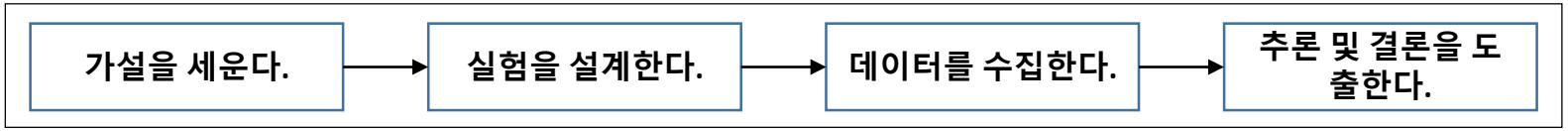
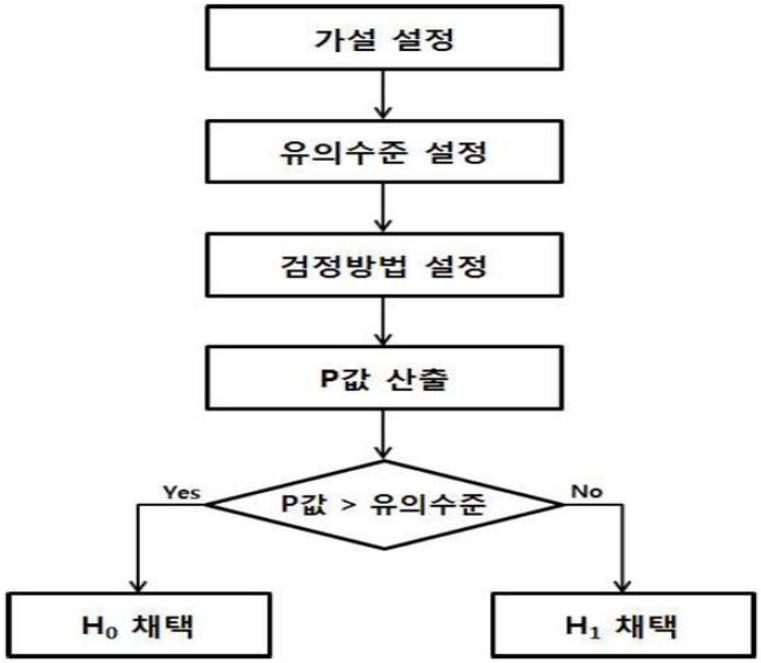
가설검증



전통적인 통계 추론 과정

- **가설 설정 및 검정의 단계**



가설검정의 단계

가설검증

- (예제) 1960년부터 1980년 사이에 미국의 남부에서 논란이 된 소송으로, 배심원 선정의 인종 편견이 문제가 되었다(일부 피고들이 배심원들의 평결에 이의를 제기했고, 소송을 제기).

1. 배심원 자격이 있는 시민들 중 50%가 흑인이었다.
2. 배심원 명부에 등재된 80명 중 흑인은 4명뿐이었다.

- 배심원 명부의 선정이 무작위였다고 가정하면,
- 80명으로 구성된 배심원명부상의 흑인수는 n=80, p=0.5인 이항확률변수 X가 되고,
- 4명의 흑인만 배심원 대상자로 선정될 확률은 $Pr(X \leq 4)=?$

- 위의 배심원의 문제는 연역적 확률 추론이다.

가설검증

- 배심원 명부의 선정이 무작위였다고 가정하면,
- 80명으로 구성된 배심원 명부상의 흑인 수는 n=80, p=0.5인 이항확률변수 X가 되고,
- 4명의 흑인만 배심원 대상자로 선정될 확률은 Pr(X≤4)= 0.00000000000000001308252

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\Pr(X \leq 4) = \binom{80}{4} 0.5^4 (1-0.5)^{80-4} = 1.308 \times 10^{-18}$$

```
choose(80, 4) * (0.5)^4 * (0.5)^(80-4)
#[1] 1.308252e-18

(x <- 0.00000000000000001308252)
#[1] 1.308252e-18
```

가설검증

- 통계적인 가설 검증의 4단계 과정
  - 1단계. 모든 가설을 세운다.
    - H_0 : 귀무가설 또는 영가설.
      - ✓ 관측된 사실이 순전히 우연의 결과라는 가설이다.
      - ✓ 앞에 소개한 소송에서, 배심원이 전체 모집단에서 무작위로 선정된 것이 H_0이고, 흑인이 선정될 확률은 p=0.5 이다.
    - H_1 : 대립가설.
      - ✓ 어떤 요소가 우연과 결합되어 나타난 것이 관측된 사실이라는 가설이다.
      - ✓ H_1는 흑인들이 배심원으로 선정될 확률은 자격이 있는 사람 중 흑인 구성비보다 적다는 것이다(p < 0.5).

가설검증

- 통계적인 가설 검증의 4단계 과정

  - 2단계. 검증통계량.
    - 영가설에 반대되는 증거를 평가할 통계량을 정한다.
      - ✓ 앞에 소개한 소송에서, 검정통계량은 p=0.5, n=80인 이항확률변수 X 이다.

  - 3단계. p값.
    - 영가설이 사실이라면, 최소한 관측치만큼 극단적인 검증통계량이 관측될 확률은 얼마인가? 라는 질문에 대한 대답이다.
      - ✓ 앞에 소개한 소송에서,
      - ✓ p값은 $\Pr(X \leq 4 \mid p=0.5 \text{ AND } n=80) = 1.308 \times 10^{-18}$

가설검증

- **통계적인 가설 검증의 4단계 과정**

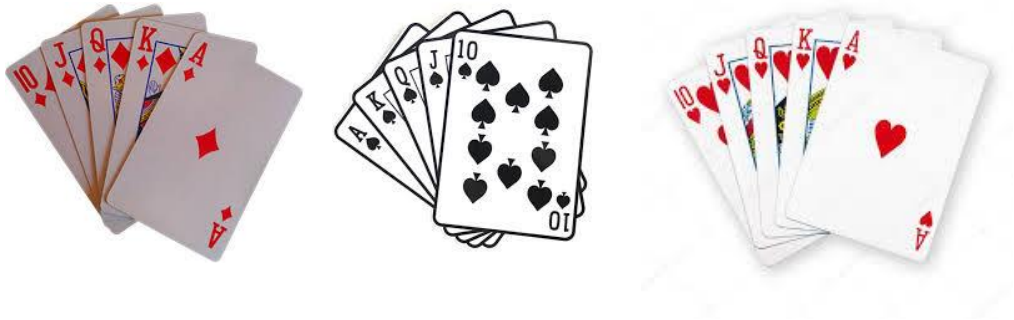
  - **4단계. p값과 정해진 유의수준 α를 비교한다.**

    - α는 어떤 결과가 통계적으로 의미 있다고 판단하는 기준이다.
    - 즉 p값 $\leq \alpha$ 이면, 영가설 H_0를 기각하고 다른 뭔가가 있다고 생각하는 것이다.

