

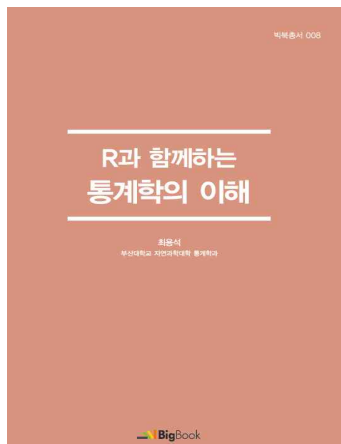
Basic Statistics & Visualizations of Data with



: EDA Exploratory Data Analysis

CDA Categorical Data Analysis

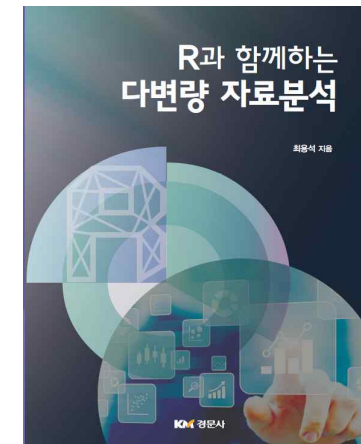
MDA Multivariate Data Analysis



최 용 석 (부산대학교 통계학과)

yschoi.pusan.ac.kr

2018.01.09.~10.



부산대학교 간호과학연구소 워크숍

Program

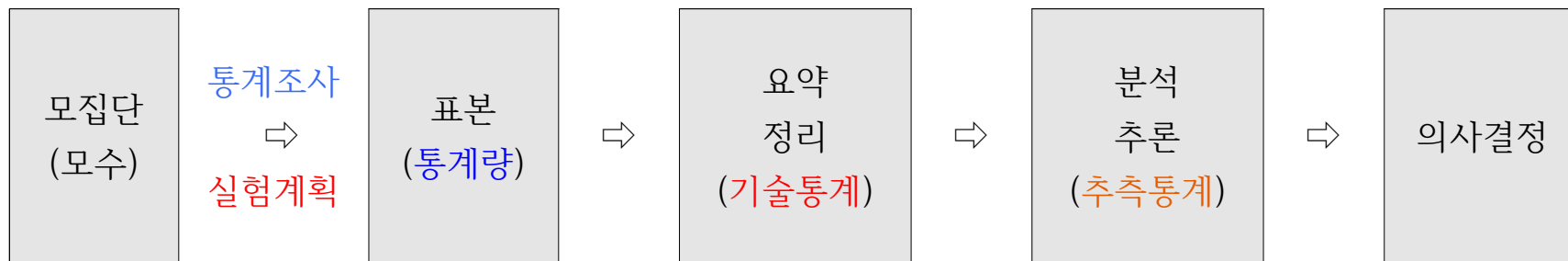
주제 1 : R과 함께하는 기초 통계

- R의 이해와 응용 : 기초 통계 모음

주제 2: 자료의 시각화

- EDA를 위한 시각화 (Visualizations for EDA) : 다중산점도, 상자그림, 줄기-잎 그림, 3차원 그림
- 범주형자료의 시각화 (Visualizations for CDA) : 단순과 다중 대응분석 그림
- 다변량자료의 시각화 (Visualizations for MDA) : 행렬도

주제 1 : R과 함께하는 기초 통계



통계자료 분석의 단계

1.1 자료의 이해

1.2 한 집단의 비교 : 모평균 검정, 모비율 검정

1.3 두 집단의 비교 : 독립표본, 대응표본, 모비율

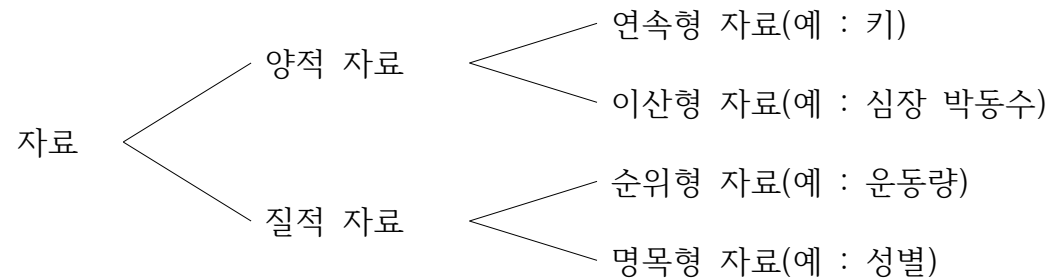
1.4 여러 집단의 비교 : 일원 분산분석

1.5 범주들의 관계 비교 : 독립성, 동질성 검정

부록 : R의 설치 및 사용법

1.1 자료의 이해

자료의 종류



[표 1.1.1] 심장 박동수 자료

학생	처음 심장 박동수	나중 심장 박동수	달리기	흡연	성별	키(cm)	몸무게(kg)	운동량
1	64	88	1	2	1	168	63.5	2
2	58	70	1	2	1	183	65.8	2
3	62	76	1	1	1	185	72.6	3
4	66	78	1	1	1	185	86.2	1
5	64	80	1	2	1	175	70.3	2
6	64	60	2	2	2	168	81.6	3
7	94	92	2	1	2	157	82.1	2
8	60	66	2	2	2	157	54.4	2
9	72	70	2	2	2	173	78.5	2
10	58	56	2	2	2	170	56.7	2

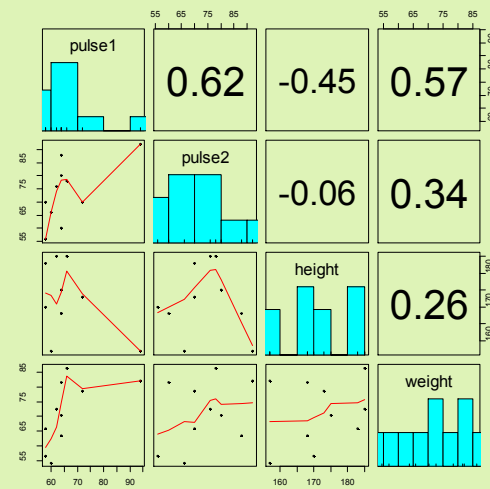
[보기 1.1.2] [표 1.1.1] 심장 박동수 자료에서 양적 자료의 요약

[표 1.1.1]에서 양적 자료는 처음 심장 박동수, 나중 심장 박동수, 키(cm), 몸무게(kg)로 이들에 대해 동시에 요약하기로 하자.

