



Day32; 20221020

날짜
유형
태그

GitHub - u8yes/R

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

<https://github.com/u8yes/R>

u8yes/R



A 1 Contributor 0 Issues 1 Star 0 Forks

```
# 표준편차(Standard Deviation:SD) : 분산 값에 루트를 적용해서 제곱을 제거한 값. ex) 23.55(sqrt(554.56))

# 자유도(degree of freedom) : 표본의 분산과 표준편차를 계산할 때 나누는 분모의 수를 (모집단-1)개로 계산하여 주어진 데이터에서 표본을 자유롭게 뽑을수 있는 경우의수를

median(score1) # [1] 86
# 분산(Variance) : score1 <- c(85, 90, 93, 86, 82)
# ((85-86)^2+(90-86)^2+(93-86)^2+(86-86)^2+(82-86)^2)/5 = 16.4
var(score1)    # [1] 18.7
sd(score1)     # [1] 4.32435

# 평균의 종류
# 1) 산술평균 : 모든 값을 더한 후 값의 개수만큼 나눈 후 나오는 값을 의미.

# 2) 상승평균/기하평균 : %로 평균 비율을 구할 때 방법.
#   ex) 회사의 연매출 10억 인 회사가 작년에 10% 성장 후 올해 2% 감소했다면 2년 평균 성장을은 어떻게 될까요?
#   ans) <루트>sqrt(1.1*0.98) = 1.04 : 4% 성장.

# 3) 제곱평균 : 각 값의 제곱의 평균을 구한 후 루트를 적용해서 구하는 평균.

# 4) 조화평균 : 주로 평균 속도를 구할 때 사용하는 방법.
#   ex) 서울에서 강원도로 휴가는 가는데 갈 때는 안 막혀서 시속 100km로 갔는데, 올 때는 너무 막혀서 시속 60km였다면 왕복 평균 속력은 얼마일까요?
#   ans) 조화 평균의 식 : 2xy / (x+y) = 2(100*60) / (100+60)
```

cost 데이터의 이상치가 너무 커서

중위수, 평균값과 대조했을 때 min값과 max값 사이가 너무 크다.

```
> str(data$survey)
int [1:300] 1 2 1 4 3 3 NA NA NA 1 ...
> # 데이터 특성(최소, 최대, 중위수, 평균, 분위수, 노이즈-NA) 제공
```

```
> summary(data)
   resident      gender       age        level        cost
  Min. :1.000  Min. :0.00  Min. :40.00  Min. :1.000  Min. :-457.200
  1st Qu.:1.000 1st Qu.:1.00  1st Qu.:48.00  1st Qu.:1.000  1st Qu.: 4.425
  Median :2.000 Median :1.00  Median :53.00  Median :2.000 Median : 5.400
  Mean   :2.233 Mean   :1.42  Mean   :53.88  Mean   :1.836 Mean   : 8.752
  3rd Qu.:3.000 3rd Qu.:2.00  3rd Qu.:60.00  3rd Qu.:2.000  3rd Qu.: 6.300
  Max.   :5.000 Max.   :5.00  Max.   :69.00  Max.   :3.000 Max.   : 675.000
  NA's    :21      NA's    :13      NA's    :30      NA's    :30
```

x - 독립변수

y - 종속변수

범주형(discrete) - 명목척도, 서열척도(바플롯, 파이, 도트차트)

연속형(continuous) - 등간척도, 비율척도(박스플롯, 플롯, 히스토그램)

- **척도(Scale)**

- 변수에 값을 부여하는 방법
- 변수 측정 단위(응답자가 선택할 수 있는 질문 항목)

정성적-질적 척도(범주형 변수)		정량적-양적 척도(연속형 변수)	
명목척도	이름이나 범주를 대표하는 의미 없는 숫자 (예 : ① 남자 ② 여자)	등간척도	속성에 대한 각 수준 간의 간격이 동일한 경우(가감산 연산) (예: 연소득이 어디에 해당되십니까?)
서열척도	측정 대상 간의 높고 낮음(서열), 순서에 대한 값 부여 (예 : 좋아하는 순위를 표시하시오.)	비율척도	등간척도의 특성에 절대원점(0)이 존재하고, 비율계산이 가능한 경우(사칙연산) (예 : 나이가 몇 세 입니까?)

- 범주형 변수

서열척도는 숫자값 자체에 크고 작음, 높고 낮음의 의미가 담겨서 표현.