      ✓ 앞에 소개한 소송에서,

      ✓ 유의수준 α를 로열 플러시가 세 번 연달아 나올 확률인 3.6×10^{-18}로 취했다.

P값($=1.308 \times 10^{-18}$) < 로열 플러시가 세번 연달아 나올 확률($= 3.6 \times 10^{-18}$)



가설검증

가설검증 개요

- 분석하고자 하는 대상 모집단의 추출된 표본으로부터 분석목적에 적합한 과학적 추론을 위해서 가설검정을 시행한다.
- 가설검정이란 모집단에 대해 통계적 가설을 세우고 표본을 추출한 다음, 그 표본을 통해 얻은 정보를 이용하여 통계적 가설의 진위를 판단하는 과정이다.
- 즉 표본을 활용하여 모집단에 대입해보았을 때 새롭게 제기된 대립 가설이 옳다고 판단할 수 있는지를 평가하는 과정이다.
- 대부분 귀무가설이 참이라는 전제하에서 표본을 통하여 귀무가설이 옳지 않다는 것을 보임으로써 귀무가설을 기각시키고 대립가설을 채택한다.
- 귀무가설이 참이 아니라는 것을 표본을 통하여 통계적으로 증명하지 못한다면 귀무가설을 기각할 수 없다.
- 귀무가설이 참이라는 전제로 아주 예외적인 표본의 통계값이 나타날 확률이 일정수준(대개, 1%, 5% 혹은 10%) 이하일 경우에만 (즉 표본 조사를 통해 얻은 통계값이 매우 극단적일 경우) 귀무가설을 기각하고 대립가설을 채택한다.

가설검증

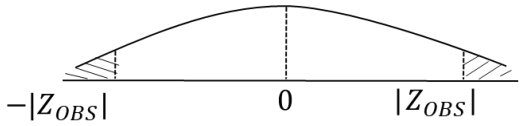
● 가설검증 방법

- 일반적으로 가설검정 방법은 대립가설의 형태에 따라서 양측검정과 단측검정이 있다.

✓ **양측검정**

모수 θ(혹은 모수들의 함수)에 대해 표본자료를 바탕으로 모수가 특정값 θ_0과 통계적으로 같은지 여부를 판단하기 위해 귀무가설을 $H_0: \theta = \theta_0$, 대립가설을 $H_1: \theta \neq \theta_0$와 같이 설정하는 경우를 양측검정이라 한다.

'양측' H_1 : p $\neq p_0$는 p값을 $\Pr(|Z| \geq |Z_{OBS}|)$로 쓴다.

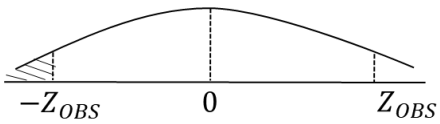


※ 수학기호 : θ Theta 세타

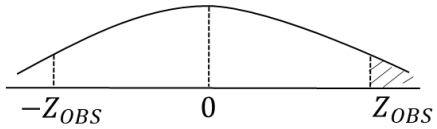
✓ **단측검정**

모수 θ에 대한 귀무가설이 $H_0: \theta = \theta_0$ 일 때, θ 가 특정 값 θ_0 보다 클 경우 (혹은 작은 경우)에만 귀무가설을 기각하게 되는 경우를 단측검정, 이 경우 대립가설 $H_1: \theta > \theta_0$ 혹은 $H_1: \theta < \theta_0$ 와 같다.

'왼쪽' H_1 : p < p_0는 p값을 $\Pr(Z < Z_{OBS})$로 쓴다.

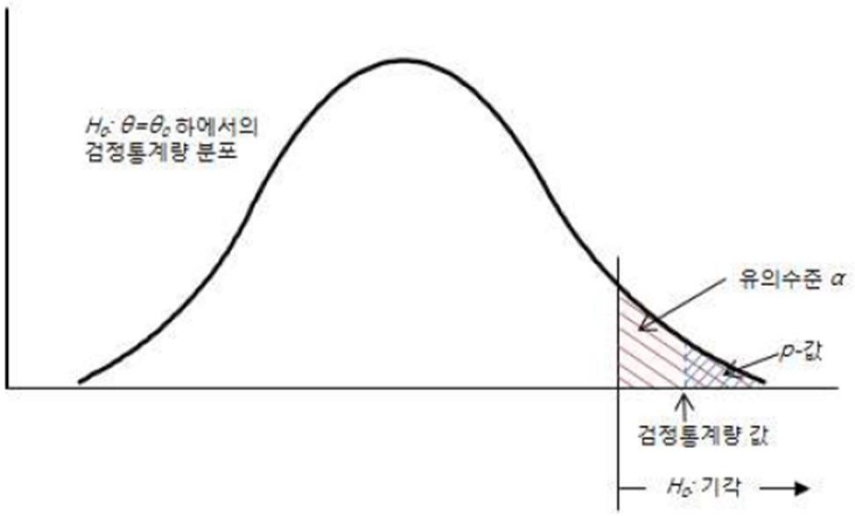


'오른쪽' H_1 : p > p_0는 p값을 $\Pr(Z > Z_{OBS})$로 쓴다.



가설검증

- **검정 통계량과 p값**
  - **검정 통계량**
    - 검정 통계량은 가설검정의 대상이 되는 모수를 추론하기 위해 사용되는 표본 통계량으로 귀무가설이 참이라는 전제하에서 모집단으로부터 추출된 확률표본의 정보를 이용하여 계산한다.
  - **P값**
    - 귀무가설이 참이라는 가정에 따라 주어진 표본 데이터를 희소 또는 극한값으로 얻을 확률값을 p-값이라고 한다.



