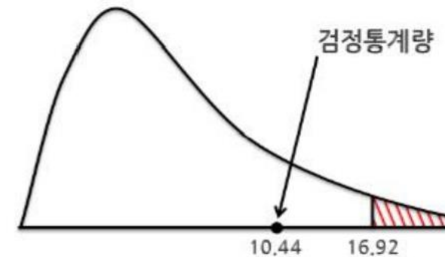
$H_1: \sigma^2 > 25$

$$= \frac{(n-1)s^2}{\sigma^2}$$

$$= \frac{(10-1) \times 29}{25}$$

$$= 10.44$$

| α | 0.995 | 0.99 | 0.975 | 0.95 | 0.9 | 0.5 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
|----------|---------|--------|-------|-------|------|------|-------|-------|-------|-------|-------|
| v | | | | | | | | | | | |
| 1 | 0.00004 | 0.0002 | 0.001 | 0.004 | 0.02 | 0.45 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 |
| 2 | 0.01 | 0.02 | 0.05 | 0.10 | 0.21 | 1.39 | 4.61 | 5.99 | 7.38 | 9.21 | 10.60 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 2.37 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 |
| 4 | 0.21 | 0.30 | 0.48 | 0.71 | 1.06 | 3.36 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 |
| 5 | 0.41 | 0.55 | 0.83 | 1.15 | 1.61 | 4.35 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 |
| 6 | 0.68 | 0.87 | 1.24 | 1.64 | 2.20 | 5.35 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 |
| 7 | 0.99 | 1.24 | 1.69 | 2.17 | 2.83 | 6.35 | 12.02 | 14.07 | 16.01 | 18.48 | 20.28 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 3.49 | 7.34 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 4.17 | 8.34 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 |



검정통계량이 채택역 안에 위치하므로 귀무가설 채택!
따라서 제품의 분산은 25보다 크다고 할 수 없다.

Chapter 12. 두 모집단의 비교에 관한 가설검정

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ INTRO

- ✓ 앞서 설명한 가설검정은 한 모집단의 특성에 관한 것
- ✓ 그러나 한 모집단의 평균에 대한 가설검정 못지 않게 두 모집단의 평균에 대한 가설검정을 할 때가 많이 존재
- ✓ 한 지역에서 표본으로 선정된 근로자들의 월급이 270만원이고 다른 지역에서의 근로자들의 월급이 280만원이라고 할 때, 두 지역의 근로자들의 월급이 차이가 난다고 볼 수 있는가 하는 문제
- ✓ 즉, 두 지역에서 계산된 표본들의 월급차가 10만원이지만 이러한 차이가 모집단 자체가 다르기에 일어난 차이인지 혹은 단순한 표본선정에서 생길수 있는 오차인지를 알아보는 것

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포

- 앞선 제 8장에서 단일 모집단의 평균에 대한 표집분포를 다룬 바 있음
- 두 모집단으로부터 뽑힌 두 표본평균의 차이에 관한 표집분포를 분석
- 첫 번째 모집단에서 뽑힌 표본들의 평균을 각각 $\bar{X}_{11}, \bar{X}_{12}, \dots, \bar{X}_{1i}$ 라 하고, 두 번째 모집단에서 뽑힌 모든 표본들의 평균을 $\bar{X}_{21}, \bar{X}_{22}, \dots, \bar{X}_{2j}$ 라 표시
- 각 모집단에서 뽑힌 표본들의 순서를 가리키지 않고, 단지 첫번째 모집단에서 뽑힌 것인지 또는 두번째 모집단에서 뽑힌 것인지만을 가리킬 때는 편의상 각각 \bar{X}_1, \bar{X}_2 로 표시
- $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포라 함은 첫번째 모집단에서 뽑힐 수 있는 표본들과 두번째 모집단에서 뽑힐 수 있는 모든 표본들의 평균차의 표집분포를 뜻함

$(\bar{X}_1 - \bar{X}_2)$ 의 표집분포의 예

| 서 여직원 | 부 | 경리과 | 총무과 |
|----------|---|-----|-----|
| | | | |
| 1 | | 20 | 20 |
| 2 | | 26 | 24 |

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포

- 어느 회사의 경리과와 총무과에 여직원이 각각 2명 있으며, 이들의 연령은 각각 위 표와 같음
- 각 과별로 여직원을 두 명 뽑을 때 두 과에서 뽑힐 수 있는 선택가능한 표본들은 아래와 같음
(복원추출로 시행)

모집단 1 (경리과)

$$(1) \quad 20, 20 : \bar{X}_{11} = 20$$

$$(2) \quad 20, 26 : \bar{X}_{12} = 23$$

$$(3) \quad 26, 20 : \bar{X}_{13} = 23$$

$$(4) \quad 26, 26 : \bar{X}_{14} = 26$$

모집단 2 (총무과)

$$(1) \quad 20, 20 : \bar{X}_{21} = 20$$

$$(2) \quad 20, 24 : \bar{X}_{22} = 22$$

$$(3) \quad 24, 20 : \bar{X}_{23} = 22$$

$$(4) \quad 24, 24 : \bar{X}_{24} = 24$$

- 경리과에서 4가지 표본, 총무과에서 4가지 표본이 뽑힐 수 있음. 그러나 두 모집단에서 뽑힌 모든 표본

평균들의 차이를 보면 아래와 같이 16가지의 경우가 일어남

(만일 경리과의 여직원이 5명, 총무과 여직원 10명이라면, 복원추출시 뽑힐 수 있는 표본은 경리과 $5 \times 5 = 25$ 개, 총무과 10×10 총 조합 2500)

$$\bar{X}_{11} - \bar{X}_{21} = 0, \quad \bar{X}_{11} - \bar{X}_{22} = -2, \quad \bar{X}_{11} - \bar{X}_{23} = -2, \quad \bar{X}_{11} - \bar{X}_{24} = -4,$$

$$\bar{X}_{12} - \bar{X}_{21} = 3, \quad \bar{X}_{12} - \bar{X}_{22} = 1, \quad \bar{X}_{12} - \bar{X}_{23} = 1, \quad \bar{X}_{12} - \bar{X}_{24} = -1,$$

$$\bar{X}_{13} - \bar{X}_{21} = 3, \quad \bar{X}_{13} - \bar{X}_{22} = 1, \quad \bar{X}_{13} - \bar{X}_{23} = 1, \quad \bar{X}_{13} - \bar{X}_{24} = -1,$$

$$\bar{X}_{14} - \bar{X}_{21} = 6, \quad \bar{X}_{14} - \bar{X}_{22} = 4, \quad \bar{X}_{14} - \bar{X}_{23} = 4, \quad \bar{X}_{14} - \bar{X}_{24} = 2,$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포

$(\bar{X}_1 - \bar{X}_2)$ 분포의 평균과 표준편차

- 두 모집단에서 뽑힌 모든 표본평균들 간의 차이분포에서 평균을 $\mu_d = \mu_1 - \mu_2$ 분산을 $\sigma_d^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ 로 계산이 가능함
- 위의 공식은 두 모집단이 서로 독립적이란 가정하에서 계산된 것, 성질은 아래와 같다
 - (1) 두 모집단의 분포가 정규분포면, 평균차의 표집분포도 정규분포를 이룸
 - (2) 두 모집단의 분포가 정규분포가 아닐지라도, 표본의 크기 즉 n_1 과 n_2 가 충분히 크면, $(\bar{X}_1 - \bar{X}_2)$ 분포는 정규분포가 됨. 이는 '중심극한 정리'으로 증명
- 앞의 예제에서 경리와 여직원의 평균 나이 $\mu_1 = 23$ 이며 $\mu_2 = 22$ 이므로 $\mu_d = \mu_1 - \mu_2 = 1$ 이 됨
이는 앞의 표에서의 계산된 차이를 모두 다 더하여 16으로 나누어도 마찬가지
- 두 모집단의 표본들이 서로 독립적이라면, 직접 평균차의 분산을 구하지 않고도 두 모집단 평균차의 분산은 6.5임을 공식에 의해 도출이 가능

$$\sigma_d^2 = \frac{9}{2} + \frac{4}{2} = 6.5$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포

예제 12-1

영식초교 6학년 학생 400명과 다니초교 6학년 학생 900명이 동일한 문제로 학업성취 검사를 보았는데 각각 $\mu_1 = 80$, $\mu_2 = 80$ 이었다. 그리고 두 학교의 분산은 다 같이 200이라는 것을 알고 있다. 이때 연희초등학교와 잠실초등학교에서 4명씩 선택하여 평균을 계산하고 두 평균의 차를 계산한다고 하자. 각 학교의 점수분포는 정규분포이고 서로 독립적이라고 한다.

A. 두 학교에서 뽑힌 표본학생의 점수차 분포는 $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포임
이 표집분포의 평균은 다음과 같음

$$\mu_d = \mu_1 - \mu_2 = 80 - 80 = 0$$

$(\bar{X}_1 - \bar{X}_2)$ 의 표집분포의 분산과 표준편차는

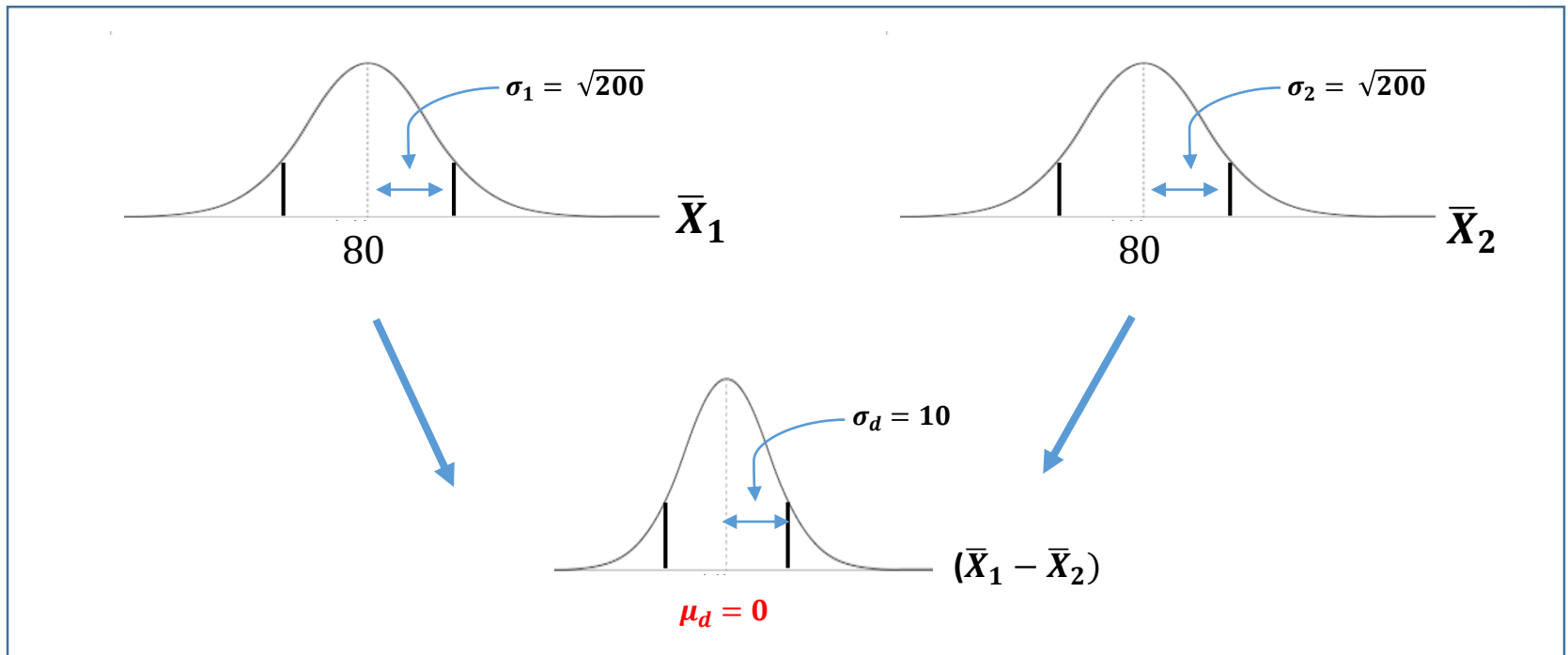
$$\sigma_d^2 = \frac{200}{4} + \frac{200}{4} = 100$$

$$\sigma_d = \sqrt{100} = 10(\text{점})$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포

$(\bar{X}_1 - \bar{X}_2)$ 분포에서 z 값



두 학교의 평균차의 표집분포가 정규분포라는 특성으로부터 특정한 $(\bar{X}_1 - \bar{X}_2)$ 의 값에 대응하는 z 값은 다음 페이지와 같이 계산됨

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포

$(\bar{X}_1 - \bar{X}_2)$ 분포에서 z 값

$(\bar{X}_1 - \bar{X}_2)$ 의 z 통계량

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma^2}$$

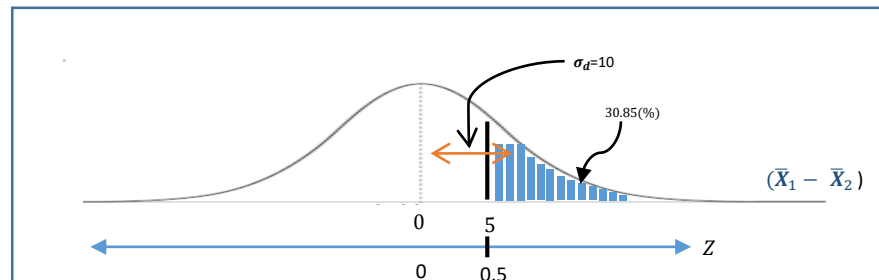
예제 12-2

앞의 1번 예제에서 4명씩 뽑아서 평균을 냈을 때, 영식초교 학생의 평균이 다니초교의 평균보다 5점 이상 높을 가능성은 얼마일까?

이 문제를 풀기 위하여 두 표본의 평균차의 분포를 그려서, $(\bar{X}_1 - \bar{X}_2) = 5$ 에 해당하는 z 값을 구해보면 다음과 같음

$$Z = \frac{5 - 0}{10} = 0.5$$

$$P(Z \geq 0.5) = 30.85(\%)$$



위의 예에서는 표본크기 n_1 과 n_2 가 서로 같았으나 서로 다른 경우에도 동일하게 적용됨

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 평균의 차이에 관한 가설설정

- 두 모집단의 평균과 관련된 가설검정은 두 모집단의 평균차에 대한 것이 대부분
- 예를 들면 A동네와 B동네의 소득수준에 차이가 있는지를 알아보기 위해서 두 동네에서 각각 표본을 뽑아 그 결과를 분석하여 '두동네의 소득수준에 차이가 있다' 혹은 "차이가 없다"라는 가설을 검정하는 것
- 아래의 예를 통해 적절한 귀무가설 및 대립가설을 설정하여 보자.

예제 12-3

서울시민 25명을 표본으로 뽑아 그들의 월 수입을 조사해본 결과 평균이 270만원이었으며, 대구시민을 상대로 36명의 표본을 뽑아보았더니 그들의 월평균 수입이 260만원이었다. 전체적으로 볼 때, 서울시민의 월평균 수입과 대구시민의 월평균 수입이 같다고

할 수 있을까?

- A. 서울시민의 평균 수입을 μ_1 이라 하고, 대구 시민의 평균 수입을 μ_2 라 하면 두 모집단의 평균이 같다는 것은 $\mu_1 = \mu_2$ 로 나타냄. 평균수입이 다르다는 것은 $\mu_1 \neq \mu_2$ 되므로 아래와 같이 귀무가설, 대립가설 설정이 가능

$$H_0 : \mu_1 = \mu_2 \text{ 또는 } \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \text{ 또는 } \mu_1 - \mu_2 \neq 0$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 평균의 차이에 관한 가설설정

예제 12-4

몇 년 동안의 연구결과, 아침식사를 하는 학생이 아침식사를 거르는 학생에 비해 학습능
률

이 더 높다고 주장한다. 과연 이는 사실일까?

- A. 아침식사를 하는 학생들의 평균 점수를 μ_1 이라 하고, 아침식사를 거르는 학생들의 평균 점수를 μ_2 라 하면 귀무
가설과 대립가설은 아래와 같이 설정이 가능

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

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

예제 12-5

해외유학을 위한 영어시험에서는 대체로 인문계 졸업생이 자연계 졸업생보다 시험점수
가

높은 편이다. 실제로 두 집단의 점수에 차이가 있는지를 알아보기 위하여 시험을 치른
인문계(μ_1)와 자연계(μ_2) 졸업생들 중에서 각각 표본을 뽑아 평균 점수를 비교하려 한다.
이때의 가설은 아래와 같이 설정한다

$$H_0 : \mu_1 \geq \mu_2 \text{ 또는 } \mu_1 - \mu_2 \geq 0$$

$$H_1 : \mu_1 < \mu_2 \text{ 또는 } \mu_1 - \mu_2 < 0$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 평균의 차이에 관한 가설설정

모집단의 분산을 알고 있을 때

대표본의 경우

한 모집단으로부터 뽑은 표본의 크기가 n_1 , 평균이 \bar{X}_1 이고, 다른 모집단으로부터 뽑은 표본의 크기가 n_2 , 평균이 \bar{X}_2 라고 하자. 이 때 두 표본의 평균 차, 즉 $(\bar{X}_1 - \bar{X}_2)$ 분포의 표준편차를 σ_d 라 하면 σ_d 는 다음과 같이 계산됨

$(\bar{X}_1 - \bar{X}_2)$ 의 표준편차 (모집단의 분산을 알고 있을 때)

$$\sigma_d = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$$

n_1 과 n_2 의 표본크기가 클 때에는 두 모집단이 정규분포인가의 여부와 관계없이 $(\bar{X}_1 - \bar{X}_2)$ 의 표집분포는 정규분포를 따름

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 평균의 차이에 관한 가설검정

모집단의 분산을 알고 있을 때

대표본의 경우

실제로 두 모집단에서 뽑힌 표본의 평균인 \bar{X}_1 과 \bar{X}_2 를 기초로 하여 두 모집단 평균의 차를 검정하기 위한 통계량 Z 는 아래와 같이 구할 수 있음

두 모집단 평균의 차이를 검정하기 위한 z-통계량

(두 모집단의 분산을 알고 있을 때)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

모집단의 분산을 알고 있을 때

대표본의 경우

예제 12-6

최근 국내에서 들어가는 다문화가정에서 자란 아이와 일반가정에서 자란 아이의 한국어 어휘능력에 차이가 있는지를 알아보기 위해 어휘능력 시험을 치러 결과를 알아보았더니 다음과 같았다. 이 어휘검사의 표준편차는 두 집단 모두 $\sigma = 5$ 로 나왔다. 두집단 아이들의 어휘 검사의 결과에 차이가 있는지 $\alpha = 0.05$ 의 유의수준에서 검정하세요

다문화 가정

$$n_1 = 50$$
$$\bar{X}_1 = 83$$

일반가정

$$n_2 = 50$$
$$\bar{X}_2 = 86$$

위의 문제를 양측검정으로 접근하여 풀려면, $(\bar{X}_1 - \bar{X}_2)$ 분포의 표준편차를 $\sigma_d = 1$

$$\sigma_d = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} = \sqrt{\frac{25}{50} + \frac{25}{50}} = 1$$

(1) $H_0 : \mu_1 = \mu_2$
집단 아이들
 $H_1 : \mu_1 \neq \mu_2$
같다고

(3) 채택영역 : $-1.96 \leq Z \leq 1.96$

기각영역 : $Z > 1.96$ 또는 $Z < -1.96$

(5) $\alpha = 0.05$ 에서 두

어휘검사 점수가

말할 수

없다.

(2) $\alpha = 0.05$

$$(4) Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_d} = \frac{(-3) - (0)}{1} = -1$$

Chapter 12 두 모집단의 비교에 관한 가설검정

모집단의 분산을 알고 있을 때

소표본의 경우

- 만약 소표본(표본의 크기가 작은 경우) 두 모집단 평균의 차에 대한 가설검정에 대해 설명을 하면 앞에서 말한 바와 같이 모집단이 정규분포일 때에는 표본의 크기가 작아도 차의 표집분포가 정규분포가 되므로, 모집단의 분산을 알고 있다면, **Z-분포를 이용할 수 있음**
- 그러나 모집단이 정규분포가 아닐 때에는 모집단의 분산을 알고 있다 하더라도 소표본인 경우에는 Z-분포를 적용하지 못함
- 따라서, 소표본인 경우에는 두 모집단의 분포가 정규분포를 이룬다는 가정하에서만 두 모집단의 평균차에 대한 가설검정이 가능

Chapter 12 두 모집단의 비교에 관한 가설검정

모집단의 분산을 알고 있을 때

소표본의 경우

예제 12-7

어느 회사에서 직업훈련을 10명에게 실시하였다. 그들이 훈련을 받은 후 제품 A를 하나 생산하는데 평균 12.1시간 걸렸으며, 훈련을 안 받은 사람 16명을 대상으로 능률을 조사하여 본 결과 제품 A를 생산하는 데 평균 14.2시간이 걸렸다. 이 결과로부터 직업훈련이 효과가 있다고 볼 수 있는가? 각 집단의 표준편차는 4시간으로 같다고 하고 $\alpha = 0.05$ 며, 두 모집단 모두 정규분포를 이룬다고 가정

훈련을 받은 사람

$$\begin{aligned}n_1 &= 10 \\ \bar{X}_1 &= 12.1\end{aligned}$$

훈련을 받지 않은 사람

$$\begin{aligned}n_2 &= 16 \\ \bar{X}_2 &= 14.2\end{aligned}$$

$$\sigma_1 = \sigma_2 = 4$$

이 때 $(\bar{X}_1 - \bar{X}_2)$ 분포의 표준편차 σ_d 는 다음과 같이 계산

$$\sigma_d = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} = \sqrt{\frac{16}{10} + \frac{16}{16}} = \sqrt{2.60} = 1.61$$

훈련받은 집단의 능률이 더 높을 것이라는 입장을 갖고 가설검정을 한다면 단측검정을 함. 두 집단이 정규분포며, 모집단의 표준편차를 알고 있으므로 Z-분포로써 가설검정을 수행

Chapter 12 두 모집단의 비교에 관한 가설검정

모집단의 분산을 알고 있을 때

소표본의 경우

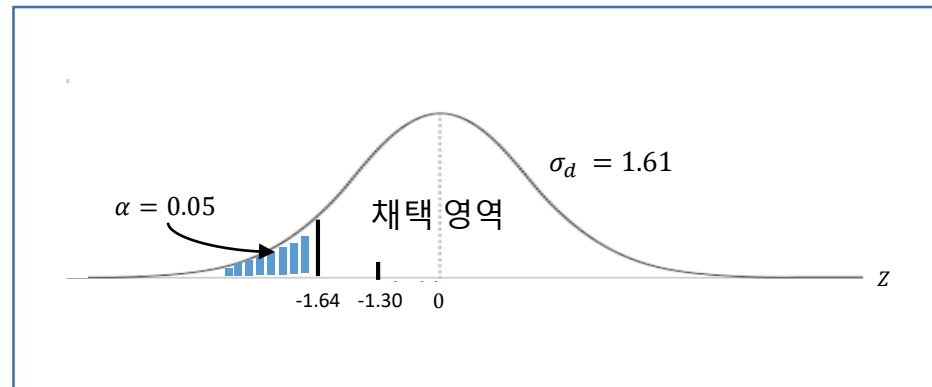
예제 12-7 - 계속

① $H_0 : \mu_1 = \mu_2$
 $H_1 : \mu_1 < \mu_2$

② $\alpha = 0.05$

③ 채택영역 : $Z \geq -1.64$

기각영역 : $Z < -1.64$



④ 표본의 결과를 Z값으로 계산하면 다음과 같음

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_d} = \frac{(-2.1) - (0)}{1.61} = -1.30$$

-1.30은 채택영역에 들어가므로 귀무가설을 기각할 수 없다

⑤ 훈련의 결과가 $\alpha = 0.05$ (단측검정)에서 효과가 있다고 할 수 없다.