연산이 의미가 없음.

서열척도(Ordinal scale)

- 측정대상 간의 크고 작음, 양의 많고 적음, 선호도의 높고 낮음

단순히 식별자(남1여0 등으로)의 의미만을 담는 것은 **명목척도**

- 연속형 변수

비율 척도는 **절대원점(0)**이 존재,

숫자 자체만으로도 의미가 있음.

- 예) 성적, 키, 무게, 인구수, 수량, 길이, 금액 등

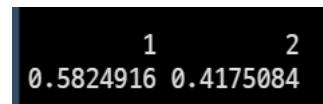
등간 척도는 연산이 의미가 있음. (매우 불만족 ~ 매우 만족 등) - 일정 비율을 유지하면서 등간 값 필요. 예) 1 3 5 7 9, 2 4 6 8 10 식으로 가야하는데 1 5 3 3 7 8 이런식은 좋지 않음.

qplot, ggplot - 3Layer로 시각화해줌

gmap, ggmap 지도를 시각화

척도

분석방법	적용분야	변수척도
빈도분석	가장 기초적이고 간단한 분석방법	모든 척도
교차분석 (카이제곱)	변수 간의 교차표 작성	명목척도, 서열척도
요인분석	• 타당성 검정 • 설명력 부족한 변수 제거	등간척도, 비율척도
신뢰도분석	추출된 요인들의 동질적인 변수 구성	등간척도, 비율척도
상관관계분석	측정변수들 간의 관계 정도를 제시	피어슨 - 등간척도, 비율척도 스피어만 - 서열척도
회귀분석	인과관계 분석	독립변수, 종속변수 : 등간척도/비율척도
t-검정	집단 간 평균 차이 검정	독립변수 : 명목척도 종속변수 : 등간척도 또는 비율척도
분산분석 (ANOVA)	3집단 이상의 평균 검정	독립변수 : 명목척도 종속변수 : 등간척도 또는 비율척도



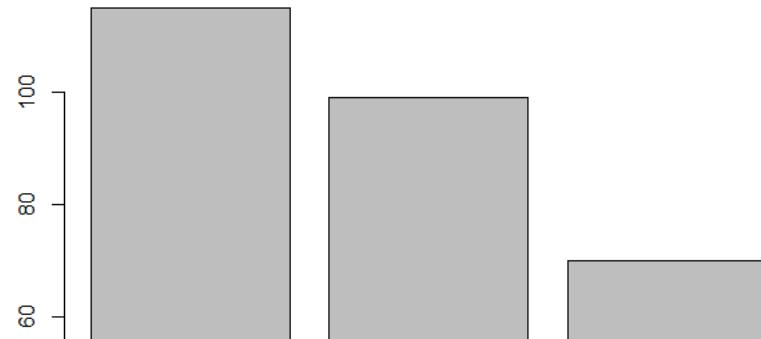
```
# 구성비율 계산
prop.table(x) # 비율계산: 0 < x < 1 사이의 값 # prop.table은 그냥 이름이다
#      1      2
#0.5824916 0.4175084
```

x11() 창을 띠운 상태에서 barplot()으로 시각화시킬 수 있음.

```
# 학력 수준(level) 변수의 빈도 수 시각화
x1 <- table(data$level) # 각 학력수준에 빈도수 저장.
x1
barplot(x1)
```

```
# 구성비율 계
y <- prop.tab
round(y*100,2)
#    1      2
#40.49 34.86

# 2.3 등간척도
# 만족도(surve
# (Untitled) ...
Terminal x Background J
6.2 · D:/heaven_dev/workspace/R/exam/level(data$level)
```