검정 통계량과 p-값과의 관계

가설검증

- 귀무가설(null hypothesis)은 '영가설' 이러고도 불리는데 쉽게 말해 아무것도 일어나지 않는다는 의미이다.
  - ✓ 예를 들어 '직업에 따라 몸무게에 차이가 없을 것이다', '나이에 따라 키 차이가 없을 것이다'

- 대립가설(alternative hypothesis)은 무언가 일어났다는 의미이다.
  - ✓ 예를 들어 '직업에 따라 몸무게에 차이가 있을 것이다', '나이에 따라 키 차이가 있을 것이다'

가설검증

- 상관관계가 '있다', '없다'의 기준이 되고 오류를 허용할 범위를 설정하는 값을 유의수준(significance level)이라 한다.
  - ✓ 쉽게 말해 유의수준 0.1은 90% 믿을 수 있다는 뜻으로 생각하면 되고, 이를 신뢰수준(confidence level)이 90%라고 표현하기도 한다.
  - ✓ 유의수준이 작아질수록 신뢰수준은 높아진다. 즉 유의수준 0.05%는 95%의 신뢰수준을 의미이다.
  - ✓ 대부분은 습관적으로 0.05의 유의수준(95% 신뢰수준)을 사용하지만,
  - ✓ 데이터에 노이즈가 많다면 0.1의 유의수준(90% 신뢰수준)을 사용하기도 하고, 고위험 분야에서는 0.01의 유의수준(99% 신뢰수준)이 필요할 수도 있다.
- 통계분석 결과 중 가장 중요한 것은 오류가 나올 수 있는 확률인 유의확률(significance probability)이며, 이를 나타내는 수치가 p 값(p-value) 이다.
- 유의수준과 분석 결과 나온 p 값을 비교하여 가설을 평가하는 것이 바로 가설검정 (test of hypothesis)이다.

가설검증

- (예시) 평소에 늘 짜장면을 먹다가 언젠가부터 맛이 없어진 것 같아 짬뽕이 먹고 싶어졌다고 가정, 이러한 생각이 의미 있는지 가설을 검정해보려 귀무가설과 대립가설을 세우면 다음과 같다.

  - ❖ 귀무가설 : 짜장면과 짬뽕의 맛은 동일하다.
  - ❖ 대립가설 : 짬뽕이 짜장면보다 맛있다.

  - ✓ 유의수준은 0.05로 결정
  - ✓ 실제로 데이터로 수집해 컴퓨터로 계산한 결과 p값이 0.03이 나왔다면,
  - ✓ p값이 유의수준 0.05보다 작다.
  - ✓ 이 경우 귀무가설은 기각되고 대립가설이 채택된다(즉 짬뽕이 더 맛있으므로 짬뽕을 먹겠다는 새로운 주장이 통계적으로 맞게 됨).

큰 표본 비율에 대한 유의성 검증

- (예제) 1000명의 유권자 중에서 550명이 A의원에게 호의적이다. A의원은 $\hat{p}$ = 0.55(55% 사람들이 A의원에게 투표할 것이다)인걸 알고, p>0.5 (A의원이 당선될 것이다) 인지 알고 싶어한다. 통계적인 가설 검증의 4단계 과정을 통하여 유의성을 검증해 보자.

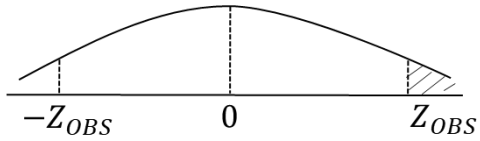
  - 1단계. 모든 가설을 세운다.
  - ✓ H_0 : 귀무가설 또는 영가설 : p = p_0
  - ✓ H_1 : 대립가설 : p > p_0

  - 2단계. 검증통계량.

$$Z_{OBS} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

  - 3단계. p값은 대립가설에 따라 다르다.
  - ✓ H_1 : 대립가설 : p > p_0

'오른쪽' H_1 : p > p_0는 p값을 $\Pr(Z > Z_{OBS})$로 쓴다.



이것은 p가 p_0에서 벗어난 정도를 측정하는 것이고, 영가설에서 Z_{OBS}는 표준정규분포를 갖는다.

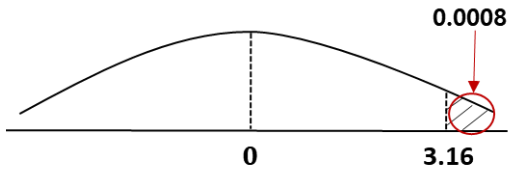
큰 표본 비율에 대한 유의성 검증

- 1. 가설은
  - ✓ $H_0 : p = p_0$
  - ✓ $H_1 : p > p_0$
- 2. 검정통계량은

$$Z_{OBS} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.55 - 0.5}{\sqrt{\frac{(0.5)(1-0.5)}{1000}}} = 3.16$$

- 3. p값은

$$Pr(Z > Z_{OBS}) = Pr(Z > 3.16) = 0.0008$$



- 4. 유의수준 α을 0.01로 택하고 살펴보니

$$Pr(Z > Z_{OBS}) = Pr(Z > 3.16) = 0.0008 < \alpha$$

  - ✓ 귀무가설을 기각하고 대립가설을 채택한다.
  - ✓ 즉 A의원은 당선될 것이다.

```
# 정규분포의 누적분포함수(cumulative function of normal distribution) 값 계산
# Pr(Z > 3.16) = 1 - F(3.16)
1 - pnorm(q=c(3.16), mean=0, sd=1, lower.tail = TRUE)
# [1] 0.0007888457
```

✓ lower.tail=TRUE 인 경우, probabilities is *P[X ≤ x]*, lower.tail=FALSE 인 경우, probabilities is *P[X > x]*.

큰 표본 비율에 대한 유의성 검증

- (예제) 시리얼을 생산하는 A회사에서 생산한 시리얼 상자의 평균 무게가 최소한 16온스(약 493.59g)라고 주장하고, B판매회사는 평균 무게가 그보다 조금이라도 작으면 반품하기로 한다. 그래서 시리얼 49상자를 무작위로 뽑아 무게를 달고 표본의 통계량을 구했다. 통계적인 가설 검증의 4단계 과정을 통하여 유의성을 검증해 보자.

  시리얼 49상자를 무작위로 뽑아 무게를 달고 표본의 통계량을 구한 결과,
  n=49, $\bar{x}$=15.90, s=0.35

  - 1단계 : 모든 가설을 세운다.