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 평균의 차이에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

- 두 모집단의 분포가 각각 정규분포를 이룬다면 두 표본의 평균차, 즉 $(\bar{X}_1 - \bar{X}_2)$ 분포도 정규분포를 이루며, 두 모집단의 분포가 정규분포가 아니더라도 표본의 크기가 충분히 크면 중심극한정리에 의해서 $(\bar{X}_1 - \bar{X}_2)$ 의 분포는 정규분포에 가까워지므로, 정규분포를 이용하여 두 모집단 평균의 차에 대한 가설 검정을 할 수 있다.
- 그러나 두 모집단의 분포가 정규분포를 이루더라도 모집단의 분산을 모르고 표본의 크기도 작을 때에는 분포를 이용해서 평균의 차에 대한 가설검정

대표본의 경우

- 두 모집단의 분포가 정규분포인 경우는 물론이고, 정규분포가 아니더라도 표본이 충분히 크면, 중심극한정리에 의해 $(\bar{X}_1 - \bar{X}_2)$ 분포는 정규분포에 가까워지므로 σ^2_1 과 σ^2_1 의 추정값으로서 s^2_1, s^2_2 을 그대로 사용할 수 있음

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 평균의 차이에 관한 가설설정

두 모집단의 분산을 모르고 있을 때

- 즉, 두 모집단에서 뽑힌 표본의 분산을 각각 s^2_1, s^2_2 라 하고 $(\bar{X}_1 - \bar{X}_2)$ 분포의 표준편차 S_d 를 계산하면 다음과 같다

두 모집단 평균의 차이를 검정하기 위한 z-통계량

(두 모집단의 분산을 알고 있을 때)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

예제 12-8

지금까지의 경험으로 보아 A회사에서 만다는 전구의 평균수명은 B회사에서 만드는 것보다 200시간이 더 길다고 한다. A 전구회사에서 169개의 표본을 뽑아 평균 수명을 조사했더니 1,400시간이었고 표준편차는 130시간이었다. B회사에서는 144개를 뽑았는데 표본의 평균 수명은 1,300시간이었고 표준편차는 120시간이었다. A회사의 전체 전구의 평균 수명이 B회사 전구보다 200시간 길다는 것을 95% 신뢰구간에서 검정하시오.

A회사

B회사

$$\bar{X}_1 = 1,400$$

$$\bar{X}_2 = 1,300$$

$$n_1 = 169$$

$$n_2 = 144$$

$$S_1 = 130$$

$$S_2 = 120$$

$$S_d = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$= \sqrt{\frac{130^2}{169} + \frac{120^2}{144}} = \sqrt{200} = 14.14$$

Chapter 12 두 모집단의 비교에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

예제 12-8 - 계속

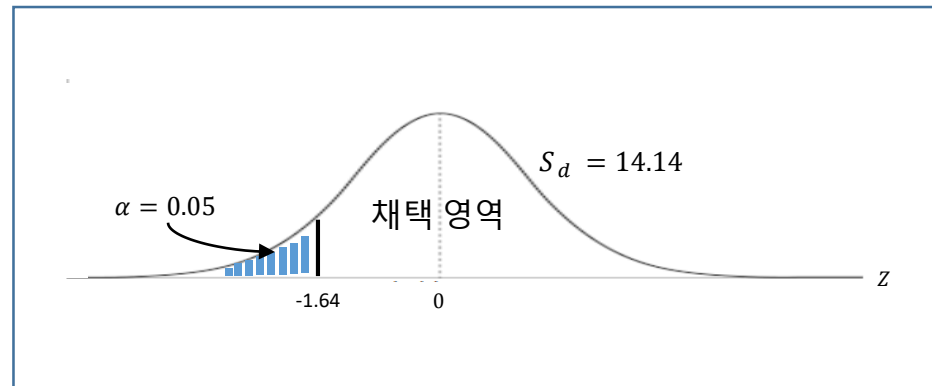
① $H_0 : \mu_1 - \mu_2 = 200$

$H_1 : \mu_1 - \mu_2 < 200$

② $\alpha = 0.05$

③ 채택영역 : $Z \geq -1.64$

기각영역 : $Z < -1.64$



④ 표본의 결과를 Z값으로 계산하면 다음과 같음

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_d} = \frac{(1,400) - (1,300)}{14.14} = -7.07$$

-7.07은 귀무가설의 기각역에 들어가므로 **귀무가설을 기각한다**

⑤ 유의수준 5%의 단측검정에서 A회사가 만든 전구의 평균 수명이 B회사의 것보다 200시간 더 길다고 할 수 없다.

Chapter 12 두 모집단의 비교에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

소표본의 경우

- 소표본의 경우에는 중심극한정리를 적용할 수 없고 이에 따라 정규분포를 이루지 못하므로 아래의 식을 적용하여 구하는 수 밖에 없음
- 아래의 식을 적용하기 위한 2가지 가정 :: 1) 두 개의 모집단이 모두 정규분포여야 함.
2) 두 모집단의 분산($\sigma^2_1 = \sigma^2_2$)이 서로 동일하다는

점.

$(\bar{X}_1 - \bar{X}_2)$ 의 표준편차 (모집단의 분산을 모르고 소표본일 경우)

$$S_d = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad S_p = \sqrt{\frac{(n_1 - 1) * S^2_1 + (n_2 - 1) * S^2_2}{n_1 + n_2 - 2}}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

소표본의 경우

$(\bar{X}_1 - \bar{X}_2)$ 의 표준편차 (모집단의 분산을 모르고 소표본일 경우)

$$S_d = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad S_p = \sqrt{\frac{(n_1 - 1) * S^2_1 + (n_2 - 1) * S^2_2}{n_1 + n_2 - 2}}$$

두 모집단의 평균의 차를 검정하기 위한 t-통계량

(모집단의 분산을 모르고 소표본일 경우)

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

식에서의 S_p 는 표준편차의 집합추정값(pooled estimate of standard deviation)이라 함

Chapter 12 두 모집단의 비교에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

소표본의 경우

- 모집단의 분산을 모르는 소표본의 경우 **두 모집단의 분산($\sigma^2_1 = \sigma^2_2$)이 서로 동일**하다는 점을 가정으로 하였으므로 우리가 가지고 있는 표본분산의 s^2_1 과 s^2_2 자유도에 따라 가중 평균한 값으로 s^2_p 을 취하는 것
- 즉, 각 표본의 크기를 n_1, n_2 라 하면 표본분산의 자유도는 $(n_1-1), (n_2-1)$ 이며, 총 자유도는 $(n_1-1) + (n_2-1) = n_1 + n_2 - 2$ 이므로 가중평균값인 s^2_p 은 다음과 같이 계산됨

$$s^2_p = \frac{(n_1-1)}{n_1 + n_2 - 2}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

두 모집단의 분산을 모르고 있을 때

소표본의 경우

예제 12-9

두 가지 종류의 암기법이 서로 차이가 있는지 알아보기 위하여, 두 집단의 어린아이에게 각각의 방법으로 외국어 단어를 가르친 다음, 암기하고 있는 단어가 몇 개 인가 알아보았더니 다음과 같았다.

암기법 A : 5, 2, 4, 7, 4, 4, 8, 3, 7, 6 ($n_1 = 10$)

암기법 B : 6, 5, 4, 9, 4, 6, 8, 5, 6, 7 ($n_2 = 10$)

위의 자료를 기초로 하여 두 암기법의 효과가 서로 차이가 있는지를 유의수준 $\alpha = 0.1$ 에서 검정하라. 두 모집단의 분포는 정규분포임을 가정.

먼저 위의 가설검정을 위해 필요한 $\bar{X}_1, \bar{X}_2, S_1^2, S_2^2$, 그리고 S_p 를 계산하면 다음과 같다

Chapter 12 두 모집단의 비교에 관한 가설검정

| 어린이 | 단어수(X_1) | $(X_1 - \bar{X}_1)^2$ |
|-----|--------------|-----------------------|
| A | 5 | 0 |
| B | 2 | 9 |
| C | 4 | 1 |
| D | 7 | 4 |
| E | 4 | 1 |
| F | 4 | 1 |
| G | 8 | 9 |
| H | 3 | 4 |
| I | 7 | 4 |
| J | 6 | 1 |
| 합 계 | 50 | 34 |

암기법 A로 가르친 결과

| 어린이 | 단어수(X_2) | $(X_2 - \bar{X}_2)^2$ |
|-----|--------------|-----------------------|
| K | 6 | 0 |
| L | 5 | 1 |
| M | 4 | 4 |
| N | 9 | 9 |
| O | 4 | 4 |
| P | 6 | 0 |
| Q | 8 | 4 |
| R | 5 | 1 |
| S | 6 | 0 |
| T | 7 | 1 |
| 합 계 | 60 | 24 |

암기법 B로 가르친 결과

$$\bar{X}_1 = \frac{\sum X_1}{n_1} = \frac{50}{10} = 5.0$$

$$S_1^2 = \frac{\sum (X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{34}{9} = 3.8$$

$$\bar{X}_2 = \frac{\sum X_2}{n_2} = \frac{60}{10} = 6.0$$

$$S_2^2 = \frac{\sum (X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{24}{9} = 2.7$$

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

$$= \sqrt{\frac{(10 - 1) * 3.8 + (10 - 1) * 2.7}{10 + 10 - 2}}$$

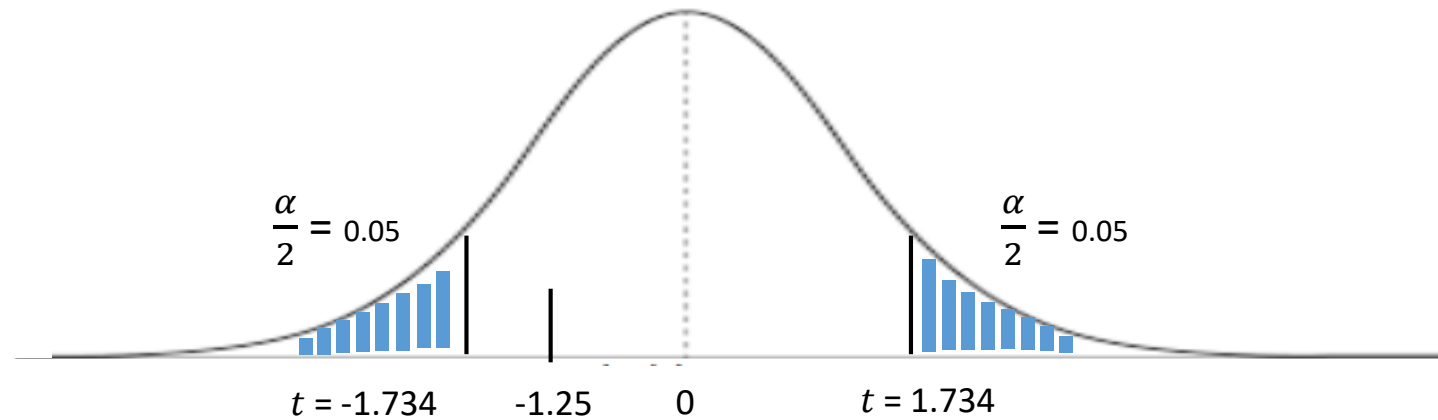
$$= 1.8$$

$$S_d = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.8 * \sqrt{\frac{1}{10} + \frac{1}{10}} = 0.8$$

Chapter 12 두 모집단의 비교에 관한 가설검정

채택영역 : $-1.734 \leq t \leq 1.734$

기각영역 : $t > 1.734$ 또는 $t < -1.734$



$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(5 - 6) - 0}{0.8} = -1.25$$

귀무가설이 채택된다는것은 유의수준 $\alpha = 0.1$ 에서 **두 암기법 사이에는 차이가 없다는 것을 의미한다**

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 짝을 이룬 표본의 차이에 대한 검정

- 앞서 설명했던 서로 다른 두개의 모집단에서 뽑은 표본평균의 차이에 대한 가설검정에 대해 언급하였다. 이와 달리 하나의 모집단에서 표본을 샘플링(Sampling)하고 그 표본으로부터 쌍으로 된 관찰 값들(paired sample)을 뽑아서 이들 간 차이에 대한 가설 검정을 해야 할 경우가 있다.
- Ex) 예를 들어 어느 회사에서 직업훈련이 근로자의 능률향상에 효과가 있는지를 알고 싶다고 하자 – 이를 위해 16명의 근로자를 뽑아서 직업훈련을 하기 전과 후의 작업능률의 점수를 알아보았더니 다음 페이지에서의 표와 같다면, 조사결과로써 훈련전과 훈련후의 능률이 같다고 할 수 있을까? (모집단에서의 차이의 분포는 정규분포라 가정하고, 짝을 이룬 표본의 차이에 대한 표집분포는 아래와 같다)

짝을 이룬 표본의 차이에 대한 표집분포

$$\text{평균} \quad \mu_d = \mu_1 - \mu_2$$

$$\text{표준편차} \quad \sigma_d = \frac{S_d}{\sqrt{n}} \quad \text{단, } S_d = \sqrt{\frac{\sum (D_i - \bar{D})^2}{(n-1)}}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 짝을 이룬 표본의 차이에 대한 검정

| 근로자 | 훈련후 (X_1) | 훈련전 (X_2) | 차이($D_i = X_1 - X_2$) | $D_i - \bar{D}$ | $(D_i - \bar{D})^2$ |
|-----|------------------|------------------|-------------------------|-----------------|---------------------|
| A | 80 | 75 | 5 | 4 | 16 |
| B | 90 | 83 | 7 | 6 | 36 |
| C | 92 | 96 | -4 | -5 | 25 |
| D | 75 | 77 | -2 | -3 | 9 |
| E | 86 | 81 | 5 | 4 | 16 |
| F | 90 | 90 | 0 | -1 | 1 |
| G | 81 | 82 | -1 | -2 | 4 |
| H | 70 | 67 | 3 | 2 | 4 |
| I | 89 | 94 | -5 | -6 | 36 |
| J | 88 | 85 | 3 | 2 | 4 |
| K | 82 | 78 | 4 | 3 | 9 |
| L | 79 | 82 | -3 | -4 | 16 |
| M | 91 | 96 | -5 | -6 | 36 |
| N | 90 | 80 | 10 | 9 | 81 |
| O | 78 | 87 | -9 | -10 | 100 |
| P | 89 | 81 | 5 | 7 | 49 |
| 합계 | | | 16 | | 442 |

$$\bar{D} = \frac{\sum(X_1 - X_2)}{n} = \frac{\sum D_i}{n} = \frac{16}{16} = 1$$

$$S_d = \sqrt{\frac{\sum(D_i - \bar{D})^2}{(n-1)}} = \sqrt{\frac{442}{15}} = \sqrt{29.47} = 5.43$$

짝을 이룬 표본의 차이검정을 위한 t-통계량

$$t = \frac{\bar{D} - (\mu_1 - \mu_2)}{S_d / \sqrt{n}}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 짝을 이룬 표본의 차이에 대한 검정

* 위의 값들을 활용하여 $\alpha = 0.05$ 에서 훈련 전과 훈련 후의 능률이 같다는 가설을 검정해보자

(1) $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

(2) $\alpha = 0.05$

(3) 자유도 df: $16-1 = 15$, $\alpha = 0.05$ 일 때
양측검정이므로

채택영역: $-2.131 \leq t \leq 2.131$

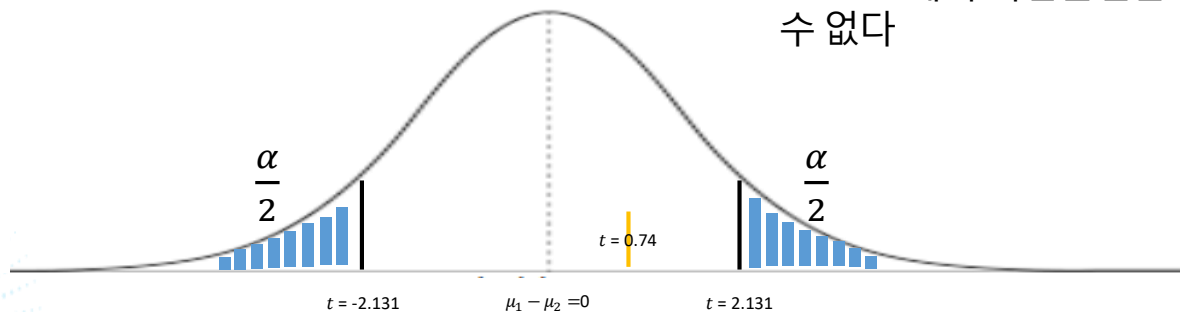
기각영역: $t > 2.131$ 또는 $t < -2.131$

(4) 표본의 결과로부터 계산된 t값은 아래와 같다

$$t = \frac{\bar{D} - (\mu_1 - \mu_2)}{S_d / \sqrt{n}} = \frac{1 - 0}{5.43 / \sqrt{16}} = 0.74$$

(5) 계산된 t 값이 0.74는 $-2.131 \leq t \leq 2.131$ 안에 포함되어 있으므로 귀무가설을 기각할 수 없다.

$\alpha = 0.05$ 에서 직업훈련은 능률향상에 효과가 있다고 할 수 없다



Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 비율의 차이에 관한 가설검정

* 단일모집단의 비율값에 대한 가설검정은 ch11. 2절에서 살펴본 바 있다. 여기서의 비율에 관한 검정은 두 모집단의 성공비율에 차이가 있느냐 없느냐에 대한 검정이다. 두 모집단의 비율의 차이에 대한 가설검정의 순서는 아래와 같다.

- (1) 두 표본으로부터 비율 P_1 과 P_2 를 산출한다.
- (2) 두 표본의 모집단의 비율이 같다고 가정을 하면, 이 추정비율($\hat{\pi}$)은 다음과 같이 표본비율로부터 가중평균으로 어림이 가능하다.

$$\hat{\pi} = \frac{n_1 * P_1 + n_2 * P_2}{n_1 + n_2}$$

- (3) 두 비율 차이의 표준편차를 다음과 같이 계산한다

$$\sigma_d = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n_1} + \frac{\hat{\pi}(1 - \hat{\pi})}{n_2}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})(n_1 + n_2)}{n_1 n_2}}$$

- (4) 표본에서 얻은 통계량을 z값으로 표준화시켜 계산하면 아래의 식과 같다
- (5) 계산된 z값을 임계값과 비교하여 의사결정을 한다

$(p_1 - p_2)$ 를 검정하기 위해 계산된 z값

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sigma_d}$$

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 비율의 차이에 관한 가설검정

예제 12-10

제 11장의 예제 5를 다시 한번 인용하여 보자. 조미료 A와 B가 경쟁이 치열하다. A를 생산하는 회사에서는 B를 사용하였던 사람들 중에 40%이하만 B를 계속 사용하며, 60%이상이 자기회사 제품으로 바꾼다는 주장을 하고 있다. 이러한 A회사의 주장을 실증적으로 조사하기 위해 B회사에서 표본을 뽑은 결과 500명 중 215명이 그대로 B제품을 사용하고 있었다. 또 다른 표본을 뽑아보았더니 400명 중에 192명이 B제품을 사용하고 있었다. 두 표본이 모두 같은 모집단에서 산출되었는가를 $\alpha = 0.01$ 에서 검정하시오.

$$\begin{aligned}n_1 &= 500 \\X_1 &= 215 \\p_1 &= 0.43\end{aligned}$$

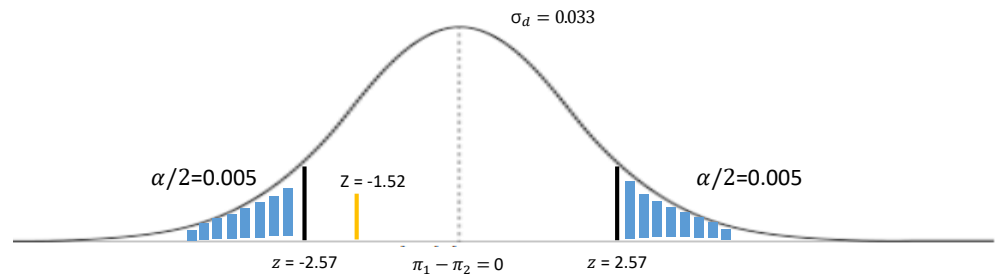
$$\begin{aligned}n_2 &= 400 \\X_2 &= 192 \\p_2 &= 0.48\end{aligned}$$

위의 결과를 가지고 가설검정을 하면 다음과 같다.

$$\begin{aligned}(1) H_0 &: \pi_1 = \pi_2 \\H_1 &: \pi_1 \neq \pi_2\end{aligned}$$

$$(2) \alpha = 0.01$$

$$\begin{aligned}(3) \text{채택영역} &: -2.57 \leq Z \leq 2.57 \\ \text{기각영역} &: Z > 2.57 \text{ 또는 } Z < -2.57\end{aligned}$$



Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 비율의 차이에 관한 가설검정

예제 12-10 - 계속

(4) 모집단의 비율 π 를 추정하기 위하여 가중평균을 내어, 이로부터 σ_d 를 구하면 다음과 같다

$$\hat{\pi} = \frac{n_1 * P_1 + n_2 * P_2}{n_1 + n_2} = \frac{215 + 192}{500 + 400} = 0.45$$

$$\sigma_d = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})(n_1+n_2)}{n_1 n_2}} = \sqrt{\frac{0.45 * 0.55 * 900}{200,000}} = 0.033$$

따라서 계산된 Z값은 아래와 같다

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sigma_d} = \frac{(0.43 - 0.48) - 0}{0.033} = -1.52$$

$Z = -1.52$ 는 채택영역안에 있으므로 귀무가설을 기각할 수 없다.