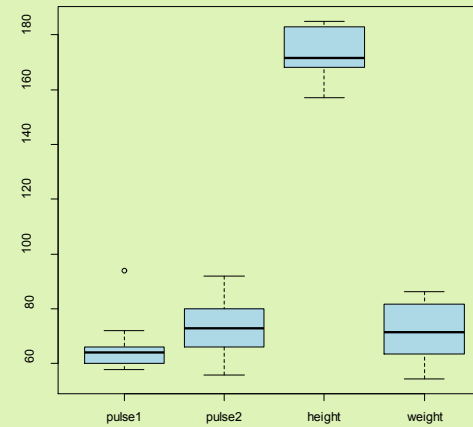
[1] 기술통계(descriptive statistics)

pulse1	pulse2	height	weight
Min. :58.0	Min. :56.0	Min. :157.0	Min. :54.40
1st Qu.:60.5	1st Qu.:67.0	1st Qu.:168.0	1st Qu.:64.08
Median :64.0	Median :73.0	Median :171.5	Median :71.45
Mean :66.2	Mean :73.6	Mean :172.1	Mean :71.17
3rd Qu.:65.5	3rd Qu.:79.5	3rd Qu.:181.0	3rd Qu.:80.83
Max. :94.0	Max. :92.0	Max. :185.0	Max. :86.20

[2] 다중 산점도(multiple scatter plot)



[3] 상자그림(box plot)



[4] 줄기-잎 그림(stem-and-leaf plot)

```

5 | 88
6 | 024446
7 | 2
8 |
9 | 4
    
```

pulse1

```

5 | 6
6 | 06
7 | 0068
8 | 08
9 | 2
    
```

pulse2

```

15 | 77
16 | 88
17 | 035
18 | 355
    
```

height

```

5 | 47
6 | 46
7 | 039
8 | 226
    
```

weight

[R-코드 1.1.1] [보기 1.1.1]과 [보기 1.1.2] 심장 박동수 자료의 표현(심장박동수자료요약.R)

```
setwd("C:/Users/stat/Desktop/개인_WORK/외부초청강의/부산대간호  
과학연구소/R-codes")  
data1.1.1<-read.table("pulse.txt", header=T)  
x<-data1.1.1  
x  
## [보기 1.1.1] 질적자료의 표현  
# 운동량: 막대그림, 원도표  
win.graph()  
par(mfrow=c(1,2))  
activity=x[,8]  
freq=table(activity)  
rel_freq=prop.table(freq)  
cbind(freq, rel_freq)  
y=round(rel_freq*100)  
z=paste(lab=c("1:적음", "2:보통", "3:많음"), ("y","%",""))  
pie(freq,labels=z)  
barplot(freq, xlab="운동량(1:적음, 2:보통, 3:많음)", ylab="빈도")  
  
## [보기 1.1.2] 양적자료의 표현 : 처음과 나중 심장 박동수, 키,  
몸무게  
x1<-x[, c(1,2,6,7)]  
  
# [1] 기술통계  
summary(x1)  
# [2] 다중산점도  
library(psych)  
pairs.panels(x1, density=F, ellipses=F)  
# [3] 상자그림  
boxplot(x1, col="lightblue")  
# [4] 줄기-잎 그림  
for (j in 1:4){  
  stem(x1[,j])  
}
```

질적 자료

- pie(), barplot() : 원도표, 막대그림
- table() : 분할표

양적 자료

- summary() : 평균(mean), 분산(variance), 등
기초통계량
- library(psych) > pairs.panels() : 다중 산점도
- boxplot() : 상자그림
- stem() : 줄기-잎 그림

1.2 한 집단의 비교 : 모평균 검정, 모비율 검정

대표본의 모평균에 대한 유의수준 α 에서 Z -검정

귀무가설	대립가설	p -값	검정 통계량
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P[Z \geq z]$	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ $\approx N(0,1)$
	$H_1 : \mu < \mu_0$	$P[Z \leq z]$	
	$H_1 : \mu \neq \mu_0$	$P[Z \geq z]$	

소표본의 모평균에 대한 유의수준 α 에서 t -검정

귀무가설	대립가설	p -값	검정 통계량
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P[T \geq t]$	$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$
	$H_1 : \mu < \mu_0$	$P[T \leq t]$	
	$H_1 : \mu \neq \mu_0$	$P[T \geq t]$	

대표본의 모비율에 대한 유의수준 α 에서 Z -검정

귀무가설	대립가설	p -값	검정 통계량
$H_0 : \mu = \mu_0$	$H_1 : p > p_0$	$P[Z \geq z]$	$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0/n}} \approx N(0,1)$
	$H_1 : p < p_0$	$P[Z \leq z]$	
	$H_1 : p \neq p_0$	$P[Z \geq z]$	

[보기 1.2.1] 대표본 발치 난이도에 대한 가설 및 검정의 이해

발치 난이도에 대한 표본자료 [표 1.2.1]에서 $n = 826$ 이고 표본평균 $\bar{x} = 3.408$ 와 표준편차 $s = 0.722$ 이므로 유의수준 0.05에서 $H_0 : \mu = 3.8$, $H_1 : \mu < 3.8$ 를 검정통계량의 관측값과 유의확률 p -값을 구하여 검정하라.