  ✓ 귀무 가설 H_0 : 시리얼 상자의 평균 무게 μ = 16 온스

  ✓ 대립 가설 H_1 : 시리얼 상자의 평균 무게 μ < 16 온스

  ※ 1oz(온스)=28.3495g=0.02835kg

  - 2단계 : 검증통계량.

$$Z_{OBS} = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

  ✓ s는 표본의 표준편차, 영가설에서 표본이 큰 경우 표준정규분포에 근사한다.

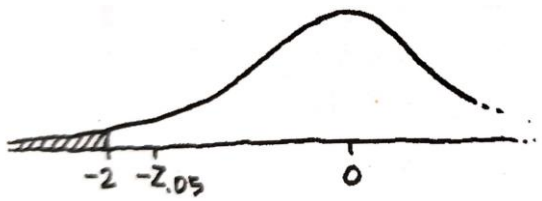
큰 표본 비율에 대한 유의성 검증

- 3단계 : p값.

  ✓ 시리얼 49상자를 무작위로 뽑아 무게를 달고 표본의 통계량을 구한 결과는 다음과 같다.

  n=49, $\bar{x}$=15.90, s=0.35

$$Z_{OBS} = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{15.90 - 16}{\frac{0.35}{\sqrt{49}}} = -2$$



$$p\text{값} = Pr(Z < Z_{OBS} \mid H_0) = Pr(Z < -2 \mid H_0) = 0.0227$$

```
# 정규분포의 누적분포함수(cumulative function of normal distribution) 값 계산
# Pr(Z < -2) = F(-2)
pnorm(q=c(-2), mean=0, sd=1)
#[1] 0.02275013
```

큰 표본 비율에 대한 유의성 검증

- 4단계 : p값과 정해진 유의수준 α를 비교한다.

  ✓ α는 어떤 결과가 통계적으로 의미 있다고 판단하는 기준점 즉 p값 $\leq \alpha$ 이면, H_0를 기각하고 H_1를 채택한다.

  ✓ 여기서 유의수준 α는 p값 ≤ 0.05 유의한 결과로 판다.

  ✓ p값 $=0.0227 < 0.05$ 이므로 H_0를 기각하고 H_1를 채택한다.

  ✓ 즉 B판매회사 시리얼 상자의 평균 무게가 16온스(약 493.59g) 보다 작아 제품을 반품한다.

가설검증

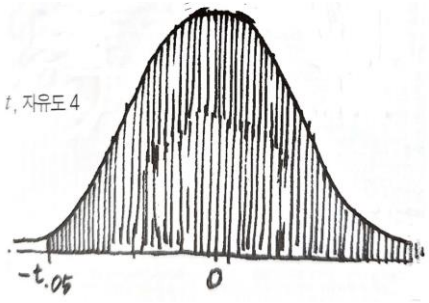
- (예제) A자동차에서 시속 10마일에서 정면충돌을 했을 경우, 평균 수리비용을 알아보기 위해 5대를 충돌실험 한 결과 각각 150달러, 400달러, 720달러, 500달러, 930달러의 수리비용이 나왔다. 평균 수리비용이 $1,000 이내여야 보험을 들어준다고 가정하자.
  유의수준 α=0.05, n=5, 표본평균 $\bar{x}$=$540, 표준편차 s=$299 통계적인 가설 검증의 4단계 과정을 통하여 유의성을 검증해 보자.

  - 1단계 : 모든 가설을 세운다.
    - ✓ 귀무 가설 H_0 : 자동차의 평균 수리비용 $\mu \geq$ 1000달러
    - ✓ 대립 가설 H_1 : 자동차의 평균 수리비용 $\mu <$ 1000달러

  - 2단계 : 검정통계량.

$$t = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

    - ✓ s는 표본의 표준편차, μ_0는 가상 평균치 $1,000 이다.
    - ✓ 측정한 t값이 왼쪽$-t_{0.05}$에 있으면(H_1를 지지하려면 $\bar{x}$-μ_0 가 음수여야 한다) H_1을 지지한다.



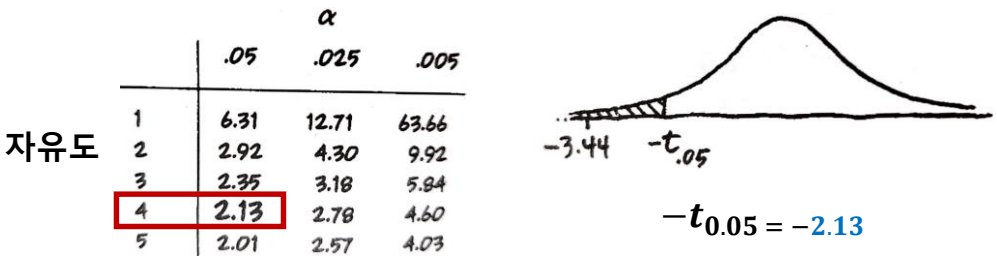
가설검증

- 3단계 : p값.

n=5, 표본평균 $\bar{x}$=$540, 표준편차 s=$299

$$t_{OBS} = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{540 - 1000}{\frac{299}{\sqrt{5}}} = -3.44$$

$$p값 = Pr(t < t_{OBS} \mid H_0) = Pr(t < -3.44 \mid H_0) = 0.00029$$



자유도	α = .05	.025	.005
1	6.31	12.71	63.66
2	2.92	4.30	9.92
3	2.35	3.18	5.84
4	2.13	2.78	4.60
5	2.01	2.57	4.03

$-t_{0.05} = -2.13$

t분포의 임계점

```
# p=0.05이고, 자유도가 4,
# t분포 분위수 함수 : qt(p, df, lower.tail = TRUE/FALSE)
qt(p=0.05, df=4, lower.tail = FALSE)
#[1] 2.131847

# 검정통계량
(540-1000)/(299/sqrt(5))
#[1] -3.440105
pt(q=c(-3.440105), df=4)
#[1] 0.0002907443
```

가설검증

- 4단계 : p값과 정해진 유의수준 α를 비교.
  - α는 어떤 결과가 통계적으로 의미 있다고 판단하는 기준점이다.
  - 즉 p값 $\leq \alpha$ 이면, H_0를 기각하고 H_1을 채택 한다.
    - ✓ 여기서 유의수준 α는 p값 ≤ 0.05 유의한 결과로 판다.
    - ✓ p값 $= 0.00029 < 0.05$ 이므로 H_0를 기각하고 H_1를 채택한다.
    - ✓ 즉 평균 수리비용이 $1,000 이내이므로 자동차 보험에 가입이 가능하다.