(5) 두 표본이 다른 모집단에서 나왔다고 할 수 없다.

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 두 모집단 분산의 차이에 관한 가설검정

* 앞서 chapter 11-3에서는 단일모집단 분산에 관한 가설검정을 설명하였으나, 두 모집단의 분산을 비교하기 위한 가설검정도 필요할 때가 있다. 그 예는 다음과 같다

* 우리나라와 대만의 인당 소득이 비슷하다는 통계가 발표되었다.

이에 대해 어떤 학생이 우리나라는 빈부의 차이가 커서, 개인소득의 분산을 보면 그 수치가 매우 크며, 대만에서는 부의 분배가 잘되어, 서로의 소득격차가 작기 때문에 국민의 개인소득의 분산이 작을 것이라는 주장을 한다고 하자.

* 이를 검정하기 위해서는 두 나라의 소득분포가 모두 정규분포를 이룰 것이라는 가정하에 두 모집단으로부터 각각 표본을 추출하여 분산 또는 표준편차를 계산하여 비교하는 방법을 택할 수 있다. 두 개의 분산을 비교하여 가설을 검정할 때에는 F-분포를 사용하게 되므로, 먼저 F-분포에 대해 간단하게 설명한다.

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ F-분포

* F-분포는 1920년대 R.A. Fisher에 의해서 규정된 분포로 그를 기념하기 위하여 후에 F-분포라 한다. F분포는 향후 제 13장에서 설명하게 될 두 개 이상의 평균차를 검정하는 분산분석방법이나 두 분산의 차이를 검정하는 경우에 적용되는 등 상당히 광범위하게 사용되는 분포다.

* F-분포는 각각의 자유도로 나누어진 두 개의 χ^2 -분포의 비율로 이루어지며, 아래와 같이 표현가능하다

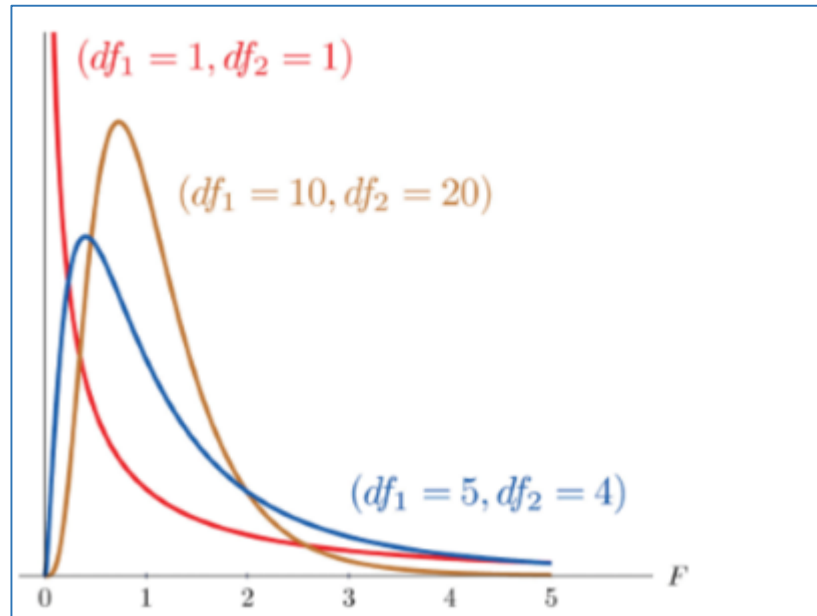
F-분포

$$F(n_1 - 1, n_2 - 1) = \frac{X_1^2 / (n_1 - 1)}{X_2^2 / (n_2 - 1)} = \frac{S_1^2}{S_2^2}$$

* F값은 언제나 +기호를 갖게 되는데 그 이유는 S_1^2 과 S_2^2 이 모두 양수이기 때문이다. F-분포는 S_1^2 의 자유도 $(n_1 - 1)$ 과 S_2^2 의 자유도 $(n_2 - 1)$ 에 따라서 그 모양이 달라진다. 분자의 자유도와 분모의 자유도에 따라 달라지는 F-분포의 모양을 몇 가지 그려보면 그림 12-2와 같다. F-분포는 n 의 크기 n_1, n_2 가 크면 정규분포에 근접한다

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ F-분포



* F값은 두 분산의 비율로써 계산이 되기 때문에 s_1^2 과 s_2^2 이 비슷하면 F값은 1에 가까워진다. 그러나 두 표본의 분산으로부터 계산된 F값이 F-분포표의 임계치보다 매우 크다면, 이 표본들은 분산 σ^2 이 서로 다른 모집단에서 뽑혔다고 할 수 있다. F-분포는 위에서 말한 바와 같이 두 개의 자유도에 의해 결정되는 확률분포이므로 두 모집단에서 뽑힌 표본의 크기 n_1 과 n_2 에 따라 가설검정의 임계치가 달라진다

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 가설의 설정과 임계값의 결정

가설의 설정

* 두 모집단의 분산을 비교하기 위한 가설검정에서도 양측검정이나 단측검정을 할 수 있다. 그러나 이 경우에는 작은 수치를 가진 표본분산을 분모로 하면 F값은 항상 1보다 크므로 F검정에서는 단측검정을 하는 것이 보통이다.

따라서 단측 검정에서의 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 < \sigma_2^2 \quad \text{또는} \quad \sigma_1^2 > \sigma_2^2$$

임계값의 결정

* 가설의 유형과 유의수준이 결정되면, 이에 따라 채택영역과 기각영역을 결정할 임계값을 찾아야 한다. F의 임계값은 두 자유도 df_1 과 df_2 에 따라, 또 유의수준에 따라 달라지므로 매우 많은 표를 필요로 한다. 따라서 표가 매우 복잡해지므로 보통 흔히 쓰이는 유의수준인 $\alpha = 0.05$ 와 $\alpha = 0.01$ 등의 표만을 제시한다.

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 가설의 설정과 임계값의 결정

임계값의 결정

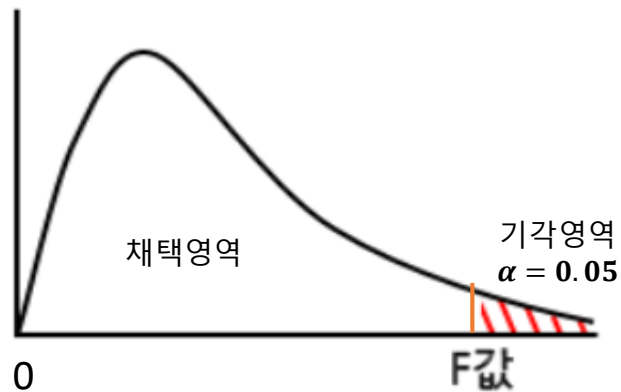
F -분포표($\alpha = 0.05$)

| $df_2 \backslash df_1$ | | 문자의 자유도(df_1) | | | | | | | | | |
|--|----|-------------------|------|------|------|------|------|------|------|------|------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 분 모 의 자 유 도 (df_2) | 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 |
| | 2 | 18.5 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 |
| | 3 | 10.1 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 |
| | 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 |
| | 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 |
| | 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 |
| | 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 |
| | 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 |
| | 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 |
| | 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 |
| | 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 |
| | 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 |
| | 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 |
| | 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 |
| | 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 |

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 가설의 설정과 임계값의 결정

* $n_1 = 10$, $n_2 = 12$ 의 표본을 뽑아, 이들 두 표본으로부터 $S_1^2 = S_2^2$ 을 계산하였다. 이를 근거로 두 모집단의 분산이 동일한가를 알아보려고 한다. $\alpha = 0.05$ 라고 하면 $df_1 = 10 - 1 = 9$, $df_2 = 12 - 1 = 11$ 이므로, 단측검정에서의 임계값에 해당하는 F값은 2.90이다. 이를 나타내면 아래와 같다.



* 앞에서 두 모집단 분산에 대한 가설검정을 할 때 F-통계량을 구하려면 항상 작은 수치의 분산을 분모로 놓아야 한다고 했다. 그럴 경우 F값은 항상 1보다 크므로 F-검정에서는 단측검정을 한다.

Chapter 12 두 모집단의 비교에 관한 가설검정

◎ 가설의 설정과 임계값의 결정

예제 12-11

영도중학교에서 1학년 학생들 성적의 차이가 2학년이 되면 더 커질 것이라는 판단하에 실제로 그러한가를 알아보려고 한다. 두 학년의 성적 분포는 정규분포일 것이라고 가정을 하였다.

1학년에서 7명을 뽑고 2학년에서 9명을 뽑아서 각각의 성적의 분산을 조사하여 본 결과 1학년의 분산은 9.0이었으며, 2학년의 분산은 19.8이었다. 두 모집단의 분산은 같다고 볼 수 있을까? (단, $\alpha = 0.05$)

(1) $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 > \sigma_2^2$

(2) $\alpha = 0.05$

(3) $F_{8,6}$ 에서 $\alpha = 0.05$ 에 해당하는

의사결정의 임계값은 $F = 4.15$ 이므로

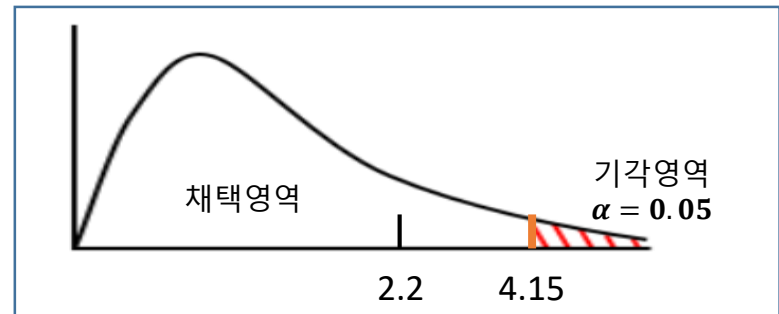
채택영역 : $F \leq 4.15$

기각영역 : $F > 4.15$ 가 된다.

(4) $F = 19.8/9 = 2.2$

$F = 2.2$ 는 채택영역 안에 있으므로 귀무가설을 기각할 수 없다.

(5) 2학년 학생의 성적의 차이가 1학년 학생의 성적의 차이보다 크다고 할 수 없다.



Chapter 13. 분산분석

Chapter 13 분산분석

◎ INTRO

- ✓ 두 모집단의 평균을 비교하기 위해 Z-검정과 t-검정을 사용, 그러나 일반적으로 여러 모집단의 평균을 동시에 비교해야 할 경우가 많이 있다.
- ✓ 만약, 3개의 매체 간 광고효과의 차이를 보기 위해 t검정이나 Z검정을 한다고 하면, ${}_3C_2 = 3$ 번의 검정이 필요함, 만약 개별광고의 방법까지 추가한다면, ${}_4C_2 = 6$ 번의 비교가 이뤄져야 할 것이다
- ✓ 이러한 세 집단 이상을 비교할 때 t-검정이나 Z-검정을 사용하면 번거로울 뿐만 아니라 또 다른 문제가 생긴다. 즉, 두 집단을 여러 번 비교하게 되면 귀무가설이 맞음에도 기각할 확률인 1종오류(α 알파오류)가 커진다
- ✓ 예를 들어 네 집단을 비교하여 생기는 가설검정의 오류가 $\alpha = 0.05$ 가 되기를 원할 때, t-검정을 한다면 여섯 번을 비교해야 하므로 $\alpha = 1 - (1 - \alpha)^6 = 1 - (1 - 0.05)^6 = 0.265$ 가 된다. 이는 $\alpha = 0.05$ 의 유의수준에서 6개의 독립된 t-검정을 할 때 α 오류를 범할 확률은 결과적으로 **26.5%**가 됨을 의미한다.

Chapter 13 분산분석

◎ 분산분석의 개념

분산분석은 F-분포를 처음 개발한 영국의 통계학자인 피셔(R.A.Fisher, 1890~1962)에 의해 소개되었다. 이 분석방법은 미리 정해진 오류를 유지하면서 3개 이상의 모집단 평균이 서로 같은지의 여부를 검증할 수 있게 해준다.

분산분석은 독립변수를 몇 개의 수준 또는 범주로 나누고 각 범주에 따라 나누어진 집단 간의 평균차이를 검정하는 것이다. 분산분석에서는 독립변수를 "요인(factor)", 범주는 "수준(level)"이라 부르기도 한다.

독립변수의 수준에 따라 나누어진 각 집단의 평균 간의 차이가 통계적으로 유의한지를 검정하는 것이므로 t-검정을 확대한 것이라 볼 수 있다.

그러나 **분산분석**이 t-검정과 다른 것은, t-검정은 집단들의 평균을 비교하는 반면에, 분산분석은 **집단의 분산을 사용하여 비교한다는 것**이다.

Chapter 13 분산분석

◎ 일원 vs 이원분산분석

독립변수가 하나일 때(일원분산분석)-하나의 독립변수를 여러 개의 수준(신문, 라디오, 텔레비전, 개별광고) 하으로 나누어 광고매체들 간의 광고효과가 차이가 있는가를 알아보는 경우

독립변수가 둘일 때(이원분산분석)- "광고매체"와 "소비자의 나이"가 광고효과에 어떤 영향을 주는 지 알고 싶다면, 이원분산분석(two-way analysis of variance)라 한다

분산분석에서는 독립변수를 요인이라 부르기도 하므로

일원분산분석은 단일요인실험(single-factor experiment),

이원분산분석은 이요인실험(two-factor experiment)라 하기도 한다

◎ 분산분석의 예

예제 13-1

어느 회사에서는 세 개의 서로 다른 기계를 사용하여 제품을 생산하고 있는데, 각각의 기계가 1시간 안에 생산하는 제품의 양을 다섯 차례 관찰하여 적은 결과가 아래의 표에 나타나 있다.

| 기계 | 각 기계의 생산량 | | | | | \bar{X}_i |
|----|-----------|----|----|----|----|-------------|
| 1 | 47 | 53 | 49 | 50 | 46 | 49 |
| 2 | 55 | 54 | 58 | 61 | 52 | 56 |
| 3 | 54 | 50 | 51 | 51 | 49 | 51 |

Chapter 13 분산분석

◎ 분산분석의 예

예제 13-1 - 계속

앞 페이지의 표에서 보면 기계 1을 1시간씩 다섯 번 조사한 결과 시간당 생산량은 각각 47,53,49,50,46이었으며, 평균은 49이다.

같은 방법으로 측정한 기계 2의 평균은 56, 기계 3의 평균은 51이다. 이 때 다섯 번의 표본생산량에 기초하여 세 기계의 평균 생산량은 동일하다고 볼 수 있는가? 이와 같은 문제에 대한 답을 제시하는 것이 “분산분석”이다.

이 예의 귀무가설은 “세 기계의 평균 생산량이 모두 동일하다”이며, 대립가설은 “평균 생산량이 모두 동일하지는 않다”가 된다. 즉 μ_1 을 기계 1의 평균 생산량, μ_2 을 기계 2의 평균 생산량, μ_3 을 기계 3의 평균 생산량이라 한다면 귀무가설과 대립가설은 다음과 같다.

분산분석의 가설설정

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_1 : 모든 평균이 동일하지는 않다
(즉, 평균이 서로 다른 기계가 있다.)

그러나, 세 집단 중에서 어느 집단이 서로 다른지는 알 수 없다. 즉, 대립가설은 “ $\mu_1 \neq \mu_2 \neq \mu_3$ ”가 아니다.

Chapter 13 분산분석

◎ 분산분석의 예

분산 분석은 위의 가설을 검정하기 위해 생산량의 변동 또는 **분산을 요인의 수준 차이에 기인한 부분**과 **우연 또는 오차에 의한 부분으로 분해한 다음**, 전자가 후자보다 충분히 클 때 요인의 수준에 따라 집단 간 차이가 있는 것으로 판단한다. **전자**가 **후자**보다 충분히 클 때 요인의 수준에 따라 집단 간 차이가 있는 것으로 판단한다. 아래의 경우를 다시 살펴보자.

표-1

| 기계 | 각 기계의 생산량 | | | | | \bar{X}_i |
|----|-----------|----|----|----|----|-------------|
| 1 | 57 | 32 | 53 | 38 | 65 | 49 |
| 2 | 36 | 49 | 64 | 71 | 60 | 56 |
| 3 | 57 | 69 | 48 | 36 | 45 | 51 |

표-2

| 기계 | 각 기계의 생산량 | | | | | \bar{X}_i |
|----|-----------|----|----|----|----|-------------|
| 1 | 48 | 49 | 49 | 49 | 50 | 49 |
| 2 | 56 | 55 | 56 | 57 | 56 | 56 |
| 3 | 50 | 51 | 51 | 52 | 51 | 51 |

Chapter 13 분산분석

◎ 분산분석의 예

앞의 두 표를 비교해 보면, 세 기계에서 만들어진 생산량의 평균은 같지만 1시간마다 조사한 개별 생산량은 다르게 나타내고 있다.

표 1에서는 **1시간 마다의 생산량에 차이가 많다**. 그러나 표 2에서는 매 시간의 생산량이 상당히 고르게 나타나 있다. 이 두 개의 표를 비교해보면 세 기계의 생산량의 평균이 서로 차이가 있는지에 대한 여부를 단순한 기계 1,2,3에서 얻은 평균만으로 단정할 수는 없다고 생각할 수 있다.

두 개의 표의 평균 생산량이 같다고 할 지라도 표 2의 경우에는 기계 1,2,3의 성능의 차이가 분명히 존재한다고 볼 수 있으며, 따라서 귀무가설은 기각될 것이 분명하다. 왜냐하면 표 2에서의 평균 생산량 \bar{x}_i 들의 차이는 우연이라고 보기가 어렵기 때문이다.

◎ 분산분석의 기본가정

- 분산분석 :: 두 집단 비교를 확장시킨 세 집단 이상을 비교하는 방법
- 분산분석의 가정 :: Z검정이나 t검정을 할 때와 동일한 가정이 적용

가정 1 각 집단에 해당되는 모집단의 분포가 정규분포다.

가정 2 각 집단에 해당되는 모집단의 분산이 같다.

가정 3 각 모집단 내에서의 오차나 모집단 간의 오차는 서로 독립적이다.

Chapter 13 분산분석

◎ 분산분석의 예

앞의 두 표를 비교해 보면, 세 기계에서 만들어진 생산량의 평균은 같지만 1시간마다 조사한 개별 생산량은 다르게 나타내고 있다.

표 1에서는 **1시간 마다의 생산량에 차이가 많다**. 그러나 표 2에서는 매 시간의 생산량이 상당히 고르게 나타나 있다. 이 두 개의 표를 비교해보면 세 기계의 생산량의 평균이 서로 차이가 있는지에 대한 여부를 단순한 기계 1,2,3에서 얻은 평균만으로 단정할 수는 없다고 생각할 수 있다.

두 개의 표의 평균 생산량이 같다고 할 지라도 표 2의 경우에는 기계 1,2,3의 성능의 차이가 분명히 존재한다고 볼 수 있으며, 따라서 귀무가설은 기각될 것이 분명하다. 왜냐하면 표 2에서의 평균 생산량 \bar{x}_i 들의 차이는 우연이라고 보기가 어렵기 때문이다.

◎ 분산분석의 기본가정

- 분산분석 :: 두 집단 비교를 확장시킨 세 집단 이상을 비교하는 방법
- 분산분석의 가정 :: Z검정이나 t검정을 할 때와 동일한 가정이 적용

가정 1 각 집단에 해당되는 모집단의 분포가 정규분포다.

가정 2 각 집단에 해당되는 모집단의 분산이 같다.

가정 3 각 모집단 내에서의 오차나 모집단 간의 오차는 서로 독립적이다.

Chapter 13 분산분석

◎ 일원분산분석

자료의 구성

분산분석을 하기 위해 계산을 하려면 자료가 어떻게 구성되어 있고 각 자료가 어떻게 표시되어 있는지를 알아야 한다. **일원분산분석은 하나의 독립변수가 여러 개의 수준**으로 나누어져 있고, 각 수준에 해당되는 집단에는 **여러 관찰 값들이 포함되어 있는 자료**를 분석하는 것이다.