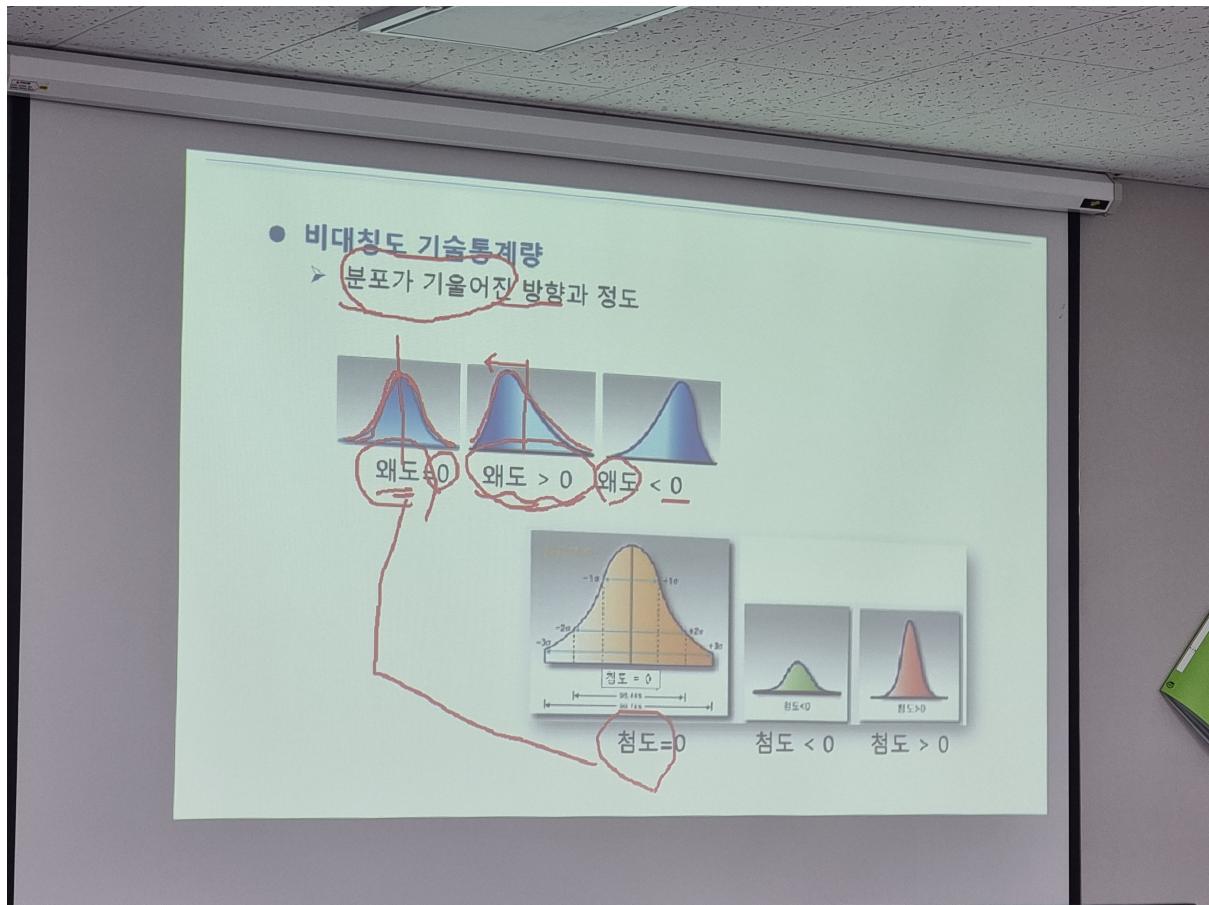


```
# 2.3 등간척도 기술 통계량
# 만족도(survey) 변수 대상 요약 통계량 구하기
survey <- data$survey
survey

summary(survey) # 만족도(5점척도)인 경우 의미 있음 -> 2.605(평균)
# Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
#1.000 2.000 3.000 2.605 3.000 5.000 112
x1 <- table(survey) # 빈도 수
x1
```

```
> survey
[1] 1 2 1 4 3 3 NA NA NA 1 2 2 2 2 NA NA NA NA NA NA 2 2 1 2 3
[27] 3 5 2 NA NA NA NA NA NA NA 2 2 3 4 3 2 2 3 4 5 4 2 NA 2
[53] 3 4 3 NA NA NA NA NA NA 3 3 3 2 2 3 3 NA NA NA 2 2 2 NA 2
[79] 2 3 NA NA 3 3 3 3 3 3 1 4 NA NA NA NA 4 3 3 4 NA NA NA NA 3
[105] 3 2 NA NA 3 NA 2 NA 2 2 5 2 NA 3 NA NA NA NA NA NA NA NA NA 2 2
[131] 4 3 4 3 3 NA NA 2 2 2 2 1 2 NA NA NA NA NA 3 3 3 3 4
[157] 3 NA 4 2 2 2 2 2 NA NA NA NA 3 3 2 NA 2 3 3 3 NA NA 3 4 3 4
[183] NA NA 3 3 4 2 1 2 4 3 3 2 5 2 2 2 1 2 4 NA 2 2 1 1 1
[209] 2 2 NA NA NA NA NA NA NA NA 2 3 4 5 3 3 4 NA 2 1 2 NA 1 2
[235] 2 1 2 2 NA NA 3 4 5 3 NA 3 4 4 5 2 2 3 NA NA 2 1 2 1 NA NA
[261] 2 3 NA 3 4 3 4 3 4 NA NA NA 2 1 2 NA NA NA NA NA 1 1 2 2 NA NA
[287] NA NA NA 2 1 2 3 NA NA NA NA
> summary(survey) # 만족도(5점척도)인 경우 의미 있음 -> 2.605(평균)
  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
  1.000 2.000 3.000 2.605 3.000 5.000 112
```