결정이론

- 가설검증의 유의성 검증을 가정용 연기탐지기에 비유해볼 수 있다.
  - 화재가 없는데도 경보가 울리는 것을 제1종 오류라고 한다.
  - 반대로 경보가 없는 화재는 제2종 오류라고 한다.
  - 제1종 오류를 피하기 위해 화재 탐지기를 제거하면 제2종 오류의 발생률을 증가시킴, 반대로 탐지기의 감도를 높여 제2종 오류의 가능성을 줄이면 오경보의 횟수가 늘어난다.

[2 X 2 결정표]

	화재 없음	화재 발생
무경보	무오류	제2종 오류
경보	제1종 오류	무오류

t-검정, t-value, p-value

- (예제) 약리학자가 한의약 기술을 이용해 부작용이 거의 없는 수면제 세 종류를 개발하고, 수면제의 수면 효과를 측정하기 위해 임상 실험을 각각 수행하여 다음 데이터를 얻었다. data1에서 30은 수면시간이 30분 늘었음을 뜻하고, -5는 5분만큼 줄었음을 뜻한다.
  - 수면제1 data1 = { 30, -5, 55, -30, -20, 45 }
  - 수면제2 data2 = { 12, 13, 12, 13, 12, 13 }
  - 수면제3 data3 = { 30, -5, 55, -30, -20, 45, 30, -5, 55, -30, -20, 45 }

$$\text{t} - \text{통계량} = \frac{\mu - \mu_0}{\frac{s}{\sqrt{n}}}$$

- ✓ 귀무 가설 $H_0 : \mu = 0$ (수면제는 효과가 없다)
- ✓ 대립 가설 $H_1 : \mu > 0$ (수면제는 효과가 있다)

t-검정, t-value, p-value

- R로 t-검정 수행 : data1, data2, data3의 평균, 표준편차

```
# 세 종류 데이터
data1 <- c(30, -5, 55, -30, -20, 45)
data2 <- c(12, 13, 12, 13, 12, 13)
data3 <- c(30, -5, 55, -30, -20, 45, 30, -5, 55, -30, -20, 45)

# data1  평균, 표준편차
mean(data1)
#[1] 12.5
sd(data1)
#[1] 35.60197
# data2  평균, 표준편차
mean(data2)
#[1] 12.5
sd(data2)
#[1] 0.5477226
# data3  평균, 표준편차
mean(data3)
#[1] 12.5
sd(data3)
#[1] 33.94514
```

t-검정, t-value, p-value

- R로 t-검정 수행 : data1의 t-검정, t-value, p-value 계산

```
# t-검정
t.test(data1, alternative = c("greater"))
#
#One Sample t-test
#
#data:  data1
#t = 0.86003, df = 5, p-value = 0.2145
#alternative hypothesis: true mean is greater than 0
#95 percent confidence interval:
#  -16.7876      Inf
#sample estimates:
#  mean of x
#  12.5

# t-value
(12.5)/(35.60197/sqrt(6))
#[1] 0.8600261
# p-value
1- pt(q=c(0.8600261), df=5)
#[1] 0.214537
```

✓ 수면제가 수면 시간을 늘려주는지 여부가 관심이므로 μ_0=0으로 설정

$$t - 통계량 = \frac{\mu - \mu_0}{\frac{s}{\sqrt{n}}}$$

t-검정, t-value, p-value

- R로 t-검정 수행 : data2의 t-검정, t-value, p-value 계산

```
# t-검정
t.test(data2, alternative = c("greater"))
#
#One Sample t-test
#
#data:  data2
#t = 55.902, df = 5, p-value = 1.732e-08
#alternative hypothesis: true mean is greater than 0
#95 percent confidence interval:
#  12.04942      Inf
#sample estimates:
#  mean of x
#       12.5

# t-value
(12.5)/(0.5477226/sqrt(6))
#[1] 55.9017
# p-value
1- pt(q=c(55.9017), df=5)
#[1] 1.732451e-08
```

✓ 수면제가 수면 시간을 늘려주는지 여부가 관심이므로 μ_0=0으로 설정

$$t-\text{통계량} = \frac{\mu - \mu_0}{\frac{s}{\sqrt{n}}}$$

t-검정, t-value, p-value

- R로 t-검정 수행 : data3의 t-검정, t-value, p-value 계산

```
# t-검정
t.test(data3, alternative = c("greater"))
#
#One Sample t-test
#
#data:  data3
#t = 1.2756, df = 11, p-value = 0.1142
#alternative hypothesis: true mean is greater than 0
#95 percent confidence interval:
#  -5.098089       Inf
#sample estimates:
#  mean of x
#       12.5

# t-value
(12.5)/(33.94514/sqrt(12))
#[1] 1.275625
# p-value
1- pt(q=c(1.275625), df=11)
#[1] 0.114184
```

✓ 수면제가 수면 시간을 늘려주는지 여부가 관심이므로 μ_0=0으로 설정

$$t - 통계량 = \frac{\mu - \mu_0}{\frac{s}{\sqrt{n}}}$$

t-검정, 분산분석(ANOVA)

- 앞의 수면제 예처럼 데이터 하나만으로 수행하는 t-검정을 단일 표본 t-검정(one-sample t-test)
- 두 집단(페이지 A, 페이지 B)의 평균 세션 시간 차이를 비교하는 예처럼 2개의 데이터에 대해 수행하는 t-검정을 두 표본 t-검정(two-sample t-test)
- 두 표본 t-검정은 기본 개념과 절차가 단일 표본 t-검정과 비슷함
- 분산분석은 3개 이상의 집단을 비교하는데 사용
- 분산분석은 많은 개념과 절차를 t-검정과 공유하기 때문에 t-검정을 제대로 이해하면 분산분석을 쉽게 이해할 수 있음

카이제곱검정

- 카이제곱 검정(Chi-Squared Test)은 종속변수가 범주형 자료(categorical data)인 경우에 사용하는 분석기법

- 범주형 자료 분석은 크게 적합도 검정(goodness of fit test), 독립성 검정(test of independence), 동질성 검정(test of homogeneity)의 3가지로 분류할 수 있음