관찰값은 x_{ij} 로 표시되는데 두 개의 하위부호 중에서 i 는 한 집단 내에서의 위치를 나타내고, j 는 몇 번째의 집단인지를 나타낸다. 예를 들어 x_{23} 은 3번째 집단의 2번째 관찰값을 말한다. 일반적으로 말해서 x_{ij} 는 j 번째 집단의 i 번째 관찰값을 나타낸다.

| 집단 관찰번호 | 집단 1 | 집단 2 | ... | 집단 j | |
|------------|-------------|-------------|-----|-------------|-----------|
| 1 | x_{11} | x_{12} | ... | x_{1j} | |
| 2 | x_{21} | x_{22} | ... | x_{2j} | |
| 3 | x_{31} | x_{32} | ... | x_{3j} | |
| ... | ... | ... | ... | ... | |
| i | x_{i1} | x_{i2} | ... | x_{ij} | |
| | \bar{x}_1 | \bar{x}_2 | ... | \bar{x}_j | \bar{x} |

\bar{x} : 전체평균

\bar{x}_j : j번째 집단의 평균

\bar{x}_{ij} : j번째 집단의 i번째 관찰값

Chapter 13 분산분석

◎ 일원분산분석

관찰값의 모형

관찰값을 x_{ij} 라 하면 x_{ij} 는 다음과 같은 요소로 구성되어 있다.

일원분산분석에서 관찰값의 모형

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

μ : 전체평균

α_j : j번째 집단의 영향

ε_{ij} : j번째 집단에 있는 관찰값 i 의 우연적 오차

각 관찰값은 전체평균 μ 와 , 수준이 다른 집단에 있기 때문에 생기는 전체평균과의 차이 α_j , 그리고 각 집단에 있는 관찰값의 개인차 또는 ε_{ij} 로 이루어져 있음을 알 수 있다. 위의 모형을 통갯값으로 표현하면 다음과 같다

$$X_{ij} = \bar{X} + (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

집단 간 차이 집단 내 차이

$(\bar{X}_j - \bar{X})$: j번째 집단의 평균과 전체평균 간의 차이

$(X_{ij} - \bar{X}_j)$: 각 관찰값과 각 집단평균 간의 차이

위의 식에서 \bar{X} 을 왼쪽 항으로 이항하여 다음과 같이 변형시킬 수 있다.

$$(X_{ij} - \bar{X}) = +(\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

Chapter 13 분산분석

◎ 일원분산분석

관찰값의 모형

$$(X_{ij} - \bar{X}) = +(\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

위의 식에서 왼쪽 식과 오른쪽 식을 각각 제곱하여 전체관찰수만큼 합하면 아래의 식을 도출할 수 있다.
여기서 " $\sum \sum$ "의 표시는 관찰 값을 합할 때 i에 해당되는 모든 관찰값과 j에 해당되는 모든 관찰값을 다 합한다는 것을 의미한다.

$$\sum \sum (X_{ij} - \bar{X})^2 = \sum \sum (\bar{X}_j - \bar{X})^2 + \sum \sum (X_{ij} - \bar{X}_j)^2$$

위의 식에서 오른쪽 항의 $\sum \sum (\bar{X}_j - \bar{X})^2$ 을 집단간 제곱합(sum of squares between groups : SSB)이라 하며, $\sum \sum (X_{ij} - \bar{X}_j)^2$ 은 집단 내 제곱합(sum of squares within groups: SSW), 그리고 왼쪽 항의 $\sum \sum (X_{ij} - \bar{X})^2$ 은 총제곱합(total sum of squares : SST)이라 하는데, 총제곱합은 집단내 제곱합과 집단간 제곱합의 합과 같다.

Chapter 13 분산분석

◎ 일원분산분석

관찰값의 모형

예제 13-2 어느 음료수회사에서 맛은 같고, 색깔은 무색, 분홍, 오렌지, 초록으로 하여 음료수를 제조/판매하고 있다. 음료수 색깔이 음료수 판매량과 관계가 있는가를 알아보기 위하여, 일정한 인구를 갖는 여섯 지역을 선택하여 각각의 판매량을 조사하였더니 아래의 표와 같다.

(단위: 만병)

| 도시 \ 색깔 | 무색(1) | 분홍(2) | 오렌지(3) | 초록(4) | 합계 |
|----------|-------------------------|-------------------------|-------------------------|-------------------------|--------------------------|
| 1 | 26 | 31 | 27 | 30 | |
| 2 | 28 | 28 | 25 | 29 | |
| 3 | 25 | 30 | 28 | 32 | |
| 4 | 29 | 27 | 24 | 31 | |
| 5 | 27 | 29 | 26 | 32 | |
| 6 | 27 | 29 | 26 | 32 | |
| 합계 평균 | 162 $\bar{X}_1 = 27$ | 174 $\bar{X}_2 = 29$ | 156 $\bar{X}_3 = 26$ | 186 $\bar{X}_4 = 31$ | 678 $\bar{X} = 28.25$ |

Chapter 13 분산분석

◎ 일원분산분석

관찰값의 모형

예제 13-2 - 계속

음료수 색깔은 독립 변수에 해당되며 여기서는 독립변수를 네 개의 범주로 나누었다. 판매량은 종속변수에 해당되며, 각 집단의 관찰 값은 여섯 개 지역에서의 음료수 색깔에 따른 판매량이 된다.

판매량의 차이를 색깔로 설명할 수 있을까? 만일 색깔이 판매량에 영향을 미치는 요인이라면 색깔별 판매량의 평균은 동일하지 않을 것이다. 이러한 사실을 규명하기 위해 먼저 귀무가설과 대립가설을 설정하여 보자.

$$H_0 : \mu_1(\text{무색}) = \mu_2(\text{분홍}) = \mu_3(\text{오렌지}) = \mu_4(\text{초록})$$

$$H_1 : \text{적어도 어느 하나는 다르다.}$$

Chapter 13 분산분석

◎ 일원분산분석

제공합

종속변수의 총 분산을 **총제공합(total sum of squares : SST)**이라 하는데, 총제공합은 모든 표본자료의 각 관찰값에서 전체표본의 평균을 뺀 것을 제곱해서 합한 것이다.

총제공합

$$SST = \sum \sum (X_{ij} - \bar{X})^2$$

예제 2에서 **총제공합 SST**는 다음과 같다.

$$\begin{aligned} SST &= \sum \sum (X_{ij} - \bar{X})^2 \\ &= (26 - 28.25)^2 + (28 - 28.25)^2 + \dots + (32 - 28.25)^2 + (32 - 28.25)^2 \\ &= 126.5 \end{aligned}$$

집단간 제공합(sum of squares between groups : SSB)은 음료수 색깔이라는 요인에 의해 설명되는 부분이다. 이는 **각 집단 평균값과 전체평균 간의 제공합**이며, 이때 각 집단의 관찰수(n_i) 만큼 곱해주어야 한다.

Chapter 13 분산분석

◎ 일원분산분석

제공합

집단간 제공합

$$SSB = \sum n_i (\bar{X}_j - \bar{X})^2$$

예제 2에서 **집단간 제공합 SSB**는 다음과 같다.

$$\begin{aligned} SSB &= \sum n_i (\bar{X}_j - \bar{X})^2 \\ &= 6 (27 - 28.25)^2 + 6 (29 - 28.25)^2 + \dots + 6 (26 - 28.25)^2 + 6 (31 - 28.25)^2 \\ &= 88.5 \end{aligned}$$

총제공합 중에서 색깔이란 변수로도 설명되지 않는 부분을 **집단내 제공합(sum of squares within groups : SSW)**이라고 하는데, 이는 각 집단내 개별 관찰값 i 의 우연적 오차를 말한다.

집단내 제공합 SSW는 각 집단에 있는 관찰값들과 그 집단의 평균

Chapter 13 분산분석

◎ 일원분산분석

제공합

집단내 제공합

$$SSW = \sum \sum (x_{ij} - \bar{x}_j)^2$$

예제 2에서 집단간 제공합 SSW는 다음과 같다.

$$\begin{aligned} SSW &= \sum \sum (x_{ij} - \bar{x}_j)^2 \\ &= \sum (X_{i1} - \bar{X}_1)^2 = (26 - 27)^2 + \dots + (27 - 27)^2 = 10 \\ &= \sum (X_{i2} - \bar{X}_2)^2 = (31 - 29)^2 + \dots + (29 - 29)^2 = 10 \\ &= \sum (X_{i3} - \bar{X}_3)^2 = (27 - 26)^2 + \dots + (26 - 26)^2 = 10 \\ &= \sum (X_{i4} - \bar{X}_4)^2 = (30 - 31)^2 + \dots + (32 - 31)^2 = 8 \end{aligned}$$

$$SSW = 10 + 10 + 10 + 8 = 38$$

Chapter 13 분산분석

◎ 일원분산분석

제공합

$$SSW = 10 + 10 + 10 + 8 = 38$$

총제공합은 집단간 제공합과 집단내 제공합으로 구성되어 있으며, 각각의 자유도는 다음과 같다

$$\begin{aligned} \text{제공합 : } SST &= SSB + SSW \\ \text{자유도 : } (N-1) &= (J-1) + (N-J) \end{aligned}$$

분산분석의 계산절차를 보여주기 위하여 **SST**, **SSB**, **SSW**를 모두 계산하였지만, 일반적으로 이들 중 두 개만 계산하여도 된다. 예를 들면 **SST**와 **SSB**만을 구한 다음 **SSW = SST - SSB**의 방식으로 구한다.

Chapter 13 분산분석

◎ 일원분산분석

평균제곱

분산분석을 하기 위해서는 제곱합으로 계산된 SSB와 SSW를 자유도로 나누어 평균제곱(mean squares)을 구해야 한다. 이 개념은 제곱합을 관찰수로 나눈다는 점에서 "분산"과 유사한 개념이다.

만일 집단의 수는 적고 각 집단의 대상자 수가 아주 많을 때 집단내 분산(SSW)은 집단간 분산(SSB)에 비해 아주 커질 것이다. 이러한 경우 집단간 분산과 집단내 분산을 해당 자유도로 나누어주면 분산의 정도에 대해 표준화된 수치가 나온다.

집단간 제곱합을 자유도로 나눈 것을 **집단간 평균제곱(mean squares between groups : MSB)**이라 하며, 집단내 제곱합을 자유도로 나눈 것을 **집단내 평균제곱(mean squares within groups)**이라 한다. 집단의 수를 J개라 하면 평균제곱을 구하는 식은 다음과 같다.

평균제곱

| | |
|-----|-------------------------|
| 집단간 | $MSB = \frac{SSB}{J-1}$ |
|-----|-------------------------|

| | |
|-----|-------------------------|
| 집단내 | $MSW = \frac{SSW}{N-J}$ |
|-----|-------------------------|

Chapter 13 분산분석

◎ 일원분산분석

평균제곱

예제 2에서 평균제곱은 아래와 같다.

평균제곱

$$\text{집단간} \quad \text{MSB} = \frac{SSB}{J-1} = \frac{88.5}{3} = 29.5$$

$$\text{집단내} \quad \text{MSW} = \frac{SSW}{N-J} = \frac{38}{20} = 1.9$$

F-통계량

분산분석은 두 종류의 분산, 즉 **집단간 분산**과 **집단내 분산** 간의 비율을 구하는 방식으로 **F-검정**을 한다. 만일 집단간 분산이 집단내 분산에 비해 그 비율이 크면 집단에 따른 차이가 크다는 것을 의미한다. 즉 독립변수를 몇 개의 수준으로 나누어 그 차이를 알아내는 것이 의미있다는 것을 말한다.

F-검정을 하기 위

분산분석에서의 F-통계량

$$F_{J-1, N-J} = \frac{MSB}{MSW}$$

Chapter 13 분산분석

◎ 일원분산분석

분산분석표와 가설검정

지금까지 설명한 분산분석의 과정을 간단하게 표로 만들어 볼 수 있는데, 이를 분산분석표 (ANOVA table)이라고 한다. 일원분산분석에서의 분산분석표를 예로 들면 아래의 표와 같다

| 분산원 | 제곱합 | 자유도 | 평균제곱 | F값 |
|-----|--|-------|-------------------------|-------------------|
| 집단간 | $SSB = \sum n_i(\bar{X}_j - \bar{X})^2$ | J-1 | $MSB = \frac{SSB}{J-1}$ | $\frac{MSB}{MSW}$ |
| 집단내 | $SSW = \sum \sum (X_{ij} - \bar{X}_j)^2$ | N-J | $MSW = \frac{SSW}{N-J}$ | |
| 합계 | $SST = \sum \sum (X_{ij} - \bar{X})^2$ | N - 1 | | |

맨 오른쪽에 있는 F값은 집단간 분산 MSB를 집단내 분산 MSW로 나눈 것이다. 이 F값은 각 집단이 정규분포인 모집단에서 추출되었으며, 각 모집단의 분산은 동일하다는 가정하에서 계산된 것이다. 청음료의 예를 가지고 표 13-6에 따라 분산분석표를 만들어 보면 아래의 표와 같게 된다.

| 분산원 | 제곱합 | 자유도 | 평균제곱 | F값 |
|-----|-------|-----|------|--------|
| 집단간 | 88.5 | 3 | 29.5 | 15.526 |
| 집단내 | 38.0 | 20 | 1.9 | |
| 합계 | 126.5 | 23 | | |

Chapter 13 분산분석

◎ 일원분산분석

분산분석표와 가설검정

| 분산원 | 제곱합 | 자유도 | 평균제곱 | F값 |
|-----|-------|-----|------|--------|
| 집단간 | 88.5 | 3 | 29.5 | 15.526 |
| 집단내 | 38.0 | 20 | 1.9 | |
| 합계 | 126.5 | 23 | | |

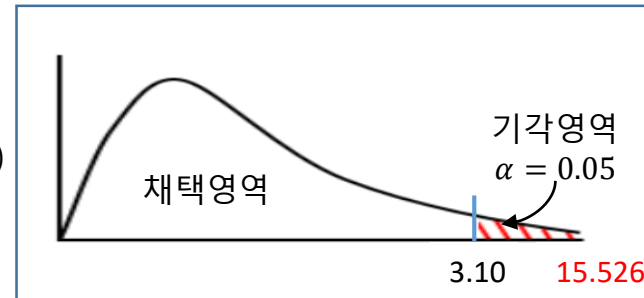
위의 분산분석표를 이용하여 가설검정을 해보자.

- (1) $H_0 : \mu_1(\text{무색}) = \mu_2(\text{분홍}) = \mu_3(\text{오렌지}) = \mu_4(\text{초록})$
 $H_1 : \text{색깔에 따라서 판매량이 다를 수 있다.}$

- (2) $\alpha = 0.05$

- (3) 임계값 $F_{(3,20)} = 3.10$

채택영역 : $F \leq 3.10$, 기각영역 : $F > 3.10$



- (4) 계산된 F값은 15.526이다

- (5) 계산된 F값이 $\alpha = 0.05$ 에서 기각영역에 있으므로 귀무가설을 기각한다.

즉, 청량음료의 색깔에 따라 판매량이 다르다

Chapter 13 분산분석

◎ 일원분산분석

관계에 대한 설명력 η^2 제공

분산분석 결과 F값이 유의하다는 것은 집단 간의 차이가 오차의 정도를 넘어설 만큼 크다는 것을 의미한다. 그러나 집단 간 차이(독립변수의 영향)때문에 생기는 분산이 총분산(종속변수)을 얼마나 설명하는가를 알려주지는 못한다. 독립변수의 설명력을 알기 위해서 자주 사용하는 지수로 η^2 (에타 제곱)이 있다

η^2 은 결정계수인 γ^2 과 마찬가지로 종속변수에 대한 독립변수의 설명력을 말해 준다. "상관비"(correlation ratio)라고 하는 η^2 은 분산분석에서는 SSB와 SST의 비율로써 독립변수에 의해 설명될 수 있는 분산과 총분산 간의 독립변수의 설명력 η^2

$$\eta^2 = \frac{SSB}{SST}$$

예제 2에서의 음료수 색깔에서의 η^2 을 계산하면 다음과 같다. $\eta^2 = SSB/SST = 88.5/126.5 = 0.70$

이는 "색깔의 종류"라는 독립변수가 종속변수인 "판매량"의 분산을 약 70% 설명한다는 것을 알 수 있다

Chapter 13 분산분석

◎ 일원분산분석

개별집단 평균차에 대한 사후검정

앞에서 우리는 분산분석을 실시하여 여러 집단 간의 차이에 대한 검정을 하였다. 분산분석에서 귀무가설을 받아들였을 때는 여러 집단 간의 평균이 같다는 것을 의미하므로 더 이상의 검정이 필요없게 된다.

그러나 만일 귀무가설을 거부하여 여러 집단 간에 차이가 있는 것으로 나타났다고 하자. 이럴 경우 **모든 집단 간에 차이가 있을 수도 있지만, 때로는 특정집단 간에만 차이가 있을 수 있다.**

그 중 어느 집단들 간에 차이가 있는지를 알기 원한다면 다음 단계의 통계적 비교를 해야 하는데, 이를 사후비교(Post hoc comparison)이라 한다. **왜냐하면, 분산분석의 결과 여러 집단 간의 차이가 있다는 것을 알고 난 후에 실시하기 때문이다.** 사후비교방법은 미리 정한 α -오류의 수준을 그대로 유지하면서 어떤 쌍의 평균차이가 유의한지를 알게 해준다.

사후비교방법에는 여러가지가 있지만, 여기서는 자주 사용되면서도 가장 간단한 방법인 **Tukey방법**을 살펴보자. Tukey검정은 **“진실로 유의있는 차의 검정(Tukey's Honestly Significant Difference)”**간단히 Tukey의 HSD라 한다. Tukey의 방법은 모든 집단의 사례수가 같고 모든 집단 간의 평균값들을 1대 1로 비교할 경우 다른 사후비교방법들보다 통계적 검정력이 강하다 평가된다.

이 방법을 사용하기 위해서는 먼저 귀무가설을 기각할 수 있는 임계치를 결정해야 한다. 두 표본평균값들 간의 차이가 임계치보다 더 크면 두 집단의 평균이 같다는 귀무가설은 기각된다. Tukey의 HSD 방법은 대체로 5%나 1%의 유의수준을 사용하여 양측검정을 수행한다.

Chapter 13 분산분석

◎ 일원분산분석

개별집단 평균차에 대한 사후검정

HSD의 임계치는 다음 공식으로 결정된다.

Tukey검정을 위한 HSD임계치

$$HSD = q * \sqrt{\frac{MSW}{n}}$$

MSW : 집단내 평균제곱

n : 각 집단내 사례수

q : α 수준과 MSW 의 df , 집단수 k 에 의한 통계값

음료수의 색깔에 관한 분산분석의 예를 통해 Tukey방법을 사용하여 사후비교를 수행해보자. 위의 예제에서는 여러 집단들 간의 평균이 같다는 귀무가설을 기각했기 때문에 각 집단의 표본평균값들 간의 차이를 검정할 수 있다.

예제 13-3

앞에서 음료수 색깔과 판매량의 예를 보면 음료수 색깔이 판매량에 영향을 끼친다고 하였다. 네 종류 중에서 어느 것이 서로 차이가 나는지 $\alpha = 0.05$ 수준에서 검정하라.

Chapter 13 분산분석

◎ 일원분산분석

개별집단 평균차에 대한 사후검정

예제 13-3 - 계속

먼저 Tukey's q 값을 확인해보면, $\alpha = 0.05$, MSW 의 자유도는 20, 집단의 수는 $K = 4$ 에서 q 값은 3.96이라는 것을 알 수 있다. 또한 앞의 예제 13-2에서 집단내 평균제곱 $MSW = 1.9$ 라는 사실이 계산되었다. 이를 기초로 HSD의 임계치를 구하면 아래와 같다.

$$HSD = q * \sqrt{\frac{MSW}{n}} = 3.96 * \sqrt{\frac{1.9}{6}} = 2.228$$

이제 표본평균값들 간의 모든 가능한 쌍들에 대한 차이를 나열해보자. 아래의 표의 각 칸 내의 수치는 평균값들 간의 차이를 나타낸다

| 평균치 | (무색) $\bar{X}_1 = 27$ | (분홍색) $\bar{X}_2 = 29$ | (오렌지색) $\bar{X}_3 = 26$ | (초록색) $\bar{X}_4 = 31$ |
|------------------|--------------------------|---------------------------|----------------------------|---------------------------|
| $\bar{X}_1 = 27$ | 0 | -2 | 1 | -4 |
| $\bar{X}_2 = 29$ | | 0 | 3 | -2 |
| $\bar{X}_3 = 26$ | | | 0 | -5 |
| $\bar{X}_4 = 31$ | | | | 0 |

Chapter 13 분산분석

◎ 일원분산분석

개별집단 평균차에 대한 사후검정

예제 13-3 - 계속

절댓값기준으로 HSD의 임계값이 2.228 이상이므로 각 쌍의 평균에 대한 차이를 살펴보면 아래와 같다

$$(\bar{X}_1 - \bar{X}_4) = -4 :: \text{무색과 초록색}$$

$$(\bar{X}_3 - \bar{X}_4) = -5 :: \text{오렌지색과 초록색}$$

$$(\bar{X}_2 - \bar{X}_3) = 3 :: \text{분홍색과 오렌지색}$$

따라서 위의 세 쌍의 평균 판매량 간의 차이만이 $\alpha = 0.05$ 에서 유의하다는 것을 알 수 있다. 사후비교는 사전의 이론적 배경에 의해서가 아니라, F-검정 결과 유의한 차이가 있을 때 그 차이가 어떤 집단의 평균에 의해서 비롯되었는지를 알기 원할 때 실시한다.

Chapter 13 분산분석

◎ 이원분산분석

기본개념

주효과와 상호작용효과

이원분산분석은 **두 개의 독립변수** 중 어느 변수에 분산의 원인이 있는지 그리고 두 변수의 상호작용은 어떠한지 등을 알아본다.

예를 들어 어느 회사에서 새로운 상품의 인지도를 높이기 위해 TV, 신문, 라디오를 통한 광고 방법 중에서 어느 방법이 더 효과적인지를 알아보려 한다. 만약, 홍보과에서는 **TV광고는 어린이에게 효과**가 있고, **라디오 광고는 청소년**, 그리고 신문광고는 중년 혹은 장년층에게 효과가 있다고 주장한다고 가정하자.

세 가지 **광고방법**에 의한 효과를 알기 위해서는 **대상자의 연령층**이라는 또 다른 변수와 함께 그 효과를 분석해야 한다. 이렇게 된다면, 결국 **독립변수가 두 개**가 되는데, 이 두 변수의 효과를 알아보는 것을 **주효과(main effect)분석**이라 한다.