[표 1.2.1] 매복치 자료 일부(molar.txt)

자료 파일 molar.txt의 34개 변수 설명

한글	영문	범주 및 변수값
성별	sex	M(1), F(0)
나이	age	6세~75세
치아번호	T	15~48번, 90번
동반 합병증	C	C1~C7: 발생(1), 미발생(0)
발치난이도	Di	1~5
높은 발치난이도 추정원인	DiR	DR1~DR9: 발생(1), 미발생(0)
발치 후 합병증	PC	PC1~PC9: 발생(1), 미발생(0)
영상 명확성	Im	명확(1), 불명확(0)
하악관과의 접촉여부	MC	접촉(1), 미접촉(0)
방 사 선 소 견	PG	1(A1), 2(A2), 3(A3), 4(B1), 5(B2), 6(B3), 7(C1), 8(C2), 9(C3)
매복된 하악 제3대구치 수직 및 전후방 관계를 고려한 분류	W	1(MA), 2(V), 3(H), 4(DA), 5(etc)
발치 난이도 예측 지수	PI	1~10

[1] $\bar{X}$ 의 분포는 근사적으로 $N(3.8, \frac{0.722^2}{826})$ 이 되고 이를 표준화한 검정통계량은 근사적으로 표준정규분포를 따른다.

$$Z = \frac{\bar{X} - 3.8}{0.722/\sqrt{826}} \simeq N(0,1)$$

[2] 표본평균이 $\bar{x} = 3.408$ 이므로 검정통계량의 관측값 :

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.408 - 3.8}{0.722/\sqrt{826}} = -15.615$$

[3] 대립가설의 형태 $H_1 : \mu < 3.8$ 에 따른 유의확률:

$$p\text{-값} = P[Z \leq z] = P[Z \leq -15.615] = 0.000$$

[4] 유의수준 $\alpha = 0.05$ 에서 $p\text{-값} < \alpha$ 이므로 귀무가설 H_0 를 기각한다.

[5] 이 자료에 따르면 새 발치 방법은 난이도 평균이 기존 발치 방법에 비해 낮음을 알 수 있다.

[R-코드 1.2.1] [보기 1.2.1] 대표본의 모평균에 대한 z -검정(매복치발치난이도-Z-검정.R)

```
setwd("C:/Users/stat/Desktop/개인_WORK/외부초청강의/부산대간호과학연구소/R-codes")
#### original data(826x34) ####
data1.2.1 <- read.table("molar.txt", header=T)
data<-data1.2.1

## [보기 1.2.1] 양적자료의 표현
발치난이도=data$Di
m=mean(발치난이도)
v=var(발치난이도)
s=sd(발치난이도)
list(m, v, s)

## Z-test
z.test = function(x, mu){
  zeta = (mean(x) - mu)/(sqrt(var(x)/length(x)))
  return(zeta)}
x=발치난이도
z=z.test(x, 3.8)

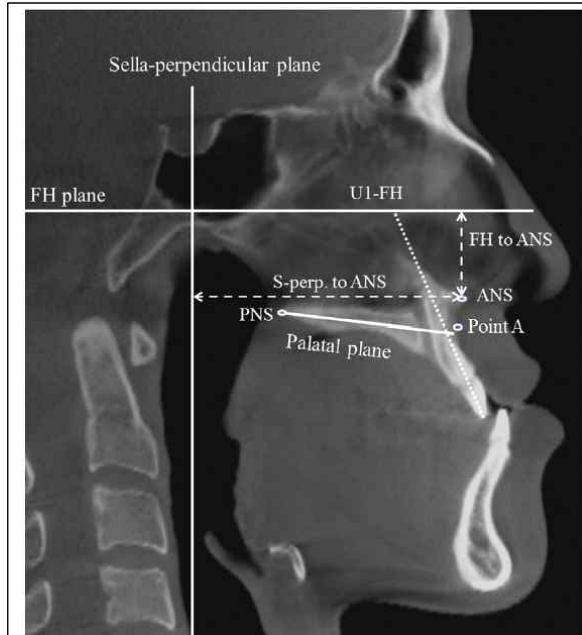
# 기각치(유의수준 5%), p-값
cvalue=qnorm(1-0.05)
pvalue=pnorm(z)
list(z, cvalue, pvalue)
```

$$H_0 : \mu = 3.8, H_1 : \mu < 3.8$$

- z.test():
- $z=z.test(x, 3.8)$: x =발치난이도, 모평균 3.8
- 발치난이도=data\$Di
- $pvalue=pnorm(z)$: $p\text{-값} = P[Z \leq -15.615]$

```
>list(z, cvalue, pvalue)
[[1]]
[1] -15.61452
[[2]]
[1] 1.644854
[[3]]
[1] 2.899154e-55
```

[보기 2.3.3] 주걱턱 수술 후 상악(윗턱) 전치부 변형 수평측 거리 변화의 자료 일부



Cone-beam computed tomographic (CT) images :

- Sella perpendicular : 두개골의 기준점인 sella에서 수직으로 내린 기준 선
- ANS(anterior nasal spine): 측면에서 바라볼 때 입천장의 가장 앞쪽 뾰족한 부분
- FH(frankfort horizontal) Plane: 정면을 바라볼 때 바닥과 평행해지는 선으로 보통 치과 영역에서 수평 기준선
- point A: 윗턱에서 ANS부터 시작되는 만곡점 중 가장 후방부
- S-perpendicular to the PNS: 수평측 거리 변화(mm)

환자 10명의 표본자료는 정규분포를 따른다.

23.46, 20.94, 20.28, 19.55, 18.05, 18.03, 23.39, 16.84, 12.47, 28.80

수평측 거리 변화가 20보다 큰 지를 귀무가설과 대립가설을 세우고 이를 유의수준 $\alpha = 0.01$ 에서 검정하여 보라.

[1] μ 를 거리 변화라 하면 $\mu > 20$ 이고 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 20, \quad H_1 : \mu > 20$$

[2] 표본의 크기가 $n=10$ 인 소표본이며 정규분포를 따르므로 t-검정을 실시할 수 있다. 검정통계량의 계산:

$$\bar{x} = 20.181, \quad s = 4.42$$

$$t = \frac{20.181 - 20}{4.42/\sqrt{10}} = \frac{0.181}{1.398} = 0.129$$

[3] 대립가설 $H_1 : \mu > 20$ 의 형태에 따른 자유도 $df = n - 1 = 9$ 일 때 유의확률:

$$p\text{-값} = P[T \geq t] = P[T \geq 0.129] = 0.450$$

[4] 유의수준 $\alpha = 0.01$ 에서 $p\text{-값} > \alpha$ 이므로 귀무가설 H_0 가 기각되지 않는다.