왜도가 왼쪽으로 쓰리면 음수, 왜도가 오른쪽으로 쓰리면 평균이 양수



예) 불수능이면 첨도가 크고 뾰족하다. 평균에 몰려있음.

교차분석과 chi-square 분석

교차분석
(카이제곱)

변수 간의 교차표 작성

명목척도, 서열척도

빈도분석결과에 대한 보충자료를 제시하는 데 효과적.

- 교차 분석에 사용되는 변수는 값이 10 미만인 범주형 변수여야 함.

feature가 너무 많으면 안 좋다.

```
# 2) 교차분할표 생성을 위한 패키지 설치
install.packages("gmodels")
library(gmodels)
```

```
# 3) 패키지를 이용한 교차 분할표 생성
CrossTable(x, y)

# 교차테이블에 카이제곱 적용
CrossTable(x, y, chisq = T)
#Pearson's Chi-squared test
#-----
#Chi^2 = 2.766951    d.f. = 2      p = 0.2507057
```

```
Cell Contents
|-----|
|           N |
| Chi-square contribution |
|       N / Row Total |
|       N / Col Total |
|       N / Table Total |
|-----|
```

Total Observations in Table: 225

x	y	실패	합격	Row Total
고졸		40	49	89
		0.544	0.363	
		0.449	0.551	0.396
		0.444	0.363	
		0.178	0.218	
대학		27	55	82
		1.026	0.684	
		0.329	0.671	0.364
		0.300	0.407	
		0.120	0.244	
대학원졸		23	31	54
		0.091	0.060	
		0.426	0.574	0.240
		0.256	0.230	
		0.102	0.138	
Column Total		90	135	225
		0.400	0.600	

Statistics for All Table Factors

```
Pearson's Chi-squared test
#-----
Chi^2 = 2.766951    d.f. = 2      p = 0.2507057 # d.f. = 2(자유도)
```

5.991(카이제곱 분포표) > 2.766951(카이제곱)

열: 0.05 유의수준(X2)

행: 자유도(DF) 2

DF	X2(.995)	X2(.99)	X2(.975)	X2(.95)	X2(.05)
1	0.000	0.000	0.001	0.004	3.841
2	0.010	0.020	0.051	0.103	5.991
3	0.072	0.115	0.216	0.352	7.815
4	0.207	0.297	0.484	0.711	9.488
5	0.412	0.554	0.831	1.145	11.071

X2 값이 5.991(카이제곱 분포표) 이상이면 귀무 가설을 기각할 수 있다는 의미

```
# 1) 일원카이제곱 검정
# (1) 적합성 검정
#-----
# 귀무가설(영가설): 기대치와 관찰치는 차이가 없다. : p >= 알파
```

```

#     예) 주사위는 게임에 적합하다.

# 연구가설(대립가설): 기대치와 관찰치는 차이가 있다. : p < 알파
#     예) 주사위는 게임에 적합하지 않다.
#-----

# 60회 주사위를 던져서 나온 관측도수/기대도수
# 기대도수 : 10, 10, 10, 10, 10, 10
# 관측도수 : 4, 6, 17, 16, 8, 9

chisq.test(c(4,6,17,16,8,9))
# X-squared = 14.2 > 11.07, df = 5, p-value = 0.01439

```

카이제곱 - 원래 기대확률대비 나온 시행확률대비가 유의미한지 보는 것.