이원분산분석은 이렇게 **두 개의 주효과 간의 상호작용효과(interaction effect)**를 알아볼 수 있다는 장점도 존재한다. 만약 일원분산분석을 두 번 한다 해도 상호작용효과는 알 수 없다.

Chapter 13 분산분석

◎ 이원분산분석

기본개념

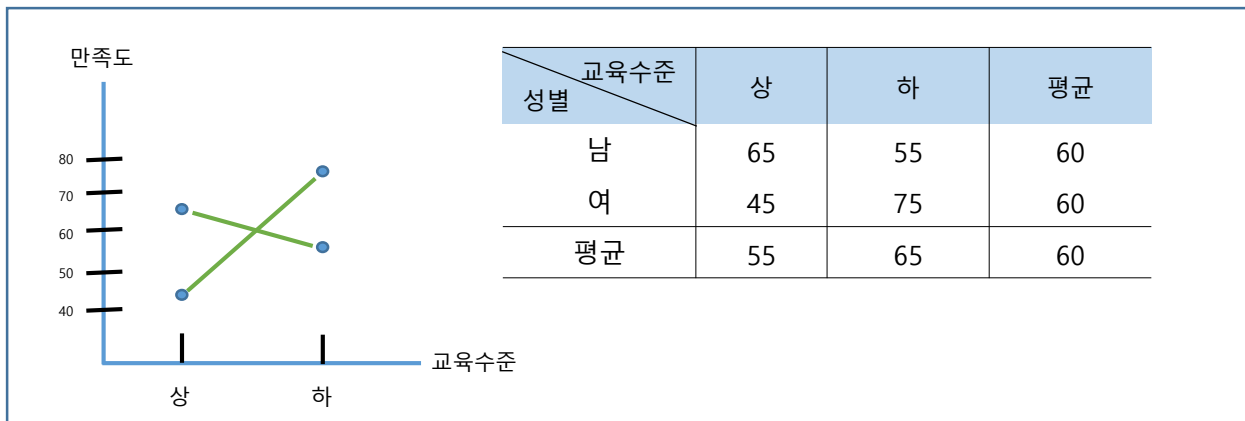
상호작용효과의 예

상호작용효과를 쉽게 이해하기 위해 다른 예를 들어보자. 결혼만족도가 교육 수준에 따라 어떻게 차이가 나는지를 알기 원한다는 사례가 있다라고 하자.

기존연구 :: 교육수준(x_1)이 높을수록 결혼만족도(y)가 낮다는 결과

다른연구 :: 성별(x_2)에 따라 결혼만족도(y)의 결과는 다르게 도출될 수 있음

상호작용효과 – ‘결혼만족도(y)’는 ‘교육수준’, ‘성별’을 모두 적용하였을 때 결혼만족도가 어떻게 달라지는지를 확인하는 것임. 결혼만족도를 알아보기 위해 여러 문항으로 구성된 검사지를 실시한 결과 아래와 같은 점수가 나왔다고 하자. 표안의 숫자는 만족도에 대한 각 집단의 평균 점수이다.



Chapter 13 분산분석

◎ 이원분산분석

기본개념

상호작용효과의 예

앞 페이지에서의 표를 보면 교육수준을 고려하지 않고 남녀의 결혼만족도(둘 다 평균이 60)는 매우 똑같다. 하지만, 또 다른 변수인 **교육수준**을 보면 **교육을 많이 받은 사람의 만족도는 55**인 데 비해 **교육수준이 낮은 사람의 만족도는 65**로 높게 나왔다.

따라서, 두 변수 중 '교육수준'만이 "결혼만족도"라는 종속변수에 영향을 끼치는 것으로 표에선 나타났다. 그러나 앞 페이지에서의 표와 그림을 살펴보면 이는 충분한 결론이라 할 수 없다. **남자의 경우에는 교육수준이 낮을 때 만족도가 낮지만, 여자의 경우에는 교육수준이 높을 때 오히려 만족도가 낮다는 것**을 알 수 있다. 이러한 경우, 즉 한 독립변수가 다른 독립변수의 수준에 따라 달리 작용하므로 상호작용의 효과가 있다는 것을 알 수 있다.

상호작용이 있는지를 쉽게 알기 위해서는 두 변수에 의한 평균값의 차를 구해보면 된다. **성별**에 따라 교육수준을 비교해보면 (남자의 교육수준의 차 :: $65-55=10$), (여자의 교육수준의 차:: $45-75=-30$)이며, **교육**에 따라 성별을 비교해보면(교육수준 '상'의 차 :: $65-45=20$), (교육수준 '하'의 차 :: $55-75 =-20$)이다.

상호작용이 있는 경우에는 한 독립변수의 수준에 따라 다른 변수의 평균값의 차의 부호가 서로 다르다. 또한 상호작용이 있는 경우 그림을 그려보면 두 선의 방향이 서로 다르거나 어긋난다.

Chapter 13 분산분석

◎ 이원분산분석

자료의 구성

이원분산분석은 두 개의 독립변수가 있고, 이 두 개의 독립변수는 또 다시 몇 개의 수준으로 나누어진다.

첫번째 독립변수를 A라 하고 나누어진 수준의 개 수를 J라 하자. 그리고 두 번째 독립변수를 B라 하고 나누어진 수준의 개수를 K라 하자. 관찰값은 X_{ijk} 로 표시하는데, i는 집단 내의 위치를 나타내고, j는 첫번째 독립변수의 수준을, 그리고 k는 두 번째 독립변수의 수준을 나타낸다.

ex) X_{231} 은 독립변수 A의 세번째 수준과 독립변수 B의 첫번째 수준에 해당되는 집단에서 두번째 관찰값을 나타냄

아래의 표는 자료구성과 분석에 필요한 기호를 나타낸다. 독립변수 A는 3개의 수준, 독립변수 B는 2개의 수준으로

| B \ A | A_1 | A_2 | A_3 | $\bar{X}_{.k}$ |
|----------------|-------------------------------------|-------------------------------------|-------------------------------------|----------------|
| B_1 | X_{111} X_{211} X_{311} | X_{121} X_{221} X_{321} | X_{131} X_{231} X_{331} | $\bar{X}_{.1}$ |
| \bar{X}_{jk} | \bar{X}_{11} | \bar{X}_{21} | \bar{X}_{31} | |
| B_2 | X_{112} X_{212} X_{312} | X_{122} X_{222} X_{322} | X_{132} X_{232} X_{332} | $\bar{X}_{.2}$ |
| \bar{X}_{jk} | \bar{X}_{12} | \bar{X}_{22} | \bar{X}_{32} | |
| $\bar{X}_{j.}$ | $\bar{X}_{1.}$ | $\bar{X}_{2.}$ | $\bar{X}_{3.}$ | \bar{X} |

\bar{X} : 전체평균
 $\bar{X}_{j.}$: 독립변수 A의 j수준에 있는
 관찰값들의 평균
 $\bar{X}_{.k}$: 독립변수 B의 k수준에 있는
 관찰값들의 평균
 \bar{X}_{jk} : 독립변수 A의 j수준과 독립
 변수 B의 k수준에 있는
 관찰값들의 평균

Chapter 13 분산분석

◎ 이원분산분석

관찰값의 모형

이원분산분석에서 관찰값 X_{ijk} 는 다음과 같은 수리적 모형을 갖는다.

이원분산분석에서 관찰값의 모형

$$X_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{ijk}$$

X_{ijk} : 독립변수 A의 j 번째 수준과 독립변수 B의 k 번째 수준의

영향을 받은 i 번째 관찰값

μ : 전체평균

α_j : 독립변수 **A의 효과**

β_k : 독립변수 **B의 효과**

$(\alpha\beta)_{jk}$: 두 독립변수 **A, B의 상호작용효과**

e_{ijk} : 관찰값 i 의 개인차 혹은 오차

위의 모형은 실제연구에서 다음과 같은 통계값으로 표현된다.

$$X_{ijk} = \bar{X} + (\bar{X}_{j.} - \bar{X}) + (\bar{X}_{.k} - \bar{X}) + (\bar{X}_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X}) + (X_{ijk} - \bar{X}_{jk})$$

위의 식에서 전체평균 \bar{X} 를 왼쪽 항으로 옮기면 다음과 같다.

$$X_{ijk} - \bar{X} = +(\bar{X}_{j.} - \bar{X}) + (\bar{X}_{.k} - \bar{X}) + (\bar{X}_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X}) + (X_{ijk} - \bar{X}_{jk})$$

$(X_{ijk} - \bar{X})$: 각 관찰값과 전체평균 간의 차이

$(\bar{X}_{j.} - \bar{X})$: 독립변수 **A의 j수준의 영향**

$(\bar{X}_{.k} - \bar{X})$: 독립변수 **B의 k수준의 영향**

$(\bar{X}_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X})$: 독립변수 **A의 j수준과 독립변수 B의 k수준의 상호작용영향**

$(X_{ijk} - \bar{X}_{jk})$: 개인차 혹은 오차

Chapter 13 분산분석

◎ 이원분산분석

관찰값의 모형

앞 페이지의 공식은 아래와 같이 제곱합으로 표현할 수 있다.

$$\text{제곱합 : } SST = SSA + SSB + SSAB + SSW$$

각 제곱합은 아래와 같이 계산이 가능하다.

$$SST = \sum \sum \sum (X_{ijk} - \bar{X})^2 = \sum \sum \sum X_{ijk}^2 - \frac{(\sum \sum \sum X_{ijk})^2}{JKn}$$

$$SSA = \sum \sum \sum (\bar{X}_{j.} - \bar{X})^2 = \sum Kn(\bar{X}_{j.} - \bar{X})^2$$

$$SSB = \sum \sum \sum (\bar{X}_{.k} - \bar{X})^2 = \sum Jn(\bar{X}_{.k} - \bar{X})^2$$

$$SSAB = \sum \sum n \sum (\bar{X}_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X})^2$$

$$SSW = \sum \sum \sum (X_{ijk} - \bar{X}_{jk})^2$$

$\sum \sum \sum X_{ij}^2$ 은 모든 관찰값의 제곱을 합한 것이며, $\sum \sum \sum X_{ijk}$ 는 모든 관찰값을 합한 것이다. n 은 각 집단의 사례수를 의미하므로 JKn 은 전체사례수 N 이 된다.

Chapter 13 분산분석

◎ 이원분산분석

제공합과 자유도

이원분산분석에서 제공합과 자유도는 다음과 같은 관계를 갖는다.

이원분산분석에서 제공합과 자유도

| | |
|-----|--|
| 제공합 | $SST = SSA + SSB + SSAB + SSW$ |
| 자유도 | $JKn-1 = (J-1) + (K-1) + (J-1)(K-1) + JK(n-1)$ |

각 효과를 검정하기 위해 먼저 제공합을 자유도로 나누어 평균제곱을 구하면 다음과 같다.

$$MSA = SSA / (J - 1)$$

$$MSB = SSB / (K - 1)$$

$$MSAB = SSAB / (J - 1)(K - 1)$$

$$MSW = SSW / JK(n - 1)$$

각 효과의 유의도를 검정하기 위해서는 각 평균제곱을 오차의 평균제곱인 MSW로 나누어 F-검정을 한다. 해당 내용은 다음 페이지의 표로 요약이 가능하다

Chapter 13 분산분석

◎ 이원분산분석

제공합과 자유도

| 분산원 | 제공합 | 자유도 | 평균제공 | F값 |
|--------|--------|--------------|-------------------|------------|
| A효과 | SSA | $J-1$ | $SSA/(J-1)$ | MSA/MSW |
| B효과 | SSB | $K-1$ | $SSB/(K-1)$ | MSB/MSW |
| AB상호작용 | $SSAB$ | $(J-1)(K-1)$ | $SSAB/(J-1)(K-1)$ | $MSAB/MSW$ |
| 집단내 | SSW | $N-JK$ | $SSW/JK(n-1)$ | |
| 합계 | SST | $N-1$ | | |

이원분산분석의 예

예제 13-4

S전자회사에서는 근로자의 생산성을 높이는 데 어떤 훈련방법이 효과적인지, 그리고 훈련 방법은 근로자의 숙련도에 따라 그 효과가 달라지는지에 대해 연구하였다. 연구대상으로 선출된 근로자들은 무작위로 각각의 집단에 배정되었다. 훈련방법은 세 종류로 구분되었고, 숙련도는 두 수준으로 구분되었다. 훈련의 결과로 나타난 생산량은 다음 페이지와 같으며 **훈련방법을 독립변수 A, 숙련도를 독립변수 B**라고 하였다.

Chapter 13 분산분석

◎ 이원분산분석

이원분산분석의 예

| | A_1 | A_2 | A_3 | $\bar{X}_{.k}$ |
|----------------------|------------------------|-----------------------|-----------------------|------------------------------------|
| B_1 | 9 8 6 10 7 | 5 8 7 6 4 | 8 7 9 6 5 | 7 |
| \bar{X}_{jk} | 8 | 6 | 7 | |
| B_2 | 3 4 6 5 2 | 4 5 2 5 4 | 7 9 6 5 8 | 5 |
| \bar{X}_{jk} | 4 | 4 | 7 | $\sum \sum \sum X_{ijk} = 180$ |
| $\bar{X}_{j.}$ | 6 | 5 | 7 | $\bar{X} = 6$ |
| $\sum \sum X^2_{ij}$ | 420 | 276 | 510 | $\sum \sum \sum X^2_{ijk} = 1,206$ |

$$SST = \sum \sum \sum (X_{ijk} - \bar{X})^2 = \sum \sum \sum X^2_{ijk} - \frac{(\sum \sum \sum X_{ijk})^2}{JKn} = 1,206 - \frac{(180)^2}{30} = 126$$

$$SSA = \sum \sum \sum (\bar{X}_{j.} - \bar{X})^2 = \sum K n (\bar{X}_{j.} - \bar{X})^2 = 10 * (6 - 6)^2 + 10 * (5 - 6)^2 + 10 * (7 - 6)^2 = 20$$

$$SSB = \sum \sum \sum (\bar{X}_{.k} - \bar{X})^2 = \sum J n (\bar{X}_{.k} - \bar{X})^2 = 15 * (7 - 6)^2 + 15 * (5 - 6)^2 = 30$$

$$SSAB = \sum \sum n \sum (\bar{X}_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X})^2 = 5 * (8 - 6 - 7 + 6)^2 + 5 * (6 - 5 - 7 + 6)^2 + 5 * (7 - 7 - 7 + 6)^2 + 5 * (4 - 6 - 5 + 6)^2 + 5 * (4 - 5 - 5 + 6)^2 + 5 * (7 - 7 - 5 + 6)^2 = 30$$

$$SSW = SST - SSA - SSB - SSAB = 126 - 20 - 30 - 20 = 56$$

Chapter 13 분산분석

◎ 이원분산분석

이원분산분석의 예

평균제곱

$$MSA = \frac{SSA}{(J-1)} = \frac{20}{2} = 10$$

$$MSAB = \frac{SSAB}{(J-1)(k-1)} = \frac{20}{2} = 10$$

$$MSB = \frac{SSB}{(K-1)} = \frac{30}{1} = 30$$

$$MSW = \frac{SSW}{N-JK} = \frac{56}{24} = 2.333$$

| 분산원 | 제곱합 | 자유도 | 평균제곱 | F값 |
|--------|-----|-----|-------|--------|
| A효과 | 20 | 2 | 10 | 4.286 |
| B효과 | 30 | 1 | 30 | 12.859 |
| AB상호작용 | 20 | 2 | 10 | 4.286 |
| 집단내 | 56 | 24 | 2.333 | |
| 합계 | 126 | 29 | | |

독립변수 A효과에 관한 가설검정

(1) $H_0 : \alpha_j = 0$ 또는 $H_0 : \text{훈련방법은 차이가 없다.}$ (4) 계산된 F값은 4.286이므로 H_0 를 기각한다.

$H_1 : \alpha_j \neq 0$ 또는 $H_1 : \text{훈련방법은 차이가 있다.}$ **즉, 훈련방법에 따라 생산성에는 차이가 있다.**

(2) $\alpha = 0.05$

(3) $F_{0.05(2,24)}$ 에서 임계값은 3.40 // 채택영역 : $F \leq 3.40$, 기각영역: $F > 3.40$

Chapter 13 분산분석

◎ 이원분산분석

이원분산분석의 예

독립변수 B효과에 관한 가설검정

- (1) $H_0 : \beta_k = 0$ 또는 $H_0 : \text{숙련도는 영향없다.}$ (4) 계산된 F값은 12.859이므로 H_0 를 기각한다.
 $H_1 : \beta_k \neq 0$ 또는 $H_1 : \text{숙련도는 영향있다.}$ 즉, 숙련도는 생산성에 영향을 준다.

(2) $\alpha = 0.05$

(3) $F_{0.05(1,24)}$ 에서 임계값은 4.26 // 채택영역 : $F \leq 4.26$, 기각영역: $F > 4.26$

독립변수 A,B의 상호작용효과에 관한 가설검정

(1) $H_0 : (\alpha\beta)_{jk} = 0$ 또는 $H_0 : \text{두 변수 간에는 상호작용이 없다.}$

$H_1 : (\alpha\beta)_{jk} \neq 0$ 또는 $H_1 : \text{두 변수 간에는 상호작용이 있다.}$

(2) $\alpha = 0.05$

(3) $F_{0.05(2,24)}$ 에서 임계값은 3.40 // 채택영역 : $F \leq 3.40$, 기각영역: $F > 3.40$

(4) 계산된 F값은 4.286이므로 H_0 를 기각한다.

즉, 훈련방법과 숙련도 두 변수 간에는 상호작용이 있다.

Chapter 13 분산분석

◎ 이원분산분석

일원분산분석과 이원분산분석에서 오차의 비교

근로자의 생산성을 높이기 위해 세 가지 훈련방법을 실시한 다음, 방법들 간에 차이가 있는지를 알아보는 것이 '일원분산분석'이다. 119page에서의 표에서 만일 **근로자의 생산성을 훈련방법(독립변수 A)의 차이만으로** 그 영향을 설명한다면, 훈련방법의 차이는 전체분산에서 다음과 같은 비율(η^2)을 차지한다.

$$\text{독립변수 A의 설명력} : \eta^2 = \text{SSA} / \text{SST} = 20 / 126 = 0.159$$

훈련방법의 차이는 종속변수의 분산 중 15.9%를 설명해준다. **나머지 84.1%의 분산은 결국 오차에 의한 분산이라고 본다.** 전체분산이 정해져 있는 상황에서, **훈련방법의 차이만을 독립변수로 생각했기 때문에**

오차가 이렇게 많을 뿐이지 만일 생산량에 영향을 줄 수 있는 또 다른 독립변수를 발견해 낼 수 있다면 오차

분산은 작아질 수 있다. 그러면 훈련방법의 차이뿐만 아니라 근로자의 숙련도(독립변수 B)의 정도에 따라 분산이 영향을 받는다고 한다면 전체분산에서 숙련도라는 독립변수의 영향만큼 뺄 수 있다. 또한 두 독립변수 간의 상호작용까지 합하면 오차분산은 더욱 작아진다.

$$\text{독립변수 B의 설명력} : \eta^2 = \text{SSB} / \text{SST} = 30 / 126 = 0.238$$

Chapter 13 분산분석

◎ 이원분산분석

일원분산분석과 이원분산분석에서 오차의 비교

독립변수 AB의 상호작용의 설명력 : $\eta^2 = SSAB/SST = 20/126 = 0.159$

따라서 독립변수 A와 B, 그리고 AB상호작용의 설명력은 다음과 같다.