[5] 10개의 표본자료는 모평균이 수평측 거리변화가 20임을 증거 한다.

[R-코드 1.2.2] [보기 1.2.2] **소표본**의 모평균에 대한 t -검정(상악전치부변형- t -검정.R)

```
## [보기 1.2.2] S-perpendicular to the PNS  
x=c(23.46, 20.94, 20.28, 19.55, 18.05, 18.03,  
23.39, 16.84, 12.47, 28.80)  
list(mean(x),sd(x))  
t.test(x, mu=20, alternative="greater")
```

참고:

- $H_1 : \mu < 20$: alternative = "less"
- $H_1 : \mu \neq 20$: alternative = "two.sided"

$H_0 : \mu = 20, H_1 : \mu > 20$

- t.test():

```
> list(mean(x),sd(x))  
[[1]]  
[1] 20.181
```

```
[[2]]  
[1] 4.423554
```

```
> t.test(x, mu=20, alternative="greater")
```

One Sample t-test

```
data: x  
t = 0.1294, df = 9, p-value = 0.4499  
alternative hypothesis: true mean is greater  
than 20  
95 percent confidence interval:  
17.61675      Inf  
sample estimates:  
mean of x  
20.181
```

[보기 1.2.3] 치아 합병증 발생률

치관주위염 있는 환자들 중 10%가 발치 후 9가지 합병증(출혈, 인접치 손상, 신경손상, 개구장애, 심한 통증 및 부종, 건성 발치, 감염, 발치 중단, 기타)의 발생 빈도가 한 가지 이상으로 알려져 있다. 이 비율이 일반적인지를 보기 위하여, 200명의 환자를 조사한 결과 25명이 합병증이 발생 하였다. 이 표본 자료에 의하면 현재 이 환자 비율이 알려진 것과 달라졌다고 할 수 있는지를 살펴보자.

[1] 모비율 p 를 합병증 발생 비율이라 하면 p 가 일반적으로 알려진 비율 0.10과 다른지에 대한 증거를 찾고 있으므로, 가설은 다음과 같다.

$$H_0 : p = 0.10, H_1 : p \neq 0.10.$$

[2] 표본의 크기 $n = 200$ 은 충분히 크므로 대표본에 해당하며 모비율 p 의 추정치는 $\hat{p} = 25/200$ 이고 $H_0 : p = 0.10$ 에서 검정통계량의 관측값은 다음과 같다.

$$z = \frac{(25/200) - 0.1}{\sqrt{0.1 \times 0.9/200}} = \frac{0.125 - 0.1}{0.021} = 1.179$$

[3] 대립가설 $H_1 : p \neq 0.10$ 의 형태에 따른 유의확률:

$$p\text{-값} = P[|Z| \geq 1.179] = 2P[Z \geq 1.179] = 0.239$$

[4] 유의수준 $\alpha = 0.05$ 에서 $p\text{-값} > \alpha$ 이므로 귀무가설 H_0 가 기각되지 않는다.

[5] 양측검정의 결과 일반적으로 알려진 합병증 발생 비율이 달라졌다고는 할 수 없다.

[R-코드 1.2.3] [보기 1.2.3] 대표본의 모비율에 대한 검정(치아합병증발생률-Z-검정.R)

대표본 모비율 검정 : [보기 1.2.3] $H_0 : p = 0.10, H_1 : p \neq 0.10$

```
# Z-검정
n=200
p0=0.10
p=25/n
z=(p-p0)/sqrt(p0*(1-p0)/n)
z
pvalue=2*(1-pnorm(z))
list(z, pvalue)

# Chiq-square test: prop.test()
prop.test(25, 200, p=0.1,
alternative="two.sided", correct=F)
```

참고 : Yates'의 연속보정: correct=TRUE

$$X^2 = Z^2 = \left[\frac{|\hat{p} - p_0| - \frac{1}{2n}}{\sqrt{p_0 q_0 / n}} \right]^2$$

```
>list(z, pvalue)
[[1]]
[1] 1.178511
[[2]]
[1] 0.2385928

>prop.test(25, 200, p=0.1,alternative="two.sided", correct=F)
      1-sample proportions test without continuity correction

data:  25 out of 200, null probability 0.1
X-squared = 1.3889, df = 1, p-value = 0.2386
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
 0.08611974 0.17801425
sample estimates:
      p
0.125
```

1.3 두 집단의 비교: 독립표본, 대응표본, 모비율

소표본과 대표본 독립표본 모평균 차의 검정 절차

표본의 크기	소표본($n_1, n_2 \leq 30$)	대표본($n_1, n_2 > 30$)
모집단의 정규성 가정	$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$	$X \sim (\mu_1, \sigma_1^2), Y \sim (\mu_2, \sigma_2^2)$
모르는 모분산 가정	$\sigma_1^2 = \sigma_2^2 = \sigma^2$	$\sigma_1^2 \neq \sigma_2^2$
검정통계량	$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$ $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} : \text{합동표본분산}$	$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$ $s_1^2, s_2^2 : \text{표본분산}$
귀무가설	$H_o : \mu_1 - \mu_2 = 0$	
대립가설	$H_1 : \mu_1 - \mu_2 > 0, H_1 : \mu_1 - \mu_2 < 0, H_1 : \mu_1 - \mu_2 \neq 0.$	
p-값	$P(T \geq t), P(T \leq t), P(T \geq t)$	$P(Z \geq z), P(Z \leq z), P(Z \geq z)$
검정방법	$p\text{-값} < \alpha$ 이면 귀무가설 $H_o : \mu_1 - \mu_2 = 0$ 를 기각한다.	