  - 적합도 검정
    - ✓ 관측값들이 어떤 이론적 분포를 따르고 있는지를 검정
    - ✓ 한 개의 요인을 대상으로 함

  - 독립성 검정
    - ✓ 서로 다른 요인들에 의해 분할되어 있는 경우 그 요인들이 관찰값에 영향을 주고 있는지 아닌지, 요인들이 서로 연관이 있는지 없는지를 검정
    - ✓ 두 개의 요인을 대상으로 함

카이제곱검정

- **동질성 검정**
  - ✓ 관측값들이 정해진 범주 내에서 서로 비슷하게 나타나고 있는지를 검정
  - ✓ 속성 A, B를 가진 부모집단(subpopulation) 각각으로부터 정해진 표본의 크기만큼 자료를 추출하는 경우에 분할표에서 부모집단의 비율이 동일한가를 검정
  - ✓ 두 개의 요인을 대상으로 함

카이제곱검정 – 적합도 검정

- 적합도 검정(goodness of fit test)
  - 적합도 검정은 k개의 범주(또는 계급)을 가지는 한 개의 요인(factor)에 대해서 어떤 이론적 분포를 따르고 있는지를 검정하는 방법
  - 기본원리는 도수분포로 각 구간에 있는 관측도수를 $O_1, O_2, \dots, O_k$라 하고, 각 범주(또는 계급)가 일어날 확률을 $p_1, p_2, \dots, p_k$라고 할 때 기대되는 관측도수는 $E_1, E_2, \dots, E_k$를 계산하여 실제 관측도수와 기대 관측도수의 차이를 카이제곱검정 통계량(Chi-squared statistics)을 활용하여 가정한 확률모형에 적합한지를 평가

✓ 귀무 가설 H_0 : 관측값의 도수와 가정한 이론 도수(기대 관측도수)가 동일하다.

✓ 대립 가설 H_1 : 적어도 하나의 범주(또는 계급)의 도수가 가정한 이론 도수(기대 관측도수)와 다르다.

✓ 검정통계량$(X^2) = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

✓ 검정 방법 : ${\chi^2}_0 \geq {X^2}_\alpha(k-1)$ 또는 $p - value = P\left(X^2 \geq {\chi^2}_0\right) < \alpha$ 이면 H_0를 기각

※ 관측된 자료로부터 계산된 검정통계량의 값을 ${\chi^2}_0$

카이제곱검정 – 적합도 검정

- (예제) 유전학자 멘델은 콩 교배에 대한 유전의 이론적 모형으로서 잡종비율을 A : B : C = 2 : 3 : 5라고 주장하였다. 이 이론의 진위를 가리기 위해 두 콩 종자의 교배로 나타난 100개의 콩을 조사하였더니 A형 20개, B형 40개, C형 40개이다. 이러한 관찰값을 얻었을 때 멘델 유전자의 이론이 맞다고 할 수 있는지를 유의수준 α = 0.05에서 검정하기.

```
obs <- c(20, 40, 40)
obs.probs <- c(2/10, 3/10, 5/10)
(g.fit <- chisq.test(obs, p=obs.probs))
#
#     Chi-squared test for given probabilities
#
#data:  obs
#X-squared = 5.3333, df = 2, p-value = 0.06948
```

  - ✓ 위 분석결과를 보면,
  - ✓ p-value가 0.06948 이므로 유의수준 α 0.05 보다 크므로 귀무가설 H0를 채택하고 대립가설 H1을 기각하여 '멘델이 주장한 콩의 잡종비율 이론적 분포는 적합하다'고 판단할 수 있다.

카이제곱검정 – 적합도 검정

```
g.fit$observed    # observed frequency
#[1] 20 40 40
g.fit$expected    # expected frequeycy
#[1] 20 30 50
g.fit$statistic   # chi-squared statistics
#X-squared
# 5.333333
g.fit$parameter   # degrees of freedom
#df
# 2
g.fit$p.value     # P-value
#[1] 0.06948345

# 카이제곱분포의 누적 분포 함수를 이용한 p-value 계산
1 - pchisq(q=c(5.3333), df=2, lower.tail=TRUE)
#[1] 0.06948461
```

카이제곱검정 – 독립성 검정

- 독립성 검정(test of independence)
  - 독립성 검정은 두 개의 범주형 변수/요인이 서로 연관성이 있는지, 상관이 있는지, 독립적인지를 카이제곱 검정을 통해 통계적으로 판단하는 방법
  - 예를 들어, 학력(고졸, 대졸, 대학원졸)이라는 범주형 변수(variable X)/요인(factor 1)와 연소득(하, 중, 상)이라는 범주형 변수(variable Y)/요인(factor 2) 간에 서로 관련성이 있는 것인지 아니면 관련이 없이 독립적인지를 판단하는 것과 같은 문제에 독립성 검정을 사용
  - 자료를 분류하는 두 변수를 X와 Y라고 하고, 변수 X는 m개, 변수 Y는 n개의 범주(또는 계급)를 가진다고 했을 때 관측도수 O_{ij}는 m개와 n개의 층으로 이루어진 표로 정리하면 다음과 같음(이를 m x n 분할표)

[독립성 검정 자료 구조]

X \ Y	b_1	b_2	…	b_n	total
a_1	O_{11}	O_{12}	…	O_{1n}	O_{1*}
a_2	O_{21}	O_{22}	…	O_{2n}	O_{2*}
…	…	…	…	…	…
a_m	O_{m1}	O_{m2}	…	O_{mn}	O_{m*}
total	$O_{.1}$	$O_{.2}$	…	$O_{.n}$	n

카이제곱검정 – 독립성 검정

- 독립성 검정(test of independence)
  - 독립성 검정에는 카이제곱(χ^2) 통계량을 사용
  - 귀무가설 H0가 사실일 때 자유도 (m-1)(n-1)인 카이제곱분포에 근사
  - 검정통계량 카이제곱(χ^2)은 각 범주의 기대도수가 5 이상인 경우에 사용하는 것이 바람직하며, 기대도수가 5 미만인 경우에는 주의 필요(5보다 작으면 인접 범주와 병합하는 것도 방법)
  - 기본원리는 관측도수 O11, O21, ..., Omn이 기대도수 E11, E21, ..., Emn과 차이가 없다면 검정통계량 $\chi_0{}^2$ 값이 '0'이 되고, 반대로 관측도수와 기대도수가 차이가 크다면 검정통계량 값 또한 커지게 된다는 것임

✓ 귀무 가설 H_0 : 두 변수 X와 Y는 서로 독립이다(관련성이 없다).