독립변수 A, B, AB 상호작용의 설명력 : $\eta^2 = (SSA + SSB + SSAB) / SST = 70 / 126 = 0.556$

오차분산의 비율 : $1 - 0.556 = 0.444 (44.4\%)$

생산량을 설명할 수 있는 독립변수가 “훈련방법” 하나일 때는 **오차 분산이 84.1%**에 달하였지만 종속변수를 설명할 수 있는 독립변수가 하나 더 포함되고 따라서 상호작용 효과도 생겨나서 **오차분산은 44.4%**로 줄었다. 다시 말하면 종속변수의 변동원인이 어디에 연유하는지를 더 잘 알게 되었다. **만일 “근무연한”과 “결혼여부”에 따라 생산성이 달라진다는 이론적 근거가 있다면, 이들을 독립변수로 추가사용하여 오차를 줄일 수도 있다.**

Chapter 13 분산분석

◎ 이원분산분석

이원분산분석에서 개별집단에 대한 사후검정

일원분산분석에서와 같이 이원분산분석에서도 F-비율이 주효과에 대해 유의한 것으로 나타날 때 어떤 집단의 평균값으로 인해 귀무가설이 거부되었는지를 결정해야 하는 문제가 남는다. 이 경우에도 일원분산분석에서 사용한 것과 마찬가지로 Tukey 방법을 적용한다. 앞의 예에서 두 독립변수인 “훈련방법”과 “숙련도” 모두 통계적으로 유의하였다. 그러나 숙련도는 2개의 수준 밖에 없으므로 개별평균에 대한 사후검정을 다시 할 필요가 없다. **그러나 훈련방법에 관해서는 3개의 평균값 중에서 어떤 쌍의 평균값들이 유의한 차이가 있는지 검정할 수 있다.** 이때의 귀무가설은 아래와 같다.

$$H_{01} : \mu_1 = \mu_2$$

$$H_{02} : \mu_1 = \mu_3$$

$$H_{03} : \mu_2 = \mu_3$$

Tukey검정을 위한 HSD임계치

$$HSD = q * \sqrt{\frac{MSW}{n'}}$$

MSW : 집단내 평균제곱

n' : 각 집단내 사례수

q : α 수준과 MSW의 df , 집단수 k 에 의한 통겅값

집단의 수 $k=3$ 이고 MSW의 자유도는 24일 때, 통계값 q 의 임계값은 [Tukey의 q 값]을 보면 $\alpha = 0.05$ 에서

3.53이다. $n' = 10$, $MSW = 2.333$ 이므로 Tukey의 HSD는 $3.53 * \sqrt{(2.333)/10} = 1.705$

Chapter 13 분산분석

◎ 이원분산분석

이원분산분석에서 개별집단에 대한 사후검정

$$(\bar{X}_{1.} - \bar{X}_{2.}) = 1.0$$

$$(\bar{X}_{1.} - \bar{X}_{3.}) = -1.0$$

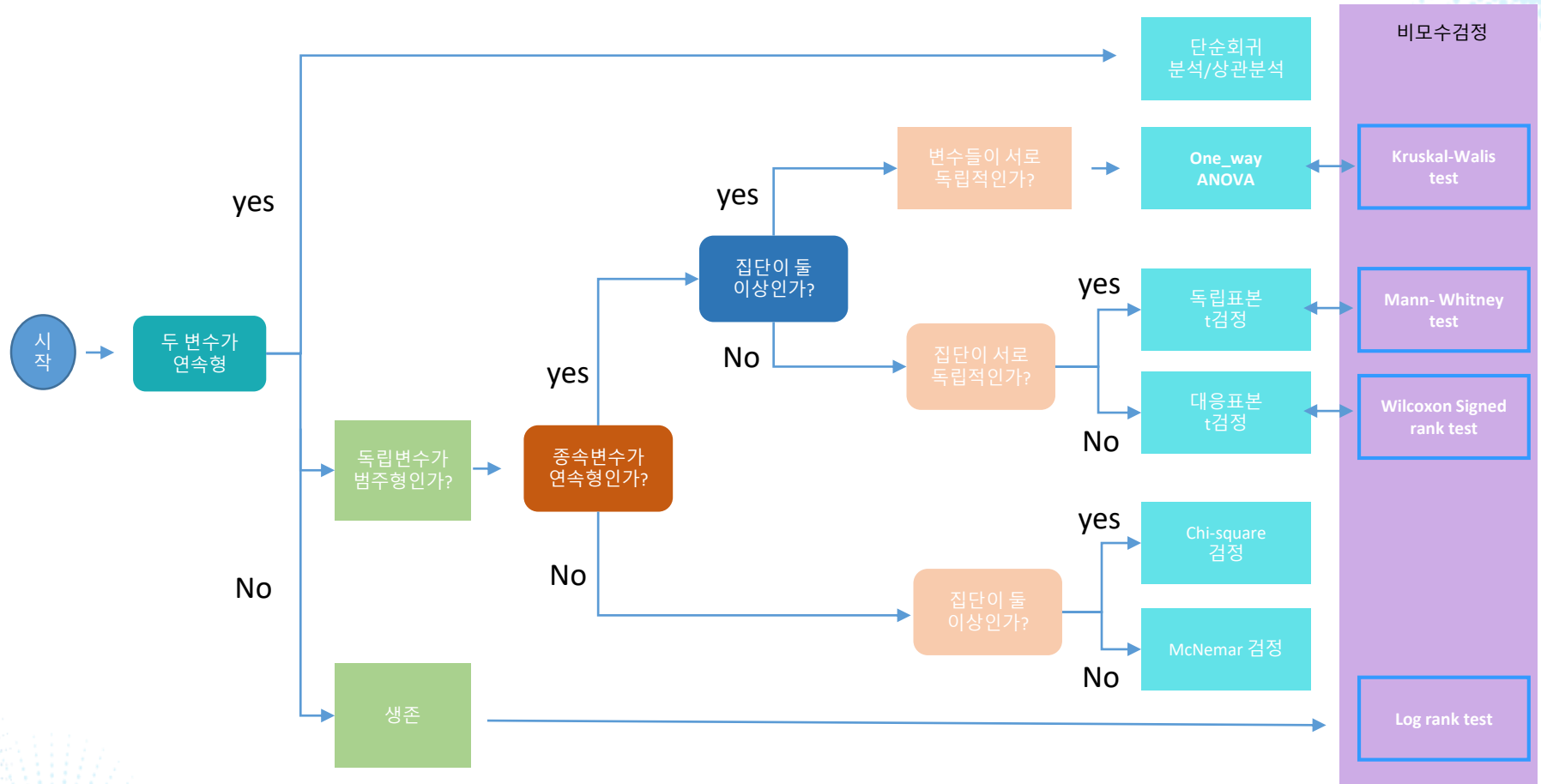
$$(\bar{X}_{2.} - \bar{X}_{3.}) = -2.0$$

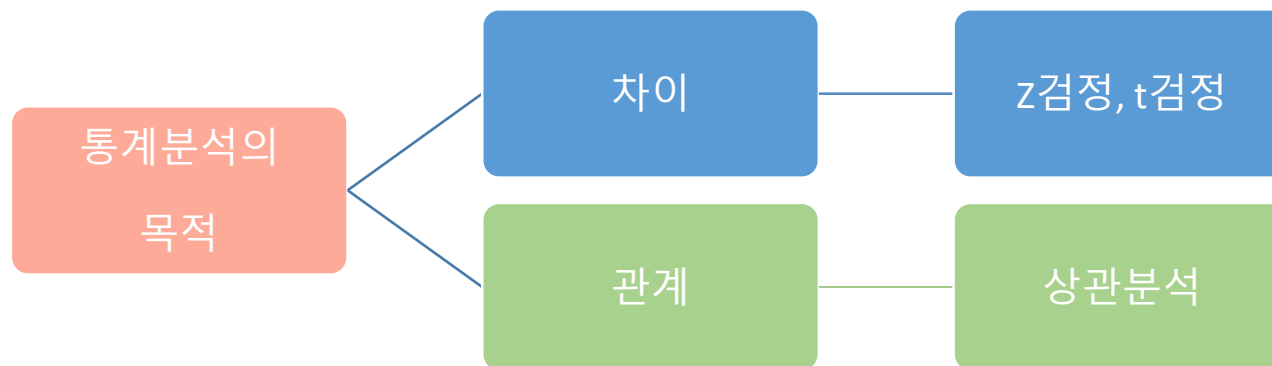
위의 각 평균의 차이에서는 유의수준 $\alpha = 0.05$ 에서 절댓값으로 $HSD = 1.705$ 보다 큰 차이는 훈련방법 2와 훈련방법 3의 평균의 차이만이 존재하므로 두 훈련방법(훈련방법 2와 훈련방법 3)만이 유의한 차이가 있으며,

그 외의 훈련방법1과 2, 훈련방법 1과 3의 경우에는 차이가 있다고 보기가 어렵다.

Chapter 14. 상관분석과 회귀분석의 기초

R 변수종류에 따른 분석모형





- 상관분석(correlation analysis)은 X값과 y값, 두 변수 간 상관관계를 탐색
- 즉 연속적인 두 변수 간의 선형관계를 탐색 및 확인하는 방법

Chapter 14 상관분석과 회귀분석의 기초

◎ INTRO

- ✓ 상관분석은 두 변수 간의 관계의 강도, 즉 얼마나 밀접하게 관련되어 있는지를 분석하는 것을 말한다.
- ✓ 그러나 때로는 관련성 뿐만 아니라 독립변수에 따라 종속변수가 어떻게 변화하는 가를 예측하기를 원하는 때가 있다.
- ✓ 회귀분석에서는 독립변수의 일정한 값에 대응되는 종속변수의 값을 예측하기 위하여 회귀방정식을 구함
- ✓ 상관분석에 의해 상관계수를 계산하고, 만일 상관계수가 높다면 두 변수 간의 관계를 회귀방정식으로 나타냄

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관분석

- 우리의 일상생활이나 학문적 연구에서는 둘 또는 그 이상의 변수들이 서로 어떤 관계를 내포하고 있는지 규명해야 함
- Ex) IQ와 수학점수 간에는 어떠한 관계가 있는가?
제품의 수요가 그 제품의 가격 및 소비자들의 소득과 어떤 관계가 있는가?
- 선형관계(linear relationship)를 규명하기 위한 상관관계 분석(Correlation)

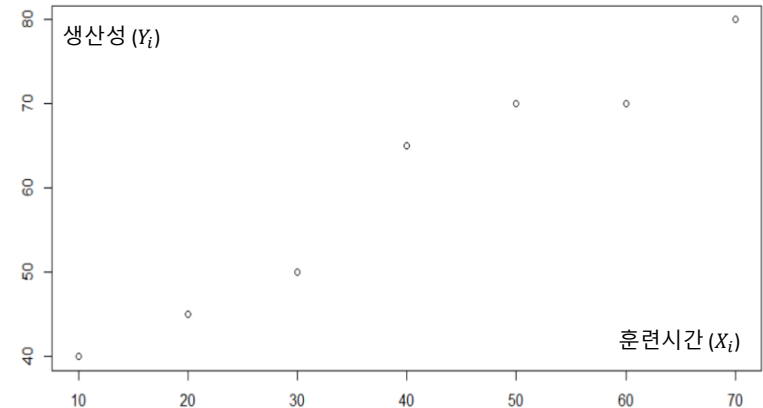
두 변수의 관계와 산포도

- 산포도(Scatter plot)란 두 변수 간의 관계를 알아보기 위하여 두 변수 값을 나타내는 점을 그래프화 한 것
- 산포도를 그리는 것은 매우 중요한데, 만일 분석할 만한 자료가 되지 못한다는 것을 알게 되면 시간과 노력의 낭비를 줄여줄 뿐 만 아니라 자료를 오판하지 않도록 함
- 다음 페이지에서 기능공의 훈련시간에 따른 숙련도를 테이블과 그래프로 살펴보자

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관분석

| 훈련시간 (X_i) | 생산성 (Y_i) |
|----------------|---------------|
| 10 | 40 |
| 20 | 45 |
| 30 | 50 |
| 40 | 65 |
| 50 | 70 |
| 60 | 70 |
| 70 | 80 |



- 이처럼 산포도만을 갖고서 두 변수간의 관계를 정확히 파악하기는 어렵다.
- 두 변수간의 관계를 정확히 파악하기 위해서는 두 변수 간의 관련성의 정도를 계수(correlation coefficient)로 알아보는 상관분석(correlation analysis)
- 두 변수 간의 함수적 관련성을 나타내는 회귀식(regression equation) 또는 예측식(prediction equation)이 필요

Chapter 14 상관분석과 회귀분석의 기초

◎ 공분산의 개념

- 상관분석을 하기 위해서는 공분산(Covariance)에 대한 이해 필요. 왜냐하면 상관분석이란 두 변수가 어떻게 함께 움직이는가를 알아보는 것인데, **공분산 역시 두 변수가 동시에 변하는 정도**를 나타냄
- 두 확률변수의 분포가 결합확률분포를 이룰 때 그 분포의 분산을 공분산이라 하며, $\text{Cov}(X,Y)$ 로 나타냄 Cov를 나타내는 식은 아래와 같음

x와 y의 공분산

$$\text{모집단 } \sigma_{XY} = \frac{\sum (X_i - \mu_X)(Y_i - \mu_Y)}{n}$$

$$\text{표본 } S_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Chapter 14 상관분석과 회귀분석의 기초

◎ 공분산의 개념

- 공분산에서 X변수가 증가할 때 Y변수가 증가하면, 즉 두 변수가 같은 방향으로 변화하면 **공분산의 수치는 +가 됨**. 만일 두 변수가 변화하는 방향이 **서로 다르면 공분산은 -부호**를 가짐
- 공분산의 단점은 두 변수의 측정단위에 따라서 커다란 차이가 나는 문제점이 내재되어 있으므로, 상대적인 강도를 나타내는 좋은 지표가 되지 못함
- Ex) 예를 들어 X와 Y의 공분산을 구할 때, 그 변수들이 센티미터(cm)로 표시되는 경우보다 미터 (m)로 표시되는 경우 **공분산의 절대값은 훨씬 작아짐**
- 위와 같은 **단위의 문제점**을 **해결**하기 위해 **피어슨의 상관계수 r** 을 이용 ::
이는 s_{xy} 를 각 변수의 표준편차 s_x 와 s_y 의 곱으로 나누어서 변수의 단위를 **표준화**

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관계수 r 의 계산

- 두 변수 간 상관과 예측에 관한 통계적 분석방법은 영국의 유전학자인 갈톤(1822 ~ 1911)에 의해 처음 제시
- 그 후 영국의 통계학자인 피어슨(1857~1936)은 이를 더욱 발전시켜 오늘날 흔히 사용하는 상관계수를 제안하게 됨
- 피어슨이 제시한 상관계수 r 은 두 변수 모두 등간 또는 비율척도에 의해 측정된 연속변수에 사용. 또한 피어슨의 r 은 정규분포를 따르는 두 변수 X 와 Y 가 일직선이라는 선형성을 가정할 수 있을 때 사용

모집단과 표본의 상관계수

| | |
|-----|---|
| 모집단 | $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X * \sigma_Y}$ |
|-----|---|

| | |
|-----|-------------------------------------|
| 표 본 | $r_{XY} = \frac{S_{XY}}{S_X * S_Y}$ |
|-----|-------------------------------------|

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관계수 r 의 계산

- 이렇듯 공분산 s_{xy} 를 s_x 와 s_y 의 곱으로 나누면 상관계수(correlation coefficient)는 $-1.0 \leq r \leq 1.0$ 의 범위에 있게 됨
- 따라서 어떠한 단위의 측정값을 사용하여도 상관성에 대한 비교와 해석이 용이. 모집단의 상관계수는 ρ (rho)로 표시하고, 표본집단의 상관계수는 r 로 표시
- 변수 X, Y 가 표준정규분포를 따를 때 s_x 와 s_y 는 1이 되므로 공분산 s_{xy} 는 r 과 같다는 것을 알 수 있음
- 두 변수의 평균치 \bar{X}, \bar{Y} 가 소수점이 나오거나 나누어 떨어지지 않을 때는 계산이 복잡하고 정확치 않으므로 다음과 같이 변형하여 계산하면 편리

상관계수 r_{XY} 의 변형식

$$r_{XY} = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}}$$

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관계수 r의 계산

| 대상자 | 훈련시간 (X_i) | 생산성 (Y_i) | X_i^2 | Y_i^2 | X_iY_i |
|-----|-----------------------|-----------------------|---------|---------|----------|
| 10 | 10 | 40 | 100 | 1,600 | 400 |
| 20 | 20 | 45 | 400 | 2,025 | 900 |
| 30 | 30 | 50 | 900 | 2,500 | 1,500 |
| 40 | 40 | 65 | 1,600 | 4,225 | 2,600 |
| 50 | 50 | 70 | 2,500 | 4,900 | 3,500 |
| 60 | 60 | 70 | 3,600 | 4,900 | 4,200 |
| 70 | 70 | 80 | 4,900 | 6,400 | 5,600 |
| 합계 | 280 $\bar{X} = 40$ | 420 $\bar{Y} = 60$ | 14,000 | 26,550 | 18,700 |

$$\begin{aligned} r_{XY} &= \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} = \frac{7 * 18,700 - 280 * 420}{\sqrt{7 * 14,000 - 280^2} * \sqrt{7 * 26,550 - 420^2}} \\ &= \frac{13,300}{\sqrt{19,600} * \sqrt{9,450}} = 0.98 \end{aligned}$$

- 상관계수가 0.98이므로, **훈련시간**과 **생산성** 간의 **선형관계가 상당히 높다**고 볼 수 있다

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관분석의 종류

- 상관분석(correlation analysis)은 X값과 y값, 두 변수 간 상관관계를 탐색
- 즉 연속적인 두 변수 간의 선형관계를 탐색 및 확인하는 방법
- 상관분석에서의 구해지는 상관 계수 r 은 두 변수의 직선적인 연관성 분석
- X가 증가하면 y도 증가하는가? 감소하는가?의 정도를 나타낼 수 있는 방법

※ 두 변수간의 관련성만을 의미할 뿐 인과관계를 밝히진 못한다.

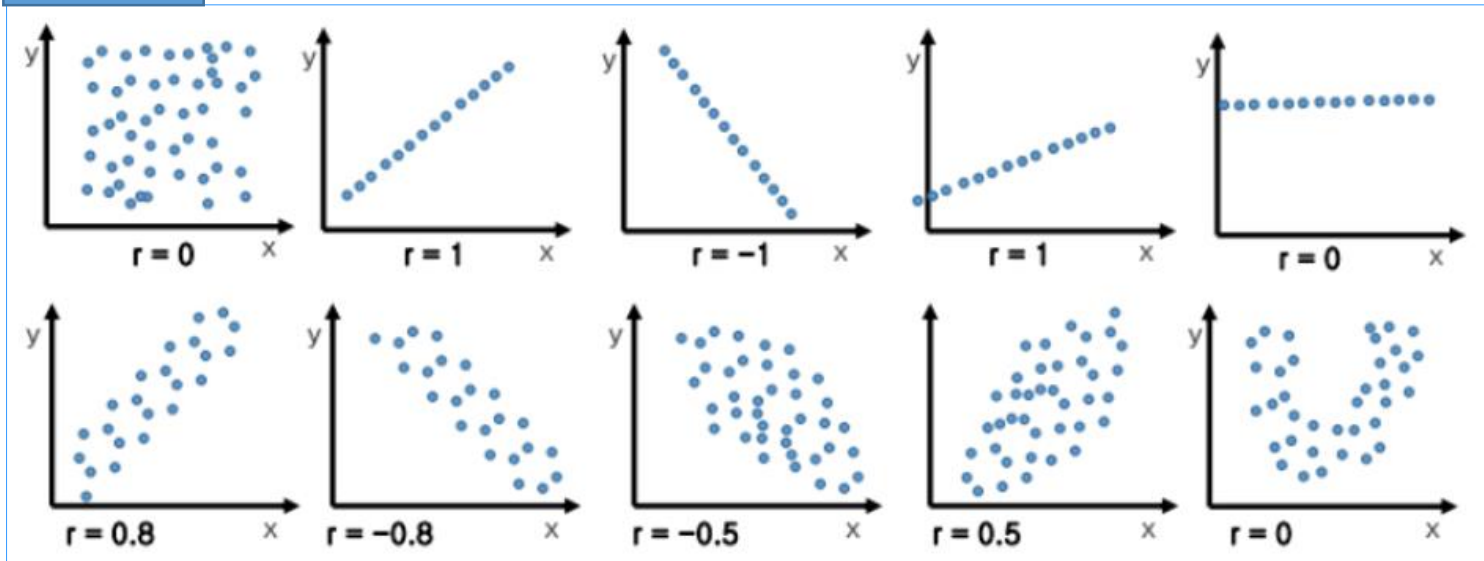
일반적으로 상관계수의 r 은 Pearson상관계수를 의미

- Pearson 상관계수 :
 - 모집단의 분포가 정규분포에 가까우면 사용
 - 두 변수가 양적자료인 경우 사용
- Spearman 상관계수 :
 - 모집단이 비정규분포를 나타낼 때 사용
 - 두 변수가 하나라도 서열척도인 경우 사용

Chapter 14 상관분석과 회귀분석의 기초

◎ 상관계수와 산포도의 모양

상관성



상관분석(correlation analysis) 공식

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \rightarrow r_{xy} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \rightarrow r_{xy} = \frac{S_{xy}}{S_x S_y}$$

◎ 상관분석실습

| 은행지점 | 홍보비용 | 예금유치액 |
|------|------|-------|
| 1 | 40 | 87 |
| 2 | 50 | 108 |
| 3 | 30 | 69 |
| 4 | 60 | 135 |
| 5 | 70 | 148 |
| 6 | 60 | 132 |
| 7 | 30 | 73 |
| 8 | 60 | 128 |
| 9 | 20 | 50 |
| 10 | 80 | 170 |

상관관계 분석적도 :

피어슨 상관계수(Pearson correlation coefficient : r)

| 피어슨 상관계수 R | 상관관계 정도 |
|------------------------|------------|
| ± 0.9 이상 | 매우 높은 상관관계 |
| $\pm 0.9 \sim \pm 0.7$ | 높은 상관관계 |
| $\pm 0.7 \sim \pm 0.4$ | 다소 높은 상관관계 |
| $\pm 0.4 \sim \pm 0.2$ | 낮은 상관관계 |
| ± 0.2 미만 | 상관관계 없음 |

※ 상관계수 r은 -1에서 +1까지의 값을 가진다. 또한 가장 높은 완전 상관관계의 상관계수는 1이고, 두 변수간에 전혀 상관관계가 없으면 상관계수는 0이다.