대응표본(소표본)의 검정 절차

	처리 대상					
	1	2	...	i	...	n
처리 1	X_1	X_2	...	X_i	...	X_n
처리 2	Y_1	Y_2	...	Y_i	...	Y_n
차	$D_1 = X_1 - Y_1$	$D_2 = X_2 - Y_2$	...	$D_i = X_i - Y_i$	...	$D_n = X_n - Y_n$
통계량	새로운 자료: $D_1, D_2, \dots, D_n$ 표본평균: $\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i$, 표본분산: $s_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$					
검정 통계량	$\text{소표본} : T = \frac{\bar{D}}{s_D / \sqrt{n}} \sim t_{(n-1)}, \quad \text{대표본} : Z = \frac{\bar{D}}{s_D / \sqrt{n}} \approx N(0,1)$					
귀무 가설	$H_0: \mu_D = 0$					
대립 가설	$H_o: \mu_D > 0, H_o: \mu_D < 0, H_o: \mu_D \neq 0$					
p -값	$P(T \geq t), P(T \leq t), P(T \geq t)$					

[보기 1.3.1] 소표본 남녀 발치 난이도 평균의 차이

매복치 발치 난이도 평균이 20대~30대 남녀 간에 같다고 알려져 있다. 그러나 새로운 연구 결과 난이도 평균이 서로 다르다고 하여 모분산이 같고 정규분포를 따르는 모집단으로부터 이를 토대로 유의수준 5%에서 남녀 집단 간의 난이도 평균의 차에 대한 검정을 해보라.

모집단	표본의 크기	난이도 평균	표준편차
남자 x	11	4.55	0.82
여자 y	11	3.55	0.82

[1] 남녀 집단의 모평균을 각각 μ_1 과 μ_2 라 하면 다음과 같은 가설을 세울 수 있다.

$$H_0: \mu_1 - \mu_2 = 0, \quad H_1: \mu_1 - \mu_2 \neq 0$$

[2] 각 집단의 표본의 크기 11은 소표본이고 모분산이 동일하며 정규분포를 따르므로 합동표본분산을 계산 :

$$s_p^2 = \frac{(11-1)0.673 + (11-1)0.673}{11+11-2} = 0.673$$

검정통계량의 관측값 계산 :

$$t = \frac{4.55 - 3.55}{0.820 \sqrt{\frac{1}{11} + \frac{1}{9}}} = 2.859$$

[3] 양측 대립가설 $H_1 : \mu_1 - \mu_2 \neq 0$ 이므로 자유도 $df = n_1 + n_2 - 2 = 20$ 에서 유의확률:

$$p\text{-값} = P(|T| \geq 2.859) = 2P(T \geq 2.859) = 0.0097$$

[4] 유의수준 $\alpha = 0.05$ 에서 $p\text{-값} < \alpha$ 이므로 귀무가설 $H_0 : \mu_1 - \mu_2 = 0$ 을 기각한다.

[5] 이 연구에 따르자면 남녀 간의 발치 난이도 평균은 차이가 있다고 할 수 있다.

[R-코드 1.3.1] [보기 1.3.1] 소표본의 두 집단의 모평균 t -검정(소표본남녀발치난이도-t-검정.R)

```
## [보기 1.3.1] 남녀 두 집단 간의 발치 난이도  
비교
```

```
x=c(3, 4, 5, 5, 3, 5, 5, 5, 5, 5)
```

```
y=c(3, 3, 3, 3, 5, 3, 3, 4, 3, 4, 5)
```

```
## 방법 1
```

```
#평균, 분산, 표준편차
```

```
mean(x);var(x);sd(x)
```

```
mean(y);var(y);sd(y)
```

```
#합동표본분산, t-검정통계량
```

```
n1=length(x)
```

```
n2=length(y)
```

```
sp2=((n1-1)*var(x) + (n2-1)*var(y))/(n1+n2-2)
```

```
sp=sqrt(sp2)
```

```
t=(mean(x)-mean(y))/(sp*sqrt(1/n1 +1/n2))
```

```
#p-value : 양측검정
```

```
pvalue=2*(1-pt(t, df=20))
```

```
list(sp2,sp,t, pvalue)
```

```
## 방법 2
```

```
# t-검정 : t.test()
```

```
t.test(x, y, var.equal=T,
```

```
alternative="two.sided")
```

$H_0 : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 \neq 0$

```
> list(sp2, sp, t, pvalue)
```

```
[[1]]
```

```
[1] 0.6727273
```

```
[[2]]
```

```
[1] 0.8201995
```

```
[[3]]
```

```
[1] 2.859314
```

```
[[4]]
```

```
[1] 0.00969419
```

```
> t.test(x, y, var.equal=T, alternative="two.sided")
```

Two Sample t-test

data: x and y

t = 2.8593, df = 20, p-value = 0.009694

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.2704671 1.7295329

[보기 2.3.2] 소표본 두 집단의 분산이 다른 경우

턱 관절염의 이환 전후 편측 관절염(OAU)와 양측 관절염(OAB) 두 집단의 좌우(horizontal)방향의 하악턱 중앙부(symphysis)의 이동량이 같다는 일반적인 사실과는 달리 집단 OAU의 이동량이 집단 OAB의 이동량보다 더 크다고 할 수 있는지를 검정하였다. 만약에 이들 두 집단의 표준편차들의 상대적 크기가 $2.34/0.84=2.79$ 로 다음과 같다면 두 모분산이 같다는 $\sigma_1^2=\sigma_2^2$ 가정을 할 수 없게 된다. 이럴 경우 유의수준 5%에서 [보기 3.2.2]와 동일한 가설에 대하여 검정하여라.

모집단	자료 크기	평균	표준편차
OAU X	$n_1 = 20$	1.04	2.34
OAB Y	$n_2 = 16$	0.63	0.84

[1] 모집단 X 와 Y 의 하악턱 중앙부 이동량 모평균을 각각 μ_1 과 μ_2 라 하면 다음과 같은 가설을 세울 수 있다.