- 일원 - 예) 주사위
- 이원 카이제곱 - 2개의 변수로 카이제곱검정을 해보려하는 것. 독립성 검증, 동질성 검증법.
동질성 검증 - 2 집단
독립성 검증 - 동일 집단

- 단일 집단간의 비율 검정·평균 검정
- 2 집단 간의 비율 검정·평균 검정
- 2 집단 유사성 대응되는 평균 검정법 (T-Test)
- 3 집단 이상 간의 비교하는 알고리즘 (분산 분석)

```

#####
# 추론통계학 분석 - 1-2. 단일집단 평균 검정(단일표본 T검정)
#####
# 방법 : 1개 집단의 평균과 어떤 특정한 값과 차이가 있는지 검증
# 작업절차
#   1. 실습파일 가져오기
#   2. 데이터 분포 및 결측치 제거(데이터 정제)
#   3. 정규분포 검정 : 모집단의 특성 반영 유무
#   4. 가설검정(모수/비모수) -> t.test()/wilcox.test()
#####

```

교차분석과 chi-square 분석

교차분석과 chi-square 분석

1) 교차표 작성/분석

- data.frame() 이용 교차표 작성
- package 이용 교차표 작성
- 교차표 분석(학력수준과 진학 여부 교차분석)

2) Chi-square 가설검정

- 교차분석/ Chi-square 보고서 작성법
 - ① 적합성 검정
 - ② 독립성 검정
 - ③ 동질성 검정

교차 분석(Cross Table Analyze)

- 범주형 자료(명목척도 또는 서열척도)를 대상으로 두 개 이상의 변수들에 대한 관련성 체크.
- 결합분포를 나타내는 교차 분할표를 작성.
- 변수 상호간의 관련성 여부를 분석하는 방법.
- 빈도분석의 특성별 차이를 분석하기 위해 수행하는 분석 방법.
- 빈도분석결과에 대한 보충자료를 제시하는 데 효과적.
- 빈도분석과 함께 고급 통계 분석의 기초 정보를 제공.

교차 분석 시 고려사항

- 교차 분석에 사용되는 변수는 값이 10 미만인 범주형 변수여야 함.
- 비율척도인 경우는 코딩변경(리코딩)을 통해서 범주형 자료로 변환하여 적용 가능.
 - ex) 나이: 10~19세는 1, 20~29세는 2, 30~39세는 3 ...

교차표 작성 / 분석

● data.frame() 이용 교차표 작성

```
setwd("c:/workspaces/Rwork/data")
data <- read.csv("cleanDescriptive.csv", header=TRUE)
data # 확인
head(data) # 변수 확인

x <- data$level2 # 학력수준 리코딩 변수
y <- data$pass2 # 대학진학 리코딩 변수

# 학력수준(독립변수) -> 진학여부(종속변수)
result <- data.frame(Level=x, Pass=y) # 데이터 프레임 생성 - 데이터 묶음

dim(result) # 차원보기 -> 248 2
table(result) # 교차표 보기
#          Pass
# Level    불합격 합격
# 고졸      40  49
# 대졸      27  55
# 대학원졸 23  31
```

교차표 작성 / 분석

● package 이용 교차표 작성

```
# 교차표 작성을 위한 패키지 설치
install.packages("gmodels")
library(gmodels) # CrossTable() 함수 사용

# diamonds 데이터 사용을 위한 ggplot2 패키지 설치
install.packages("ggplot2")
library(ggplot2)

# diamond의 cut과 color에 대한 교차표 생성
CrossTable(x=diamonds$color, y=diamonds$cut, chisq = TRUE)
```

교차표 작성 / 분석

● package 이용 교차표 작성

Total observations in Table: 53940

제목 없음 - 메모장							
		파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)					
diamonds\$color	diamonds\$cutter						
		Good	Very Good	Premium	Ideal	Row Total	
D	Fair	163 7.607 0.024 0.101 0.003	662 3.403 0.098 0.135 0.012	1513 0.014 0.223 0.125 0.028	1603 9.634 0.237 0.116 0.030	2834 5.972 0.418 0.132 0.053	6775
E	Fair	224 16.009 0.023 0.139 0.004	933 1.973 0.095 0.190 0.017	2400 19.258 0.245 0.195 0.044	2337 11.245 0.239 0.169 0.043	3903 0.032 0.398 0.181 0.072	9797
F	Fair	312 2.596 0.033 0.194 0.006	909 1.949 0.095 0.185 0.017	2164 0.333 0.227 0.179 0.040	2331 4.837 0.244 0.169 0.043	3826 0.049 0.401 0.178 0.071	9542
G	Fair	314 1.575 0.028 0.185 0.006	871 23.708 0.077 0.118 0.016	2239 20.968 0.204 0.190 0.043	2924 0.473 0.259 0.172 0.054	4884 30.745 0.433 0.209 0.081	11282
H	Fair	303 12.268 0.036 0.104 0.006	702 3.758 0.086 0.143 0.013	1824 0.697 0.251 0.151 0.034	2360 26.432 0.284 0.171 0.044	3115 12.390 0.375 0.155 0.058	8304
I	Fair	175 1.141 0.032 0.109 0.003	522 1.688 0.083 0.106 0.010	1204 0.090 0.223 0.100 0.022	1428 1.257 0.233 0.104 0.026	2093 2.479 0.316 0.087 0.039	5422
J	Fair	19 14.772 0.042 0.074 0.002	307 10.447 0.089 0.063 0.008	678 3.310 0.241 0.056 0.013	808 11.100 0.298 0.059 0.015	896 45.441 0.319 0.042 0.017	2808
Column Total		1510 0.030	4916 0.081	12082 0.224	19791 0.256	21551 0.400	53940