Chapter 14 상관분석과 회귀분석의 기초

◎ 결정계수

- 상관계수를 해석하는 또 다른 방법으로 상관계수의 제곱, 즉 결정계수 r^2 을 사용하기도 함
- 결정계수는 예측변수가 종속변수를 예측(또는 설명)할 수 있는 비율을 말해줌 $r=0.50$ 일 때

결정계수와 y분산 간의 관계

$$y\text{에 대한 }x\text{의 설명 부분} = r_{XY}^2 S_Y^2$$

$$y\text{에 대한 }x\text{의 설명되지 않는 부분} = (1 - r_{XY}^2) S_Y^2$$

- 만일 두 변수를 표준점수로 변환시키면 $S_Y^2=1$ 이 되므로 아래와 같이 표현이 가능

$$y\text{에 대한 }x\text{의 설명 부분} = r_{XY}^2$$

$$y\text{에 대한 }x\text{의 설명되지 않는 부분} = (1 - r_{XY}^2)$$

Chapter 14 상관분석과 회귀분석의 기초

◎ 단순회귀분석

회귀분석의 개념

- 앞 절에서 살펴본 상관분석은 단순히 두 변수 간 선형관련성을 측정하는 데 의의
- 하지만, 한 변수가 다른 변수에 미치는 영향을 알아보고자 하는 경우가 존재 ::
제품의 가격에 따라 수요의 변동, 금리가 상승하면 주가가 하락
- 영향을 받는 변수를 종속변수(dependent var)라 하고, 다른 변수에 영향을 주는 변수를 독립변수(independent var)
- **단순**회귀분석(simple regression analysis):: **한 독립변수**와 한 종속변수의 관계
- **다중**회귀분석(multiple regression analysis):: **여러 독립변수**와 한 종속변수의 관계

R 회귀분석-단순선형회귀분석

상관관계분석 (Correlation Analysis)

두 변수 간의
선형 관계를
조사하는 것

단순회귀분석 (Simple regression analysis)

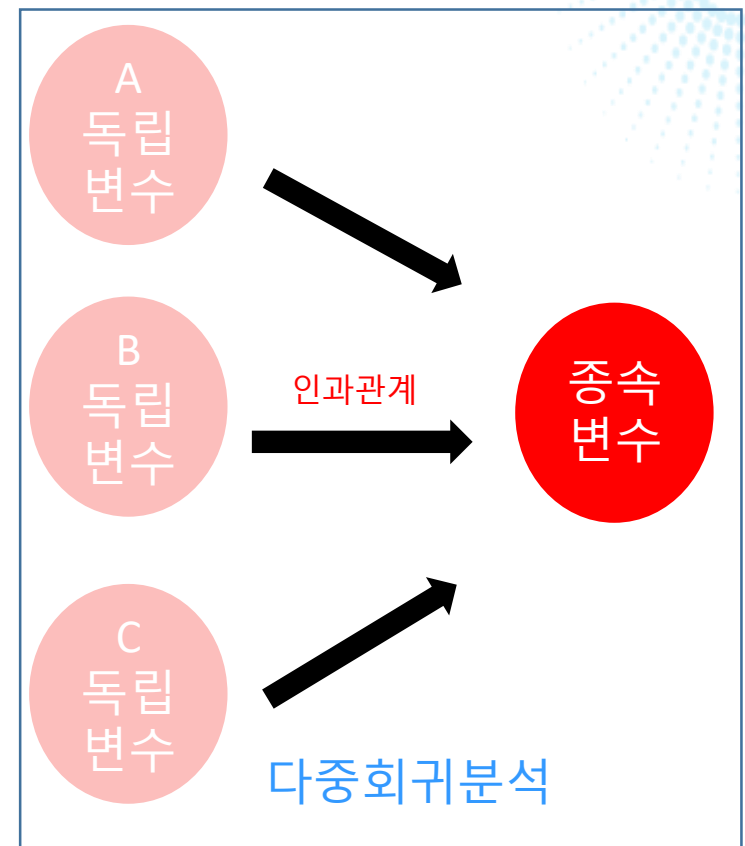
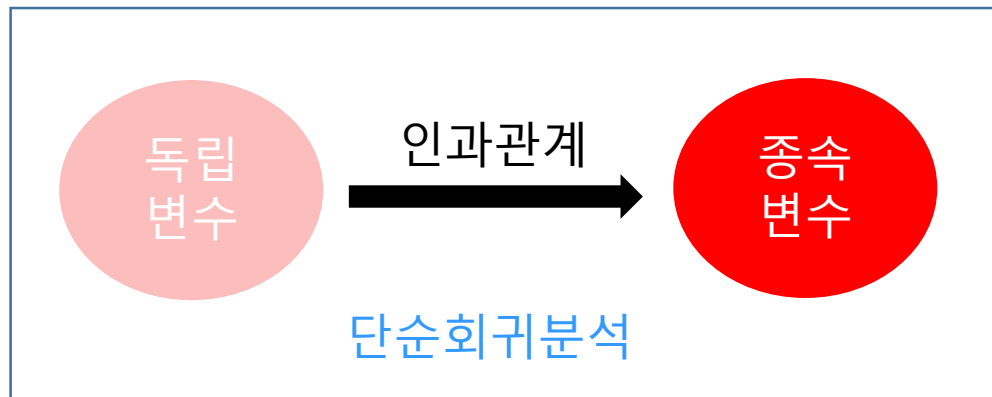
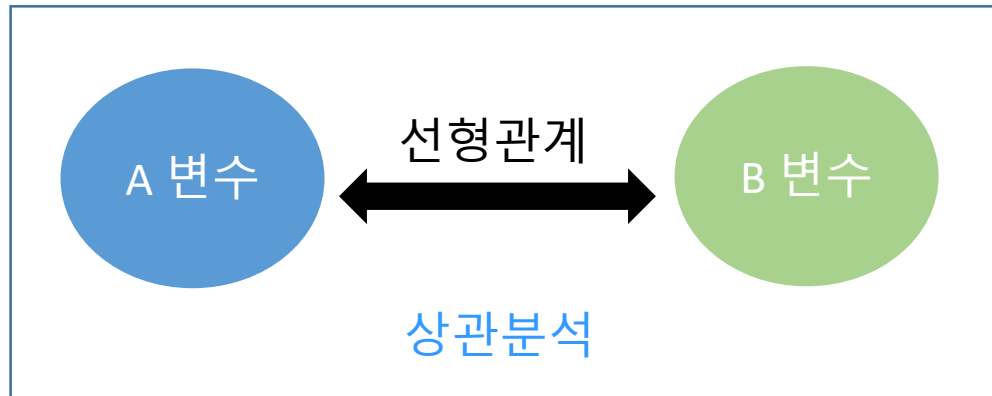
두 변수 간의
인과관계(Casua
l Relationship)를
조사하는 것

다중회귀분석 (Multiple regression)

두 개 이상의
독립 변수들과
하나의
종속 변수 간의
관계를
분석하는
기법으로
단순회귀분석
을 확장한 것

출처 : <https://blog.naver.com/PostList.naver?blogId=y4769&from=postList&categoryNo=39>

R 회귀분석-단순선형회귀분석



출처 : <https://blog.naver.com/PostList.naver?blogId=y4769&from=postList&categoryNo=39>

R 회귀분석을 하는 이유

장점 :

- 가장 기본적인 분석으로 모형 결과가 선형 1차 방정식으로 매우 단순
- 의외로 1차 방정식이 잘 적용됨
- 입력 변수의 회귀 계수를 이용해 각 변수들의 영향력을 쉽게 파악 가능
- 적은 데이터로도 모델링이 가능하며, 빅데이터에도 모델링이 적용 가능
- 넓은 범용성에 비해 시간 절약이 가능

단점 :

- 하지만 선형적이지 않는 데이터에 적용이 힘들
- 충족시켜야 하는 가정이 매우 많음
- 가정을 만족시키기 위한 변환과정이 힘들

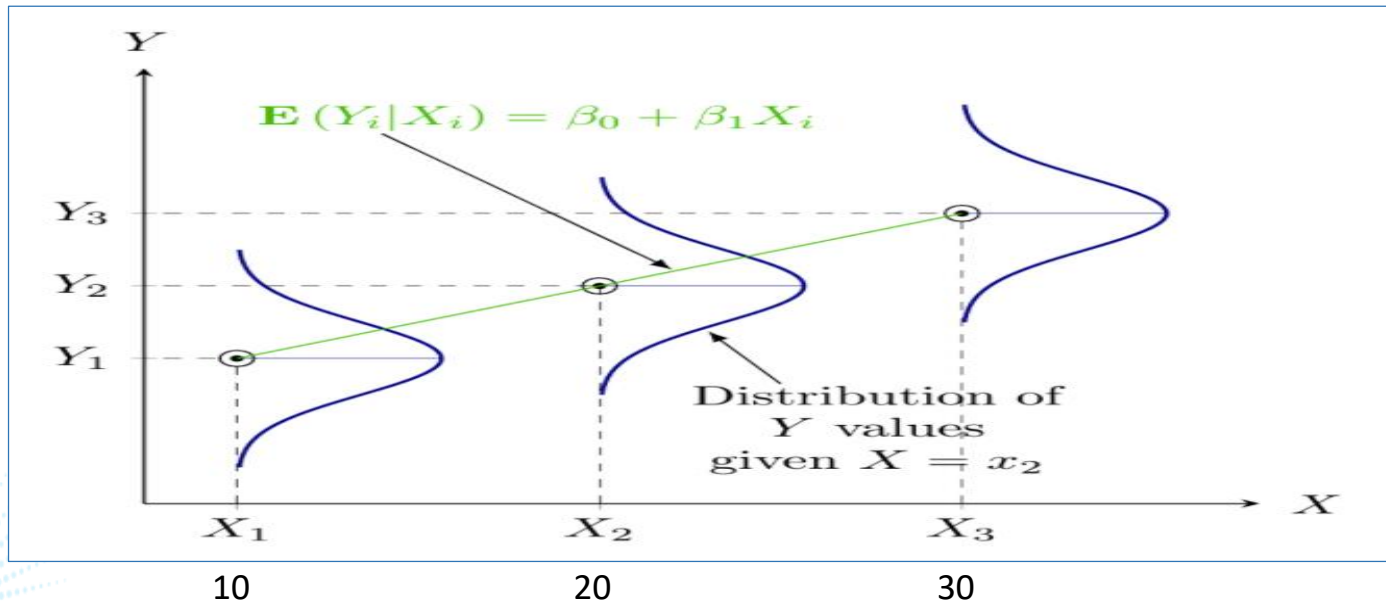
기본가정

- 선형성 :: X (독립변수)와 y (종속변수)의 관계가 선형적 관계여야 한다
- 정규성 :: Q-Q plot이 직선으로 표현되어 잔차가 정규분포를 따른다는 것이다
- 독립성 :: 오차항은 서로 독립적이어야 한다.
더빈 왓슨(Durbin Watson) 검정을 수행한다. 검정 결과 더빈 왓슨 통계량이 2에 가까울수록 오차항의 자기상관이 없음을 의미.
더빈왓슨 통계량이 2에 가까울수록 오차항이 독립적(== 자기상관성이 없다)
- 등분산성 :: X (독립변수)에 대한 잔차의 산점도를 그렸을 때 잔차들이 일정해야 함
- 비상관성 :: 관측치들의 잔차들끼리 상관성이 없어야 한다

Chapter 14 상관분석과 회귀분석의 기초

◎ 단순회귀모형과 단순회귀식

- 독립변수와 종속변수 간의 1차함수관계 또는 선형관계를 가정할 때, 회귀모형은 모집단의 경우 1차식의 확정적 함수관계를 나타내는 부분($\alpha + \beta X_i$)과 확률적 오차항(ε_i)을 결합한 형태로 나타낼 수 있음
- 오차항에 대한 구체적 가정은 다시금 말미에 설명하기로 하고, 여기서는 간단히 오차항의 평균이 0이고 오차항의 분산이 일정한 정규분포라 가정



Chapter 14 상관분석과 회귀분석의 기초

◎ 단순회귀모형과 단순회귀식

- 일례로 훈련시간을 10시간으로 했을 때 생산성은 30, 40, 50 등 어떠한 값이라도 될 수 있는데, 그 값들이 정규분포를 따른다는 것.
- 만약 생산성의 평균값이 40이라면, 훈련시간이 10시간일 경우 생산성은 평균 40을 중심으로 하는 정규분포를 따르게 되고, 이때 평균값은 훈련시간이 10이라는 조건하에서의 평균이므로 μ_{y*10} 로 표시

모집단의 경우

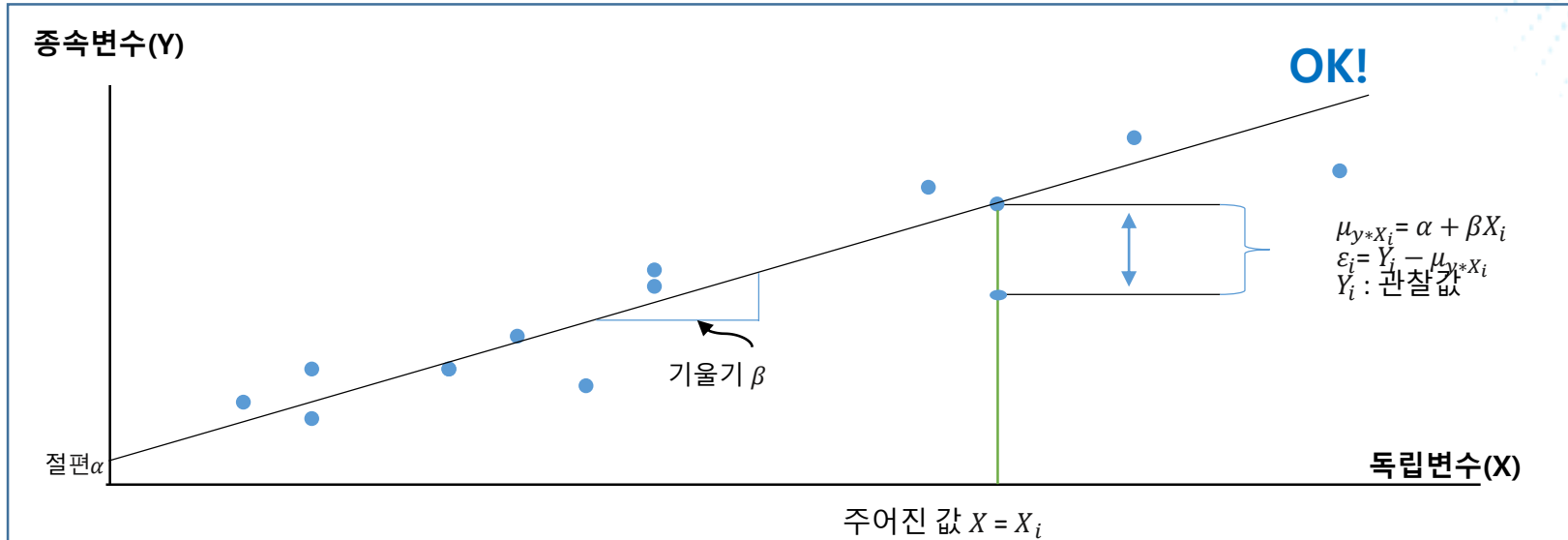
단순회귀모형 $Y_i = \alpha + \beta X_i + \varepsilon_i$

단순회귀식 $\mu_{y*X_i} = \alpha + \beta X_i$

- α 와 β 를 회귀계수(regression coefficient)라 함. 회귀식은 절편을 α 로 하고 기울기가 β 인 직선이 됨을 알 수 있음

Chapter 14 상관분석과 회귀분석의 기초

◎ 단순회귀모형과 단순회귀식



- 모집단의 회귀식을 구하는 것은 실제로 불가능한 경우가 대부분이다. 우리는 표본으로부터 회귀식을 구하여 모수를 추정하여야 함
- 표본의 절편은 a , 기울기는 b 라 하고, 오차는 잔차(residual)라고 부르며 e_i 로 나타냄

표본의 경우

단순회귀모형 $Y_i = a + bX_i + e_i$

단순회귀식 $\hat{Y}_i = a + bX_i$

Chapter 14 상관분석과 회귀분석의 기초

◎ 단순회귀모형과 단순회귀식

- 표본의 회귀식에서 종속변수 \hat{Y}_i 은 회귀식을 통해 구해지는 수치 따라서 실제의 측정값 Y_i 와 모집단의 회귀식에 의해 구해진 값 \hat{Y}_i 은 다름.
- 독립변수 X 에 대응하는 측정된 값 Y 와 예측된 값 \hat{Y}_i 의 차를 예측오차 또는 잔차(residual)라 하며 다음 식과 같음

$$e_i = Y_i - \hat{Y}_i$$

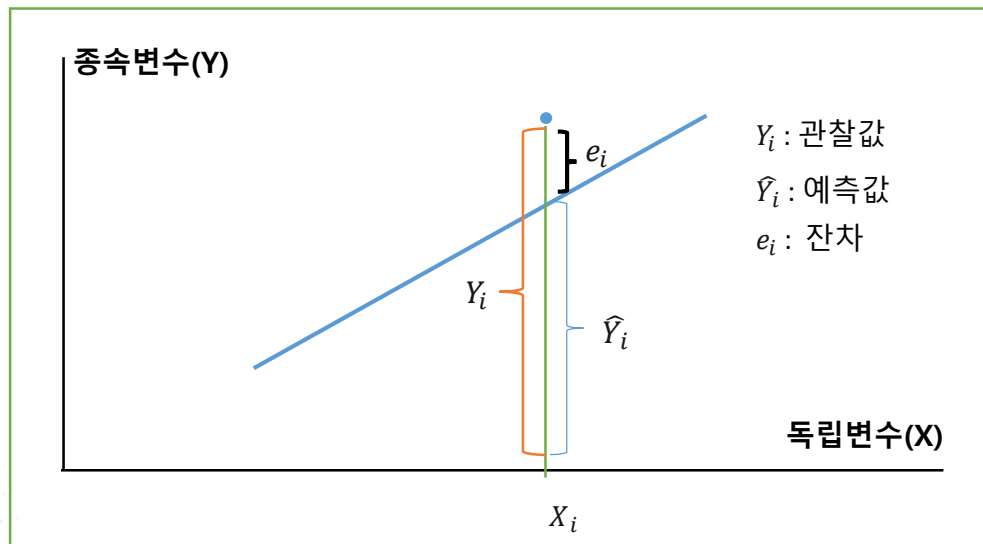
◎ 오차항에 대한 가정

- 가정1 오차항의 평균값은 0이다.
- 가정2 두 개의 오차항은 서로 독립적이다.
- 가정3 오차항의 분산은 일정하다.
- 가정4 독립변수는 확률변수가 아니다.
- 가정5 오차항은 정규분포를 따른다.

Chapter 14 상관분석과 회귀분석의 기초

◎ 최소제곱법

- 회귀분석이란 두 변수 간 관련성을 선형함수로 규정하고 이에 필요한 모수인 α 와 β 를 표본자료를 통해 규명하려는 것임을 살펴봄.
- 그렇다면 두 변수 간 선형관계를 가장 잘 나타내는 회귀선은 어떤 것일까?
- 가장 좋은 표본회귀식은 전체적으로 예측오차, 즉 잔차를 가장 작게 해주는 모형이 좋은 것
- 잔차(residual)란 실제로 관찰된 값 Y_i 과 표본회귀식으로부터 얻은 예측값의 차 \hat{Y}_i



- 잔차를 작게 해주는 회귀식을 구하는 방법은 여러 가지가 있을 수 있으나, 회귀분석에서는 잔차의 제곱합이 최소가 되도록 모수를 추정하는 방법을 주로 사용

Chapter 14 상관분석과 회귀분석의 기초

◎ 최소제곱법

$$\min \sum e_i^2 = \min \sum (Y_i - \hat{Y}_i)^2$$

- 위의 방법은 최소제곱법(method of least squares)이라 하며, 회귀식을 결정하는 가장 좋은 방법으로 받아들여지고 있음
- 이 방법에 의하면 다른 방법에 의해 구한 회귀식보다 통계학적으로 그 성질이 우수한 α, β 의 추정값을 얻을 수 있음
- 통계학적으로 우수하다는 것은 표본에서 통계값을 구할 때 그 통계값이 모집단의 모수를 가장 잘 설명하는 추정값이라는 것을 의미

◎ 최소제곱법에 의한 추정량

- 최소제곱법에 의한 회귀식 도출

$$\hat{Y}_i = a + bX_i$$

$$\min \sum e_i^2 = \min \sum (Y_i - a - bX_i)^2$$

Chapter 14 상관분석과 회귀분석의 기초

◎ 최소제곱법에 의한 추정량

- 최소제곱법에 의한 회귀식의 결정이란 만족시키는 a 와 b 를 구하는 것
- $\sum(Y_i - a - bX_i)^2$ 을 최소로 하는 a 와 b 를 구하기 위해 a 와 b 에 대해 각각 편미분하면 다음과 같은 두 식을 얻을 수 있음

$$\sum Y_i = na + b \sum X_i$$

$$\sum X_i Y_i = a \sum X_i + b \sum X_i^2$$

- 두 연립방정식을 회귀분석의 정규방정식(normal equation)이라 부름 이 두 식을 충족시키는 표본의 회귀계수 a 와 b 를 구하기 위하여 연립방정식을 풀면 다음과 같음

표본의 회귀계수

$$b = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sum X_i^2 - n \bar{X}^2}$$

$$a = \bar{Y} - b\bar{X}$$

- a 와 b 를 회귀모수 α 와 β 에 대한 통상최소제곱추정량(ordinary least square estimator) 혹은 최소제곱추정량이라 함

Chapter 14 상관분석과 회귀분석의 기초

◎ 최소제곱추정량의 특성

- 최소제곱추정량이 통계적으로 우수한 이유:

1) 최소제곱법에 의해 도출된 추정량은 불편성을 갖게 됨 즉, 추정량 a 와 b 의 평균값은 모수의 회귀계수인 α 와 β 값과 일치하게 됨

$$E(a) = \alpha$$

$$E(b) = \beta$$

- 2) 최소제곱추정량 a 와 b 는 종속변수 Y 의 1차함수, 즉 직선의 관계를 지님
- 3) 최소제곱추정량은 모든 가능한 선형불편추정량 중에서 최소의 분산값을 가짐
- 4) 즉, 최소제곱추정량 는 에 대한 선형불편추정량들 중 분산이 가장 작아 가장 효율적인 추정량 따라서 a 와 b 를 α 와 β 에 대한 최량선형불편추정량(best linear unbiased estimator)이라고도 함

Chapter 14 상관분석과 회귀분석의 기초

◎ 최소제곱추정량의 특성

- 최소제곱추정량이 통계적으로 우수한 이유:

1) 최소제곱법에 의해 도출된 추정량은 불편성을 갖게 됨 즉, 추정량 a 와 b 의 평균값은 모수의 회귀계수인 α 와 β 값과 일치하게 됨

$$E(a) = \alpha$$

$$E(b) = \beta$$

- 2) 최소제곱추정량 a 와 b 는 종속변수 Y 의 1차함수, 즉 직선의 관계를 지님
- 3) 최소제곱추정량은 모든 가능한 선형불편추정량 중에서 최소의 분산값을 가짐
- 4) 즉, 최소제곱추정량 는 에 대한 선형불편추정량들 중 분산이 가장 작아 가장 효율적인 추정량 따라서 a 와 b 를 α 와 β 에 대한 최량선형불편추정량(best linear unbiased estimator)이라고도 함