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$$

[2] 두 표본은 소표본에 해당하며 두 표준편차의 비가 $2.34/0.84=2.79$ 로 두 모평균이 같다고 보기 어렵다. 따라서 수정된 자유도는 다음과 같이 계산된다.

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{1}{n_1 - 1} \left[\frac{s_1^2}{n_1} \right]^2 + \frac{1}{n_2 - 1} \left[\frac{s_2^2}{n_2} \right]^2} = \frac{\left[\frac{2.34^2}{20} + \frac{0.84^2}{16} \right]^2}{\frac{1}{19} \left[\frac{2.34^2}{20} \right]^2 + \frac{1}{15} \left[\frac{0.84^2}{16} \right]^2} = 24.799$$

검정통계량의 관측값:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1.04 - 0.63}{\sqrt{\frac{2.34^2}{20} + \frac{0.63^2}{16}}} = 0.727$$

[3] 대립가설 $H_1 : \mu_1 - \mu_2 > 0$ 에서 자유도 $df = 24.899$ 에서 유의확률:

$$p\text{-값} = P(T \geq 0.727) = 0.237$$

[4] 유의수준 $\alpha = 0.05$ 에서 $p\text{-값} > \alpha$ 이므로 귀무가설 $H_0 : \mu_1 - \mu_2 = 0$ 을 채택

[5] 집단 X 의 하악턱 중앙부 이동량이 집단 Y 의 이동량보다 더 많다고 할 수 없다.

[R-코드 1.3.2] [보기 1.3.2] 소표본의 두 집단(분산이 다름)의 모평균 t -검정(소표본불등분산터관절
염이동량비교-t-검정.R)

```
## [보기 1.3.2] 편측 관절염(OAU)과 양측  
관절염(OAB) 집단의 비교 : 분산이 다름  
#평균,표준편차  
mx=1.04;sx=2.34  
my=0.63;sy=0.84
```

```
# t-검정통계량 : 소표본 분산이 다를때  
n1=20  
n2=16  
#t-test : 소표본  
t=(mx-my)/sqrt(sx^2/n1 + sy^2/n2)  
wdf=((sx^2/n1 +  
sy^2/n2)^2)/((sx^2/n1)^2/(n1-1) +  
(sy^2/n2)^2/(n2-1))
```

```
#p-value : 단측검정  
pvalue1=1-pt(t, df=wdf)
```

```
#p-value : df=min(n1-1, n2-1)  
pvalue2=1-pt(t, df=15)
```

```
list(t, wdf, pvalue1, pvalue2)
```

$$H_o : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 > 0$$

```
> list(t, wdf, pvalue1, pvalue2)
```

```
[[1]]
```

```
[1] 0.7271973
```

```
[[2]]
```

```
[1] 24.79893
```

```
[[3]]
```

```
[1] 0.2369574
```

```
[[4]]
```

```
[1] 0.2391511
```

[보 기 1.3.3] 주걱턱 수술 전과 후의 변화

골격성 Class III 기형((하악이 나온 보통 주걱턱)을 양악수술한 후 상악(윗턱) 전치부 변형 중 수평측 거리 변화인 S-perpendicular to the PNS의 10명 환자 수술 전 x 와 수술 후 y 의 거리 변화를 살펴보기 위해 x 와 y 간에 차이가 있다고 할 수 있는지를 유의수준 5%에서 검정하라.

수술	처리 대상									
전 x	23.46	17.26	17.32	20.94	19.72	17.44	20.28	20.84	19.55	19.37
후 y	28.96	17.70	18.69	21.96	20.40	18.91	22.72	20.86	20.57	19.43
차 D	-5.50	-0.44	-1.37	-1.02	-0.68	-1.47	-2.44	-0.02	-1.02	-0.06

[1] 수술 전 X 와 수술 후 Y 의 S-perpendicular to the PNS의 거리 변화 차이의 모평균 μ_D 에 대한 귀무가설과 대립가설:

$$H_0 : \mu_D = 0, H_1 : \mu_D \neq 0$$

[2] 대응표본의 크기 10은 소표본:

$$\bar{D} = \frac{1}{10} \sum_{i=1}^n D_i = -1.403, \quad s_D^2 = \frac{1}{10-1} \sum_{i=1}^n (D_i - \bar{D})^2 = 2.590$$

[3] 귀무가설 $H_0 : \mu_D = 0$ 하에 검정통계량의 관측값:

$$t = \frac{-1.422}{1.609/\sqrt{10}} = -2.757$$

[4] 대립가설 $H_1 : \mu_D \neq 0$ 에 대한 자유도 $df = n - 1 = 9$ 에서 유의확률:

$$p\text{-값} = 2P(T \leq -2.757) = 0.022$$

[5] 유의수준 $\alpha = 0.05$ 에서 $p\text{-값} < \alpha$ 이므로 귀무가설 $H_0 : \mu_D = 0$ 을 기각한다.

[6] 이 연구에 따르자면 수술 전과 후 거리 변화 차이는 있다고 할 수 있다.

[R-코드 1.3.3] [보기 1.3.3] 대응표본에 대한 검정 t -검정(소표본상악전치부변형대응표본- t -검정.R)

```
## [보기 1.3.3] 소표본 대응자료의 비교
X<-c(23.46, 17.26, 17.32, 20.94, 19.71,
17.44, 20.28, 20.84, 19.55, 19.37)
Y<-c(28.96, 17.70, 18.69, 21.96, 20.40,
18.91, 22.72, 20.86, 20.57, 19.43)
D=X-Y
```

방법 1

```
## t-test
n=length(D)
m=mean(D)
s=sd(D)
t=m/(s/sqrt(n))
```

```
#p-value : 양측검정
pvalue=2*pt(t, df=n-1)
list(t, pvalue)
```

방법 2

```
# t-검정 : t.test()
t.test(X, Y, paired=T,
alternative="two.sided")
```

$H_0 : \mu_D = 0, H_1 : \mu_D \neq 0$

```
> list(t, pvalue)
```

```
[[1]]
```

```
[1] -2.756709
```

```
[[2]]
```

```
[1] 0.0222315
```

```
> t.test(X, Y, paired=T, alternative="two.sided")
```

Paired t-test

data: X and Y

t = -2.7567, df = 9, p-value = 0.02223

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.5543027 -0.2516973

sample estimates:

mean of the differences

-1.403

독립표본의 모비율 비교 검정 단계 : 독립표본의 크기가 n_1 과 n_2 가 대표본인 경우

[1단계] H_o : 두 모집단의 비율(모비율)은 차이가 없다. $\Leftrightarrow H_o : p_1 - p_2 = 0$.