교차표 작성 / 분석

● 학력수준과 대학진학여부 교차분석(Package 이용)

학력수준(독립변수) : y -> 진학여부(종속변수) : x

학력수준이 대학 진학에 영향을 미친다.

x <- data\$level2 # 행 - 리코딩 변수 이용

y <- data\$pass2 # 열 - 리코딩 변수 이용

CrossTable(x,y) # x:학력수준, y:대학진학

교차표 작성 / 분석

● 부모의 학력수준과 자녀의 대학진학 여부

Total Observations in Table: 225

x	y	실패	합격	Row Total
고졸	40	49		89
	0.544	0.363		
	0.449	0.551	0.396	
	0.444	0.363		
	0.178	0.218		
대졸	27	55		82
	1.026	0.684		
	0.329	0.671	0.364	
	0.300	0.407		
	0.120	0.244		
대학원졸	23	31		54
	0.091	0.060		
	0.426	0.574	0.240	
	0.256	0.230		
	0.102	0.138		
Column Total	90	135		225
	0.400	0.600		

- 기대치 비율 예 (1행2열)
- 기대치 : $89(\text{행합}) * 135(\text{열합}) / 225(\text{전체합}) = 53.4$
- 기대치 비율 : $(49 - 53.4)^2 / 53.4 = 0.363$

관측치
기대치비율(χ^2)=(관측치-기대치) 2 /기대치

행비율

열비율

셀비율

관측치
(관측치-기대치) 2 /기대치
행비율
열비율
셀비율

관측치
(관측치-기대치) 2 /기대치
행비율
열비율
셀비율

전체 관측치
전체 열비율

교차표 작성 / 분석

❖ 논문에서 교차분석에 대한 해설 예

<교차분석 해설>

부모의 학력수준에 따른 자녀의 대학진학여부를 설문 조사한 결과 학력수준에 상관없이 대학진학 합격률이 평균 60%로 학력수준별로 유사한 결과가 나타났다. 전체 응답자 225명을 대상으로 고졸 39.6% (89명) 중 55.1%가 진학에 성공하였고, 대졸 36.4%(82명) 중 68.4%가 성공했으며, 대학원졸은 24%(54명) 중 57.4%가 대학진학에 성공하였다. 특히 대졸 부모의 대학진학 합격률이 평균보다 조금 높고, 고졸 부모의 대학진학 합격률이 평균보다 조금 낮은 것으로 분석된다.

chi-square 검정

● Chi-square 검정

- 교차 분석으로 얻어진 교차 분할표를 대상으로 유의 확률을 적용하여 변수 간의 독립성 및 관련성 여부 등을 검정하는 분석 방법.
- 범주(Category)별로 관측빈도와 기대빈도가 차이가 있는지 검정.
- 카이 제곱 분포에 기초한 통계적 방법(카이 제곱 분포표 이용).
- $\chi^2 = \sum (\text{관측값} - \text{기댓값})^2 / \text{기댓값}$.
- 분석을 위해서 교차 분할표 작성.
- 교차분석은 검정 통계량으로 카이 제곱 사용(=카이 제곱 검정).
- 검증 유형 분류 : 일원 카이 제곱 검정, 이원 카이 제곱 검정