Chapter 14 상관분석과 회귀분석의 기초

예제 14-1

10명의 기능공이 일하는 공장에서 그들에게 생산성을 높이는 훈련을 실시하였다. 훈련시간의 정도에 따른

생산성을 예측하기 위해서 회귀식을 구하고자 한다. 생산성은 하루에 만들 수 있는 제품의 수로 측정하였

| 기능공 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------|----|----|----|----|----|----|----|----|----|----|
| 훈련시간 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| 생산성 | 10 | 14 | 17 | 19 | 21 | 22 | 24 | 26 | 24 | 23 |

| 대상 | X_i | Y_i | X_i^2 | Y_i^2 | $X_i * Y_i$ | \hat{Y}_i | $Y_i - \hat{Y}_i$ | $(Y_i - \hat{Y}_i)^2$ |
|-----|-------|-------|---------|---------|-------------|-------------|-------------------|-----------------------|
| 1 | 2 | 10 | 4 | 100 | 20 | 13.25 | -3.25 | 10.5625 |
| 2 | 4 | 14 | 16 | 196 | 56 | 14.75 | -0.75 | 0.5625 |
| 3 | 6 | 17 | 36 | 289 | 102 | 16.25 | 0.75 | 0.5625 |
| 4 | 8 | 19 | 64 | 361 | 152 | 17.75 | 1.25 | 1.5625 |
| 5 | 10 | 21 | 100 | 441 | 210 | 19.25 | 1.75 | 3.0625 |
| 6 | 12 | 22 | 144 | 484 | 264 | 20.75 | 1.25 | 1.5625 |
| 7 | 14 | 24 | 196 | 576 | 336 | 22.25 | 1.75 | 3.0625 |
| 8 | 16 | 26 | 256 | 676 | 416 | 23.75 | 2.25 | 5.0625 |
| 9 | 18 | 24 | 324 | 576 | 432 | 25.25 | -1.25 | 1.5625 |
| 10 | 20 | 23 | 400 | 529 | 460 | 26.75 | -3.75 | 14.0625 |
| 합 계 | 110 | 200 | 1,540 | 4,228 | 2,448 | | | 41.625 |

Chapter 14 상관분석과 회귀분석의 기초

$$b = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{10 * 2,448 - 110 * 200}{10 * 1,540 - 110^2} = 0.75$$

$$a = \bar{Y} - b\bar{X} = 20 - 0.75 * 11 = 11.75$$

- 위에서 계산된 a와 b를 대입하면 다음과 같은 표본회귀식을 얻을 수 있음

$$\hat{Y}_i = 11.75 + 0.75 * X_i$$

- 이 회귀식에서 기울기가 0.75라는 것은 1시간의 훈련을 받으면 생산성이 하루에 0.75개만큼 늘어난다는 것만일 어느 기능공이 15시간의 훈련을 받는다면 그 사람의 예상 생산량이 23개일 것이라는 것을 아래와 같이 예측할 수 있다

$$\hat{Y}_i = 11.75 + 0.75 * 15 = 23$$

- 즉 15시간 훈련을 받은 사람은 하루에 23개의 제품을 만들 것이라고 예측한다. 그러나 15시간의 훈련을 받은 사람이 반드시 23개를 만든다는 것이 아니라, 대체로 23개를 중심으로 흩어진 분포를 이루고 있다는 것임

Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석

- 반응변수 Y 와 설명변수 X_1, X_2, \dots, X_k 사이에 선형 관계 가정

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, \dots, n$$

- 오차항 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 가정: $N(0, \sigma^2)$, 서로 독립
- 선형 및 오차항 가정
 - 회귀모형의 추정 및 추론의 정당성 보장
 - 가정 위반 시 추론 결과에 대한 신뢰성 저하

Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석

```
> cor(states)
```

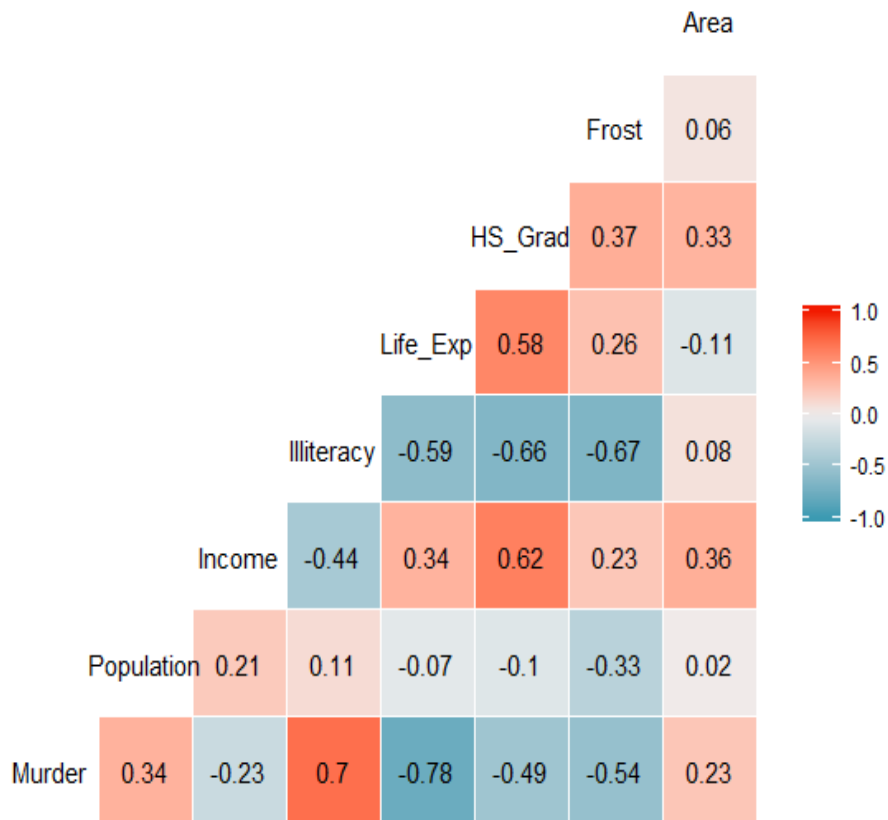
| | Population | Income | Illiteracy | Life_Exp |
|------------|-------------|------------|-------------|-------------|
| Population | 1.00000000 | 0.2082276 | 0.10762237 | -0.06805195 |
| Income | 0.20822756 | 1.0000000 | -0.43707519 | 0.34025534 |
| Illiteracy | 0.10762237 | -0.4370752 | 1.0000000 | -0.58847793 |
| Life Exp | -0.06805195 | 0.3402553 | -0.58847793 | 1.0000000 |
| Murder | 0.34364275 | -0.2300776 | 0.70297520 | -0.78084575 |
| HS Grad | -0.09848975 | 0.6199323 | -0.65718861 | 0.58221620 |
| Frost | -0.33215245 | 0.2262822 | -0.67194697 | 0.26206801 |
| Area | 0.02254384 | 0.3633154 | 0.07726113 | -0.10733194 |

| | Murder | HS_Grad | Frost | Area |
|------------|------------|-------------|------------|-------------|
| Population | 0.3436428 | -0.09848975 | -0.3321525 | 0.02254384 |
| Income | -0.2300776 | 0.61993232 | 0.2262822 | 0.36331544 |
| Illiteracy | 0.7029752 | -0.65718861 | -0.6719470 | 0.07726113 |
| Life Exp | -0.7808458 | 0.58221620 | 0.2620680 | -0.10733194 |
| Murder | 1.0000000 | -0.48797102 | -0.5388834 | 0.22839021 |
| HS Grad | -0.4879710 | 1.0000000 | 0.3667797 | 0.33354187 |
| Frost | -0.5388834 | 0.36677970 | 1.0000000 | 0.05922910 |
| Area | 0.2283902 | 0.33354187 | 0.0592291 | 1.0000000 |

- 상관계수 행렬: 변수의 개수가 많아지면 변수 사이 관계 파악이 어려움

Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석



Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석

- 선형보다는 2차가 더 적합한 것으로 보임
- 다항회귀모형 (차수가 높아지는 분석)
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_p X_i^p + \varepsilon_i$$
- 차수 p 를 너무 높이면 다중공선성의 문제가 발생할 수 있음
- 3차를 넘지 않는 것이 일반적

Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석

- 회귀모형: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$
- 회귀계수에 대한 가설
 - 1) $H_0: \beta_1 = \cdots = \beta_k = 0$
 - 2) $H_0: \beta_q = \beta_{q+1} = \cdots = \beta_r = 0, \quad q < r \leq k$
 - 3) $H_0: \beta_i = 0, H_1: \beta_i \neq 0$
- 회귀계수의 신뢰구간
- 회귀모형 적합 정도에 대한 통계량
 - 결정계수 (R^2), 수정된 결정계수 $\text{adj}(R^2) ::$ 높을수록 좋은 지표
 - AIC, BIC :: 낮을수록 좋은 지표

Chapter 14 상관분석과 회귀분석의 기초

```
> fit <- lm(Murder ~ ., states)
> summary(fit)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|--------|
| -3.4452 | -1.1016 | -0.0598 | 1.1758 | 3.2355 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|------------|------------|---------|----------|-----|
| (Intercept) | 1.222e+02 | 1.789e+01 | 6.831 | 2.54e-08 | *** |
| Population | 1.880e-04 | 6.474e-05 | 2.905 | 0.00584 | ** |
| Income | -1.592e-04 | 5.725e-04 | -0.278 | 0.78232 | |
| Illiteracy | 1.373e+00 | 8.322e-01 | 1.650 | 0.10641 | |
| Life_Exp | -1.655e+00 | 2.562e-01 | -6.459 | 8.68e-08 | *** |
| Hs_Grad | 3.234e-02 | 5.725e-02 | 0.565 | 0.57519 | |
| Frost | -1.288e-02 | 7.392e-03 | -1.743 | 0.08867 | . |
| Area | 5.967e-06 | 3.801e-06 | 1.570 | 0.12391 | |
| --- | | | | | |

Residual standard error: 1.746 on 42 degrees of freedom

Multiple R-squared: 0.8083, Adjusted R-squared: 0.7763

F-statistic: 25.29 on 7 and 42 DF, p-value: 3.872e-13

- 개별 회귀계수 추정 및 검정
- \sqrt{MSE}
- 결정계수 및 수정된 결정계수
- 모든 회귀계수의 유의성 검정 :: t검정

Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석

● 회귀모형 적합 정도에 대한 통계량

- 결정계수(R^2): 반응변수의 변량 중 회귀모형으로 설명되는 변량의 비율
 - 모형에 포함된 설명변수의 개수가 증가하면 증가하는 특성이 있음
 - 설명변수의 개수가 같은 모형 비교에는 의미가 있는 통계량
- 수정 결정계수(adj. R^2): 추가된 설명변수가 모형 적합도에 도움이 되는 경우에만 증가
- AIC & BIC: 설명변수의 개수가 p 인 모형
 - $AIC = n \log \left(\frac{SSE}{n} \right) + 2p$
 - $BIC = n \log \left(\frac{SSE}{n} \right) + p \log(n)$
 - AIC, BIC가 작은 모형이 더 적합도가 높은 모형

Chapter 14 상관분석과 회귀분석의 기초

◎ 다중회귀분석

- 반응변수의 변동을 설명할 수 있는 많은 설명변수 중 '최적'의 변수를 선택하여 모형에 포함시키는 절차
- 검정에 의한 방법
 - 변수의 유의성 검정을 이용하여 단계적으로 모형 선택
 - 후진소거법, 전진선택법, 단계별 선택법
- 모형선택 기준에 의한 방법
 - 모형의 적합도 등을 측정하는 통계량을 기반으로 모형 선택
 - 결정계수, 수정결정계수, 잔차제곱합, C_p 통계량, AIC, BIC 등등
- 어떤 모형이 '최적' 모형인가?

Chapter 14 상관분석과 회귀분석의 기초

- 함수 `regsubsets()`으로 적합

```
> library(leaps)
> fits <- regsubsets(Murder ~ ., states)
```

- 설명변수가 k인 모형 중 결정계수가 가장 높은 모형

```
> summary(fits)
```

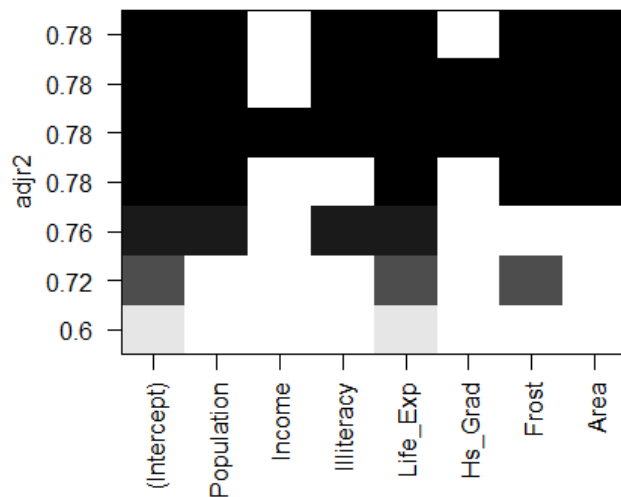
| | | Population | Income | Illiteracy | Life_Exp | Hs_Grad | Frost | Area |
|---|-------|------------|--------|------------|----------|---------|-------|------|
| 1 | (1) | " " | " " | " " | "*" | " " | " " | " " |
| 2 | (1) | " " | " " | " " | "*" | " " | "*" | " " |
| 3 | (1) | "*" | " " | "*" | "*" | " " | " " | " " |
| 4 | (1) | "*" | " " | " " | "*" | " " | "*" | "*" |
| 5 | (1) | "*" | " " | "*" | "*" | " " | "*" | "*" |
| 6 | (1) | "*" | " " | "*" | "*" | "*" | "*" | "*" |
| 7 | (1) | "*" | "*" | "*" | "*" | "*" | "*" | "*" |

Chapter 14 상관분석과 회귀분석의 기초

- 모든 가능한 회귀의 적합 결과 확인

1) 함수 plot()에 의한 확인

```
> plot(fits, scale="adjr2")
```



scale: 디폴트 "bic"

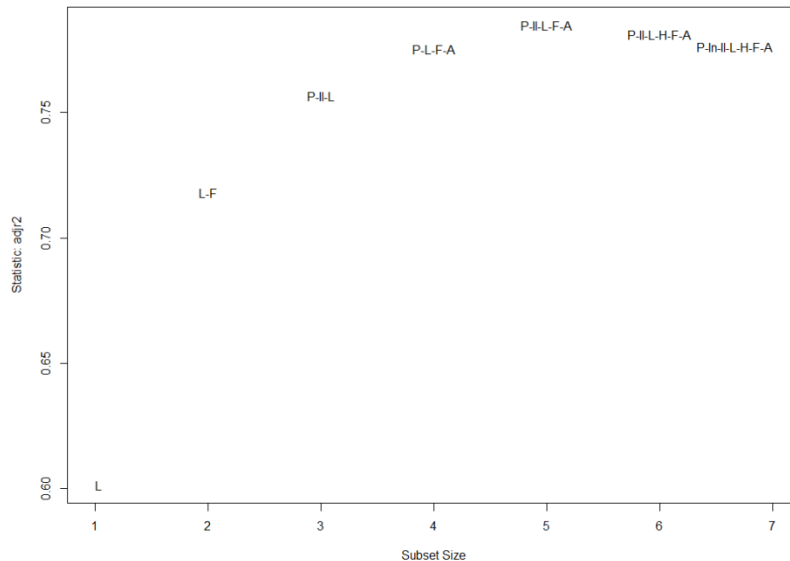
- 각 행: 하나의 모형을 의미
- 색이 채워진 직사각형: 모형에 포함된 변수
- Y축: 각 모형의 adj. R^2 값
- 위에서 첫 번째 모형: adj. R^2 의 값이 가장 큰 모형
- Population, Illiteracy, Life_Exp, Frost, Area 포함 모형 선택

Chapter 14 상관분석과 회귀분석의 기초

```
> subsets(fits, statistic="adjr2", legend=FALSE)
```

Abbreviation

| | |
|------------|----|
| Population | P |
| Income | In |
| Illiteracy | Il |
| Life_Exp | L |
| Hs_Grad | H |
| Frost | F |
| Area | A |



- P-Il-L-F-A 모형 선택

Chapter 14 상관분석과 회귀분석의 기초

Start: AIC=131.59
Murder ~ 1

| | Df | Sum of Sq | RSS | AIC |
|--------------|----|-----------|--------|---------|
| + Life_Exp | 1 | 407.14 | 260.61 | 86.550 |
| + Illiteracy | 1 | 329.98 | 337.76 | 99.516 |
| + Frost | 1 | 193.91 | 473.84 | 116.442 |
| + Hs_Grad | 1 | 159.00 | 508.75 | 119.996 |
| + Population | 1 | 78.85 | 588.89 | 127.311 |
| + Income | 1 | 35.35 | 632.40 | 130.875 |
| + Area | 1 | 34.83 | 632.91 | 130.916 |
| <none> | | | 667.75 | 131.594 |

Step: AIC=86.55
Murder ~ Life_Exp

| | Df | Sum of Sq | RSS | AIC |
|--------------|----|-----------|--------|---------|
| + Frost | 1 | 80.10 | 180.50 | 70.187 |
| + Illiteracy | 1 | 60.55 | 200.06 | 75.329 |
| + Population | 1 | 56.62 | 203.99 | 76.303 |
| + Area | 1 | 14.12 | 246.49 | 85.764 |
| <none> | | | 260.61 | 86.550 |
| + Hs_Grad | 1 | 1.12 | 259.48 | 88.334 |
| + Income | 1 | 0.96 | 259.65 | 88.366 |
| - Life_Exp | 1 | 407.14 | 667.75 | 131.594 |

Step: AIC=70.19
Murder ~ Life_Exp + Frost

| | Df | Sum of Sq | RSS | AIC |
|--------------|----|-----------|--------|---------|
| + Population | 1 | 23.710 | 156.79 | 65.146 |
| + Area | 1 | 21.084 | 159.42 | 65.976 |
| <none> | | | 180.50 | 70.187 |
| + Illiteracy | 1 | 6.066 | 174.44 | 70.477 |
| + Income | 1 | 5.560 | 174.94 | 70.622 |
| + Hs_Grad | 1 | 2.068 | 178.44 | 71.610 |
| - Frost | 1 | 80.104 | 260.61 | 86.550 |
| - Life_Exp | 1 | 293.331 | 473.84 | 116.442 |

Step: AIC=65.15
Murder ~ Life_Exp + Frost + Population

| | Df | Sum of Sq | RSS | AIC |
|--------------|----|-----------|--------|---------|
| + Area | 1 | 19.040 | 137.75 | 60.672 |
| + Illiteracy | 1 | 11.826 | 144.97 | 63.225 |
| <none> | | | 156.79 | 65.146 |
| + Hs_Grad | 1 | 1.821 | 154.97 | 66.561 |
| + Income | 1 | 0.739 | 156.06 | 66.909 |
| - Population | 1 | 23.710 | 180.50 | 70.187 |
| - Frost | 1 | 47.198 | 203.99 | 76.303 |
| - Life_Exp | 1 | 296.694 | 453.49 | 116.247 |

Step: AIC=60.67
Murder ~ Life_Exp + Frost + Population + Area

| | Df | Sum of Sq | RSS | AIC |
|--------------|----|-----------|--------|---------|
| + Illiteracy | 1 | 8.723 | 129.03 | 59.402 |
| <none> | | | 137.75 | 60.672 |
| + Income | 1 | 1.241 | 136.51 | 62.220 |
| + Hs_Grad | 1 | 0.771 | 136.98 | 62.392 |
| - Area | 1 | 19.040 | 156.79 | 65.146 |
| - Population | 1 | 21.666 | 159.42 | 65.976 |
| - Frost | 1 | 52.970 | 190.72 | 74.940 |
| - Life_Exp | 1 | 272.927 | 410.68 | 113.290 |

Step: AIC=59.4
Murder ~ Life_Exp + Frost + Population + Area + Illiteracy

| | Df | Sum of Sq | RSS | AIC |
|--------------|----|-----------|--------|--------|
| <none> | | | 129.03 | 59.402 |
| - Illiteracy | 1 | 8.723 | 137.75 | 60.672 |
| + Hs_Grad | 1 | 0.763 | 128.27 | 61.105 |
| + Income | 1 | 0.026 | 129.01 | 61.392 |
| - Frost | 1 | 11.030 | 140.06 | 61.503 |
| - Area | 1 | 15.937 | 144.97 | 63.225 |
| - Population | 1 | 26.415 | 155.45 | 66.714 |
| - Life_Exp | 1 | 140.391 | 269.42 | 94.213 |

Chapter 14 상관분석과 회귀분석의 기초

- 후진소거법에 의한 단계별 선택

```
> stepAIC(fit_full, trace=FALSE)
```

```
Call:
```

```
lm(formula = Murder ~ Population + Illiteracy + Life_Exp + Frost +  
    Area, data = states)
```

```
Coefficients:
```

| | | | | |
|-------------|------------|------------|------------|------------|
| (Intercept) | Population | Illiteracy | Life_Exp | Frost |
| 1.202e+02 | 1.780e-04 | 1.173e+00 | -1.608e+00 | -1.373e-02 |
| Area | | | | |
| 6.804e-06 | | | | |

Chapter 14 상관분석과 회귀분석의 기초

- BIC에 의한 단계별 선택

```
> stepAIC(fit_full, k=log(nrow(states)), trace=FALSE)
```

Call:

```
lm(formula = Murder ~ Population + Life_Exp + Frost + Area, data  
= states)
```

| Coefficients: | (Intercep t) | Population | Life_Exp | Frost | Area |
|---------------|-----------------|------------|------------|------------|-----------|
| | 1.387e+02 | 1.581e-04 | -1.837e+00 | -2.204e-02 | 7.387e-06 |

AIC에 의한 단계별 선택과는 다른 결과

End of This Chapter!



주요 이력

現) (주)W사 Recommendation System, Time Series etc) ing
前) (주)RTMC The Head of Strategic Planning Department
前) (주)biz사 Web Analysis & Data Voucher Business
前) H금속 FX, Amortization & Depreciation Planning
前) B건설 Mgmt Performance Report & Footnote
前) K문고 CRM VIP Clustering Strategy
前) L백화점 CRM Alert Strategy

학력

BSL(Business School of Lausanne) Big Data MBA
ASSIST Big Data statistic MBA

現) 대학교 및 관계기관 Big-Data 다수 강의
現) 코리아IT아카데미 Big-Data R
現) 코리아IT아카데미 Big-Data Python
現) 코리아IT아카데미 Big-Data Principles of Statistics