✓ 대립 가설 H_1 : 두 변수 X와 Y는 서로 독립이 아니다(관련성이 있다).

✓ 검정통계량$(X^2) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

✓ 검정 방법 : $\chi^2{}_0 \geq X^2{}_\alpha[(m-1)(n-1)]$ 또는 $p-value = P(X^2 \geq \chi^2{}_0) < \alpha$ 이면 H_0를 기각

※ 관측된 자료로부터 계산된 검정통계량의 값을 $\chi^2{}_0$

카이제곱검정 – 독립성 검정

- (예제) 학급(class 1, class 2, class 3)과 빅데이터 분석 성적(score A, B, C, F) 간의 관련성이 있는지를 조사한 아래의 분할표를 사용하여 유의수준 α = 0.05 로 검정하기.

[학급과 빅데이터 분석 성적 분할표]

	Score A	Score B	Score C	Score F
Class 1	7	13	9	12
Class 2	13	21	10	19
Class 3	11	18	12	13

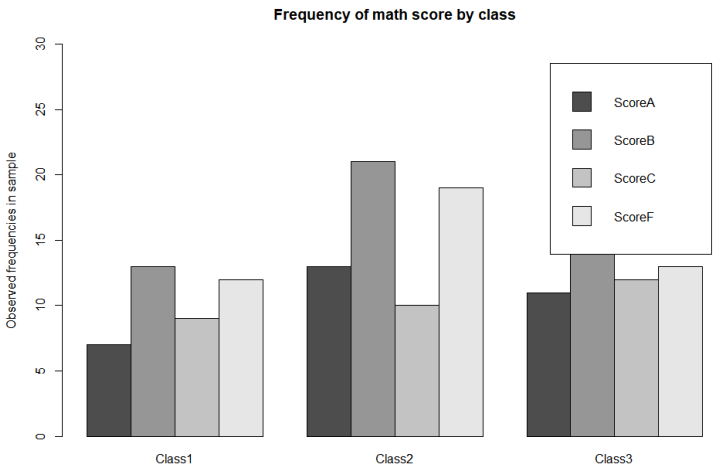
카이제곱검정 – 독립성 검정

```
# 카이제곱검정 - 독립성 검정
raw_data <- c(7, 13, 9, 12, 13, 21, 10, 19, 11, 18, 12, 13)
data_mtx <- matrix(raw_data, byrow=TRUE, nrow=3)
data_mtx
#      [,1]  [,2] [,3]  [,4]
#[1,]    7   13    9   12
#[2,]   13   21   10   19
#[3,]   11   18   12   13

dimnames(data_mtx) <- list("Class" = c("Class1", "Class2", "Class3"),
                           "Score" = c("ScoreA", "ScoreB", "ScoreC", "ScoreF"))
data_mtx
#        Score
#Class    ScoreA ScoreB ScoreC ScoreF
#  Class1      7     13      9     12
#  Class2     13     21     10     19
#  Class3     11     18     12     13

# marginal distribution : addmargins()
addmargins(data_mtx)
#        Score
#Class    ScoreA ScoreB ScoreC ScoreF Sum
#  Class1       7      13       9     12  41
#  Class2      13      21      10     19  63
#  Class3      11      18      12     13  54
#  Sum         31      52      31     44 158
```

카이제곱검정 – 독립성 검정

```
# proportional distribution : prop.table()
addmargins(prop.table(data_mtx))
#           Score
#Class          ScoreA       ScoreB       ScoreC       ScoreF         Sum
#  Class1  0.04430380  0.08227848  0.05696203  0.07594937  0.2594937
#  Class2  0.08227848  0.13291139  0.06329114  0.12025316  0.3987342
#  Class3  0.06962025  0.11392405  0.07594937  0.08227848  0.3417722
#  Sum     0.19620253   0.32911392  0.19620253  0.27848101 1.0000000
```

```
# bar plot
barplot(t(data_mtx), beside=TRUE, legend=TRUE,
        ylim=c(0, 30),
        ylab="Observed frequencies in sample",
        main="Frequency of math score by class")
```



카이제곱검정 – 독립성 검정

```
# chisquared test : chisq.test()
(i.fit <- chisq.test(data_mtx))
#
#        Pearson's Chi-squared test
#
#data:  data_mtx
#X-squared = 1.3859, df = 6, p-value = 0.9667

# 카이제곱분포의 누적 분포 함수를 이용한 p-value 계산
1- pchisq(q=c(1.3859), df=6, lower.tail=TRUE)
#[1] 0.9667105
```

- ✓ 위 분석결과를 보면,
- ✓ p-value가 0.9667 이므로 유의수준 α 0.05 보다 크므로 귀무가설 H0를 채택하여 '학급과 빅데이터 분석 성적 간에는 서로 관련성이 없다. 즉 독립적이다.'고 판단할 수 있다.

카이제곱검정 – 동질성 검정

- 동질성 검정(test of homogeneity)
  - 동질성 검정은 r개의 행과 c개의 열을 가진 두 변수 X와 Y로부터 작성된 분할표의 각 열분포에서 행들이 균일한 값을 가지는지 즉, 각 열에서 행들의 동질성를 검정하는 것임
  - 동질성 검정은 부모집단(subpopulation)을 먼저 설정한 후 각 부모집단으로부터 정해진 표본의 크기만큼 무작위로 추출하여 분할표에서 부모집단의 비율이 동일한가를 검정
  - 가령, 소득수준에 따라 지지 정당이 동일한지 여부를 검정한다고 할 때, 우선 소득수준을 부모집단으로 설정하고, 각 소득수준별로 정해진 크기의 표본을 무작위로 추출하는 것임
  - r개의 행과 c개의 열을 가진 두 변수 X와 Y로부터 작성된 r x c 분할표를 이용한 동질성 검정을 위한 자료 구조는 다음과 같음

[동질성 검정 자료 구조]

x \ Y	b_1	b_2	…	b_c	total
a_1	O_{11}	O_{12}	…	O_{1c}	O_{1*}
a_2	O_{21}	O_{22}	…	O_{2c}	O_{2*}
…	…	…	…	…	…
a_r	O_{r1}	O_{r2}	…	O_{rc}	O_{r*}
total	$O_{.1}$	$O_{.2}$	…	$O_{.c}$	n

카이제곱검정 – 동질성 검정

- 동질성 검정(test of homogeneity)
  - 검정통계량 카이제곱(χ^2)은 귀무가설 H_0가 사실일 때 근사적으로 자유도(r-1)(c-1)인 카이제곱 분포를 따름

  ✓ 귀무 가설 $H_0 : p_{1j} = p_{2j} = \ldots = p_{rj}$, j=1, ..., c

  ✓ 대립 가설 $H_1 : H_0$가 아니다.