chi-square 검정

1. 일원카이제곱 : 교차분할표 이용 안함([한 개 변인](#))
 - 적합성 검정 : 실제 표본이 내가 생각하는 분포와 같은가? 다른가?
예) 관찰도수가 기대도수와 일치하는지를 검정
2. 이원카이제곱 : 교차분할표 이용
 - 1) 독립성 검정 : [두 변인](#)는 서로 관련성이 있는가 없는가?
 - 한 모집단으로부터 하나의 표본이 추출된 경우
 - 예) 흡연량과 음주량 사이에 관련성이 있는가?
 - 귀무가설 : 흡연과 음주량은 관련성이 없다.(독립적이다.)
 - 2) 동일성 검정 : [두 집단](#)의 분포가 동일한가? 다른 분포인가?
 - 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법
 - [두 개 이상의 모집단에서 각 표본이 추출된 경우](#)
 - 귀무가설 : 집단 간의 비율이 동일하다.

chi-square 검정

● Chi-square 검정 절차

1. 가설을 설정한다.
2. 유의수준을 결정한다.
3. 기각값(카이제곱 분포표 참조)을 결정한다.

➤ 자유도(df)와 유의수준으로 기각값 결정

4. 관찰도수에 대한 기대도수를 구한다.
5. 검정통계량 χ^2 의 값을 구한다.
6. 귀무가설의 채택 또는 기각 여부를 판정한다.
7. 카이제곱 검정 결과를 설명한다.

chi-square 검정

● 카이제곱 분포표

CHI-SQUARE TABLE: VALUES OF CHI-SQUARE (ALPHA) OF THE CHI-SQUARE DISTRIBUTION									
DF	X ²								
	.995	.99	.975	.95	.05	.025	.01	.005	
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879	
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597	
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838	
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860	
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750	
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548	
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559	
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	

chi-square 검정

1. 일원카이제곱 검정 - 관측빈도와 기대빈도의 차이를 통해서 확률 모형이 주어진 자료를 얼마나 잘 설명하는지를 검정하는 통계적 방법.

- (1) 적합성 검정 - chisq.test() 이용

- 귀무가설 : 기대치와 관찰치는 차이가 없다.

- 예) 도박사의 주사위는 게임에 적합하다.

- 대립가설 : 기대치와 관찰치는 차이가 있다.

- 예) 도박사의 주사위는 게임에 적합하지 않다.

- # 주사위의 관찰치가 기대치와 차이가 있는가? 또는 없는가?

- # 60회 주사위를 던져서 나온 관측도수/기대도수

- # 관측도수 : 4, 6, 17, 16, 8, 9

- # 기대도수 : 10, 10, 10, 10, 10, 10

- chisq.test(c(4,6,17,16,8,9)) # p-value = 0.01439

- # 해설 : 주사위는 게임에 적합하지 않다.

chi-square 검정

- p값 해석 방법

- <해설> p값이 0.05미만이기 때문에 유의미한 수준에서 귀무 가설을 기각할 수 있다.

- 따라서 '도박사의 주사위는 게임에 적합하지 않다.'라는 대립가설을 채택한다.

- (귀무 가설 기각, 대립가설 채택)

- 유의수준과 유의확률

- # 유의수준(Confidence level) : 0.05(100개 중 5개(100*0.05) 허용 기준치(허용 오차))

- # 유의확률 : p-value 귀무 가설이 나올 수 있는 확률

- # p-value < 0.05 경우 : 유의확률은 유의수준 보다 적다.(귀무 가설 기각)

- 검정통계량 해석 방법

- # 검정통계량 : X-squared = 14.2, df = 5

- # 자유도(df) : 관측치가 n 인 경우 df = n - 1

- # 자유도(degree of freedom)란 검정을 위해서 n개의 표본(관측치)를 선정할 경우

- # n번째 표본은 나머지 표본이 정해지면 자동으로 결정되는 변인의 수를 의미

- # 자유도(df) 5인 경우, X-squared 기각값(역) : $\chi^2 >= 11.071$ (chi-square 분포표 참고)

- # χ^2 값이 11.071 이상이면 귀무 가설을 기각할 수 있다는 의미

chi-square 검정

(2) 선호도 분석

귀무가설 : 기대치와 관찰치는 차이가 없다.

예) 맥주의 선호도에 차이가 없다.

대립가설 : 기대치와 관찰치는 차이가 있다.