[2단계] $H_o : p_1 - p_2 = 0$ 하에 합동 추정치 계산 :

$$\hat{p} = \frac{X + Y}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

[3단계] 검정통계량 :
$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

[4단계] 유의수준 α 에서 Z -검정

귀무가설	대립가설	기각역	p -값
	$H_1 : p_1 - p_2 > 0$	$z \geq z_\alpha$	$P(Z \geq z)$
$H_o : p_1 - p_2 = 0$	$H_1 : p_1 - p_2 < 0$	$z \leq -z_\alpha$	$P(Z \leq z)$
	$H_1 : p_1 - p_2 \neq 0$	$ z \geq z_{\alpha/2}$	$P(Z \geq z)$

참고 :

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \simeq N(0,1) \Rightarrow Z^2 \simeq X_{df=1}^2$$

[보기 1.3.4] 독감 예방 백신에 의한 남녀 항체율 비교

60세에서 65세 노인: 독감 예방 백신 접종 항체의 생성 비율이 남성 노인보다 여성 노인 쪽이 더 높다고 할 수 있는지를 유의수준 5%에서 검정하고 p -값에 의한 결과와 비교하라.

모집단	표본의 크기	항체 생성
남자	113	$X = 34$
여자	139	$Y = 54$

[1] 남성 노인과 여성 노인의 항체 생성 모비율을 각각 p_1 과 p_2 :

$$H_0: p_1 - p_2 = 0, \quad H_1: p_1 - p_2 < 0.$$

[2] $n_1 = 113$, $n_2 = 139$: 대표본 $\Rightarrow$ 합동 추정치: $\hat{p} = \frac{34+54}{113+139} = 0.349$

[3] 검정통계량의 관측값 : $z = \frac{0.301 - 0.388}{\sqrt{0.349 \times (1 - 0.349)}} \sqrt{\frac{1}{113} + \frac{1}{139}} = -1.45$

[4] 유의수준 5% = 0.05에서 기각역은 $z \leq -1.64$ 이다. $z = -1.45$ 는 기각역을 벗어나므로 귀무가설 $H_0: p_1 - p_2 = 0$ 을 기각하지 못하게 된다.

[5] 여성의 항체 생성률이 남성의 항체 생성률 보다 높다고 할 수 없다.

[R-코드 1.3.4] [보기 1.3.4] 대응표본에 대한 t -검정(독감백신-t-검정.R)

```
## 독감 예방 백신에 의한 항체율 비교
## 방법 1
z.prop = function(x1,x2,n1,n2){
  numerator = (x1/n1) - (x2/n2)
  p.common = (x1+x2) / (n1+n2)
  denominator = sqrt(p.common * (1-p.common)
    * (1/n1 + 1/n2))
  z.prop.ris = numerator / denominator
  return(z.prop.ris)}
n1 <- 113
n2 <- 139
x <- 34
y <- 54
z=z.prop(x, y, n1, n2)
z
pvalue1=pnorm(z)
pvalue1

## 방법 2
# Chisquare test
chi=z^2
pvalue2=(1-pchisq(chi, df=1))/2
list(chi, pvalue2)

## 방법 3
# prop.test()
prop.test(x=c(34, 54), n=c(113,
139),alternative="less", correct=F)
```

$$H_0 : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0$$

```
> list(z, pvalue1)
[[1]]
[1] -1.450804
```

```
[[2]]
[1] 0.07341719
```

```
> list(chi, pvalue2)
[[1]]
[1] 2.104833
```

```
[[2]]
[1] 0.07341719
```

```
> prop.test(x=c(34, 54), n=c(113, 139),alternative="less", correct=F)
```

2-sample test for equality of proportions without continuity correction

```
data:  c(34, 54) out of c(113, 139)
X-squared = 2.1048, df = 1, p-value = 0.07342
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01068364
sample estimates:
   prop 1   prop 2 
0.3008850 0.3884892
```

1.4 여러 집단의 비교 : 일원 분산분석

일원 분산분석을 적용할 g 개 처리(그룹)의 독립표본의 구조				
처 리(그룹)	1	2	...	g
자 료	y_{11}	y_{21}	...	y_{g1}
	y_{12}	y_{22}	...	y_{g2}
	$\vdots$	$\vdots$		$\vdots$
	y_{1n_1}	y_{2n_2}	...	y_{gn_g}
평 균	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_g$
총평균	$\bar{y} = \frac{n_1\bar{y}_1 + \cdots + n_g\bar{y}_g}{n}$			

일원 분산분석 모형 : $y_{ij} = \mu_i + e_{ij}, j = 1, \dots, n_i, i = 1, \dots, g$

μ_i : i 번째 처리에 대한 효과(effect)로 평균 반응

e_{ij} : 서로 독립이며 정규분포 $N(0, \sigma^2)$ 을 따르는 오차항

- 귀무가설 : $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$

- 변동의 분해

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^g n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$\Leftrightarrow TSS = SST + SSE$$

일원 분산분석표

요 인 (Source)	자유도 (df)	제곱합 (Sum of Squares)	평균제곱합 (Mean Square)	분산비 (F value)	p-값
처 리 (Model)	$g - 1$	SST	$MST = SST / (g - 1)$	$F = \frac{MST}{MSE}$	$P[F \geq f]$
오 차 (Error)	$n - g$	SSE	$MSE = SSE / (n - g)$		
전 체 (Total)	$n - 1$	TSS			

$$- F = \frac{MST}{MSE} = \frac{SST / (g - 1)}{SSE / (n - g)} \sim F_{(g-1, n-g)}$$

[보기 1.4.1] 4개 회사 칫솔의 효과

16명의 치과병원 환자를 4명씩 임의추출하여 4개 회사의 칫솔을 사용하게 하여 치태지수를 조사하였다(임희정, 2008, 4.3절). 4개 회사 칫솔 효과에 대한 차이를 유의수준 5%에서 검정하라.