  ✓ 검정통계량$(X^2) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

  ✓ 검정 방법 : ${\chi^2}_0 \geq {X^2}_{\alpha}[(r-1)(c-1)]$ 또는 $p - value = P(X^2 \geq {\chi^2}_0) < \alpha$ 이면 H_0를 기각

※ 관측된 자료로부터 계산된 검정통계량의 값을 ${\chi_0}^2$

카이제곱검정 – 동질성 검정

- (예제) 대학교 4학년 남학생 100명과 여학생 200명을 무작위로 추출하여 데이터 사이언스 교과목 관련 선호도 조사를 수행하고, 유의수준 α = 0.05에서 남학생과 여학생의 선호 교과목이 동일한지 검정하기.

[데이터 사이언스 교과목 선호도 조사 결과]

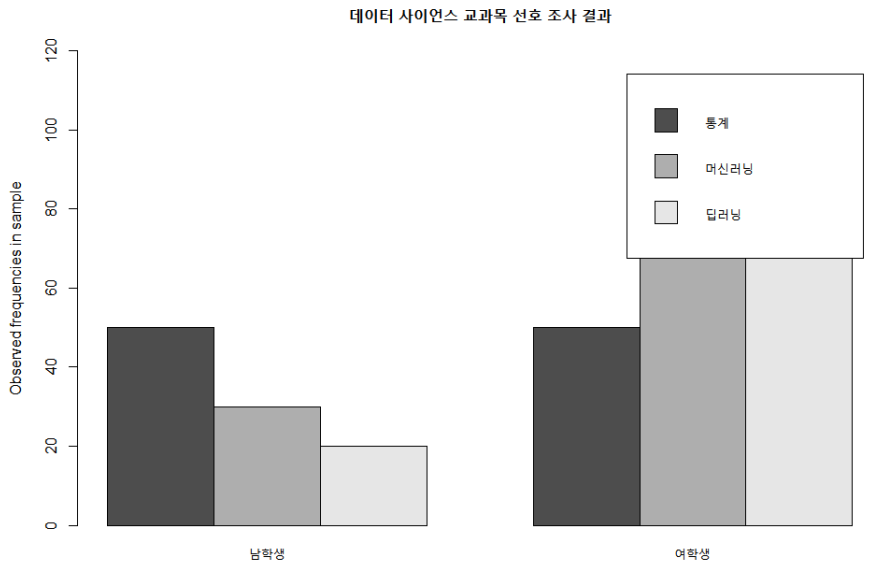
	통계	머신러닝	딥러닝	Row total
남학생	50	30	20	100
여학생	50	80	70	200
Column total	100	110	90	300

카이제곱검정 – 동질성 검정

```
# 카이제곱검정 - 동질성 검정
raw_data <- c(50, 30, 20, 50, 80, 70)
data_mtx <- matrix(raw_data, byrow=TRUE, nrow=2)
data_mtx
#      [,1]  [,2]  [,3]
#[1,]   50   30   20
#[2,]   50   80   70

dimnames(data_mtx) <- list("성별" = c("남학생", "여학생"),
                           "DS교과목" = c("통계", "머신러닝", "딥러닝"))
data_mtx
#          DS교과목
#성별      통계 머신러닝 딥러닝
#  남학생   50        30      20
#  여학생   50        80      70

# marginal distribution : addmargins()
addmargins(data_mtx)
#          DS교과목
#성별      통계 머신러닝 딥러닝 Sum
#  남학생    50        30      20   100
#  여학생    50        80      70   200
#  Sum      100      110      90   300
```

카이제곱검정 – 동질성 검정

```
# proportional distribution : prop.table()
addmargins(prop.table(data_mtx))
#           DS교과목
#성별              통계     머신러닝        딥러닝        Sum
#  남학생  0.1666667 0.1000000 0.06666667 0.3333333
#  여학생  0.1666667 0.2666667 0.23333333 0.6666667
#  Sum     0.3333333 0.3666667 0.30000000 1.0000000

# bar plot : barplot()
barplot(t(data_mtx), beside=TRUE, legend=TRUE,
        ylim=c(0, 120),
        ylab="Observed frequencies in sample",
        main="데이터 사이언스 교과목 선호 조사 결과")
```



카이제곱검정 – 동질성 검정

```
# chisquared test : chisq.test()
(h.fit <- chisq.test(data_mtx))
#
#      Pearson's Chi-squared test
#
#data:  data_mtx
#X-squared = 19.318, df = 2, p-value = 6.384e-05

# 카이제곱분포의 누적 분포 함수를 이용한 p-value 계산
1- pchisq(q=c(19.318), df=2, lower.tail=TRUE)
#[1] 6.384834e-05
```

- ✓ 위 분석결과를 보면,
- ✓ 카이제곱 통계량 값이 19.318이 나왔고 p-value가 6.384e-05 이므로 유의수준 α=0.05 보다 훨씬 작기 때문에 귀무가설 H_0를 기각하고 대립가설 H_1을 채택하여 '남학생/여학생별 선호하는 데이터 사이언스 교과목이 동일하지 않다'고 판단할 수 있다.

연속확률분포 - 카이제곱분포

- 다음은 카이제곱분포의 확률 밀도 함수, 누적 분포 함수, 분위수 함수, 난수 발생 등을 위한 R 함수 및 모수는 다음과 같다.

구분		정규분포 R 함수/모수
밀도 함수 Density function	d	dchisq(x, df)
누적 분포 함수 Cumulative distribution function	p	pchisq(q, df, lower.tail=TRUE/FALSE)
분위수 함수 Quantile function	q	qchisq(p, df, lower.tail=TRUE/FALSE)
난수 발생 Random number generation	r	rchisq(n, df)

- ✓ lower.tail=TRUE 인 경우, probabilities is $P[X \leq x]$.
- ✓ lower.tail=FALSE 인 경우, probabilities is $P[X > x]$.