칫솔회사	A	B	C	D
자 료	0.46	0.73	0.99	0.66
	0.74	0.54	0.78	0.49
	0.26	0.47	0.63	0.23
	0.55	0.59	0.81	0.09
평 균	0.50	0.58	0.80	0.37

[1] 4개 회사의 칫솔 A, B, C, D는 처리가 $g=4$ 개임을 말하며 이들의 효과가 같다는 다음의 귀무가설을 세울 수 있다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

[2] $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ 을 F -검정하기 위한 일원 분산분석표는 다음과 같다.

요 인	자유도	제곱합	평균제곱합	분산비	p -값
처 리	3	0.3985	0.1328	3.8177	0.039
오 차	12	0.4175	0.0348		
전 체	15	0.8160			

[3] 분산분석표로부터 검정통계량 관측값 $f = 3.8177$ 은 자유도 (3, 12)을 따르며 유의확률이 0.039로 유의수준 0.05에서 귀무가설을 기각한다.

[4] 분산분석의 결과 4개 회사 칫솔의 효능에는 차이가 있다고 볼 수 있다. 실제 평균을 보면 D 회사의 칫솔이 비교적 치태를 제거하는 데 성능이 우수함을 알 수 있다.

[R-코드 1.4.1] [보기 1.4.1] 4개 회사 치태지수 비교(4개회사치솔사용후치태지수비교-anova.R)

```
## [보기 4.2.2] 4개 회사 치솔 사용 후  
치태지수 비교
```

```
A<-c(0.46,0.74,0.26,0.55)
```

```
B<-c(0.73,0.54,0.47,0.59)
```

```
C<-c(0.99,0.78,0.63,0.81)
```

```
D<-c(0.66,0.49,0.23,0.09)
```

```
data<-c(A, B, C, D)
```

```
# 세 집단에 의한 자료
```

```
group<-c(rep("A",4),rep("B",4),rep("C",  
4),rep("D",4))
```

```
data<-data.frame(data,group)
```

```
# 일원 분산분석
```

```
fit<-anova(lm(data~group,data=data)  
);fit
```

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

```
> fit<-anova(lm(data~group,data=data));fit
```

Analysis of Variance Table

Response: data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	3	0.39848	0.132825	3.8177	0.03937 *
Residuals	12	0.41750	0.034792		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

1.5 범주들의 관계 비교 : 독립성, 동질성 검정

[보기 1.5.1] 독립성 검정

어느 제약회사에서 489명의 고객들이 어떤 종류의 약을 선호하는지를 알기 위해 설문조사를 실시하여 고객들의 나이와 먹는 약의 종류라는 2개의 범주형 변수에 따라 자료를 표로 정리하였다. 나이는 20세 이상 30세 미만, 30세 이상 50세 미만, 50세 이상인 세 범주로 구성되어 있고 먹는 약의 종류는 캡슐형과 정제형인 두 범주로 구성되어 있다.

나 이	먹는 약의 종류		합 계
	캡슐형	정제형	
20세 이상 30세 미만	38	79	117
30세 이상 50세 미만	87	118	205
50세 이상	78	89	167
합 계	203	286	489

[보기 1.5.2] 동질성 검정

두 가지 식이요법 A와 B의 효과를 비교하기 위해 150명의 환자를 대상으로 임의 추출된 80명에게는 식이요법 A를 나머지 70명에게는 식이요법 B를 적용하였다. 얼마간의 시간이 흐른 뒤에 식이요법과 각 환자의 건강상태인 2개의 범주형 변수에 따라 자료를 표로 정리하였다. 식이요법은 범주 A와 B로, 환자의 건강상태는 양호, 보통, 불량으로 세 범주로 구성되어 있다.

식이요법	환자의 건강상태			합 계
	양 호	보 통	불 량	
A	37	24	19	80
B	17	33	20	70
합 계	54	57	39	150

$r \times c$ 분할표의 행과 열의 독립성과 동질성 검정

- 피어슨의 카이제곱 검정 통계량 :

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(\text{관찰도수} - \text{기대도수})^2}{\text{기대도수}} = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

$$\text{여기서 } E_{ij} = \frac{i\text{번째 행합계} \times j\text{번째 열합계}}{\text{총합계}} = \frac{O_{i+} \times O_{+j}}{N}$$

- 자유도 : $df = (r-1)(c-1)$

- 유의확률 : $p\text{-값} = P[X^2 > \chi^2] < \alpha \Rightarrow$ 귀무가설 $\langle H_0 : \text{서로 독립(서로 동질)} \rangle$ 을
유의수준 α 에서 기각한다.

[R-코드 1.5.1] [보기 1.5.1] 고객들의 나이와 약의 종류에 대한 독립성 검정(나이약종류-독립성 검정.R)

```
row1 = c(38,79)
row2 = c(87,118)
row3 = c(78,89)
data.table = rbind(row1,row2,row3)
chisq.test(data.table)
```

H_0 : 먹는 약의 종류의 선호도와 나이는 관련이 없다

Pearson's Chi-squared test

data: data.table

X-squared = 5.8608, df = 2, p-value = 0.05338

[R-코드 1.5.2] [보기 1.5.2] 두 식이요법 간의 동질성 검정(식이요법-동질성검정.R)

```
row1 = c(37,24, 19)
row2 = c(17, 33, 20)

data.table = rbind(row1,row2)
chisq.test(data.table)
```

H_0 : 두 식이요법 간에는 차이가 없다

Pearson's Chi-squared test

data: data.table

X-squared = 8.224, df = 2, p-value = 0.01638