예) 맥주의 선호도에 차이가 있다.

```
data <- textConnection(  
  "맥주종류 관측도수  
  1 12  
  2 30  
  3 15  
  4 7  
  5 16")  
x <- read.table(data, header=T)
```

```
chisq.test(x$관측도수) # X-squared = 18.375, p-value = 0.001042  
# 해설 : 맥주의 선호도에 차이가 있다.
```

chi-square 검정

● 선호도 분석 결과

▪ 검정통계량 :

X-squared = 18.375, df = 4

▪ p-value 해석 :

p값이 0.05미만이기 때문에 유의미한 수준에서 귀무가설을

기각할 수 있다. 따라서 '맥주의 선호도에 차이가 있다.'라는

대립가설을 채택할 수 있다. (귀무가설 기각, 대립가설 채택)

chi-square 검정

2. 이원카이제곱 검정

1) 독립성 검정(관련성 검정) - 교차테이블 이용

- 동일 집단의 두 변인을 대상으로 관련성이 있는가? 없는가?를 검정하는 방법.

예) 귀무가설 : 부모의 학력수준과 자녀의 대학진학 여부와 관련성이 없다.

- 두 변인은 독립적이다.

예) 대립가설 : 부모의 학력수준과 자녀의 대학진학 여부와 관련성이 있다.

- 두 변인은 독립적이지 않다.

CrossTable(x, y, chisq = TRUE) # p = 0.2507057

chi-square 검정

● 독립성 검정(관련성 검정) 결과

x	y	실패	합격	Row Total
고졸	40	49	89	
	0.544	0.363		0.396
	0.449	0.551		
	0.444	0.363		0.218
대학	27	55	82	
	1.026	0.684		0.364
	0.329	0.671		
	0.300	0.407		0.244
	0.120			
대학원졸	23	31	54	
	0.091	0.060		0.240
	0.426	0.574		
	0.256	0.230		0.138
	0.102			
Column Total	90	135	225	
	0.400	0.600		

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 2.766951 d.f. = 2 p = 0.2507057 |

<검정 결과 해설>

✓ $\text{Chi}^2 = \sum [(\text{관측값} - \text{기댓값})^2 / \text{기댓값}]$
✓ d.f. = (행수-1)*(열-1) = (3-1)*(2-1) = 2
-> 두 값만 구하면 나머지는 저절로 구해진다.
✓ p = 유의수준 : 0.05이하이면 귀무가설 기각

자유도에 따른 Chi^2 분포도
-> 자유도가 클수록 정규분포에 가까워진다.
유의수준 0.05에서,
-> 자유도 : 2인 경우, 기각역 : $\text{x}^2 \geq 5.99$,
-> 자유도 : 6인 경우, 기각역 : $\text{x}^2 \geq 12.59$

자유도가 2인 경우 x^2 값이 5.99이상이면
귀무가설 기각(카이제곱 분포표 참조)

해설 : Chi^2 값이 5.99 이하이고, 유의수준이
0.05 이상으로 분석되어 귀무가설을 기각할 수
없다. 따라서 부모의 학력수준과 자녀의 대학
진학 변인 간의 관련성은 없는 것으로 분석된다.

chi-square 검정

❖논문에서 교차분석표와 Chi-square 검정에 대한 해설 예

<교차분석표와 카이제곱 검정결과 해설>-----

'부모의 생활수준과 자녀의 대학진학 여부와 관련성이 있다.'를 분석하기 위해서 자녀를 둔 A회사 225명의 부모를 표본으로 추출한 후 설문 조사하여 교차분석과 카이제곱 검정을 실시하였다.

분석 결과를 살펴보면 부모의 생활수준과 자녀의 대학진학 여부의 관련성은 유의미한 수준에서 차이가 없는 것으로 나타났다.($\chi^2=2.767$, $p>0.05$)

따라서 귀무가설을 기각할 수 없다. 다음 <표>에서 부모의 생활 수준과 자녀의 대학 진학 여부에 대한 교차표와 카이제곱 검정결과를 제시하고 있다.

chi-square 검정

<논문에서 카이제곱 검정 결과 제시방법>

카이제곱 검정결과를 논문에서 제시할 경우 교차표와 카이제곱 검정통계량 함께 제시

학력수준		실패	진학	X-squared	유의확률(p)
고졸	관찰빈도	40	49	2.766951	0.2507057
	기대빈도	36	54		
대졸	관찰빈도	27	55	2.766951	0.2507057
	기대빈도	33	49		
대학원졸	관찰빈도	23	31	2.766951	0.2507057
	기대빈도	21	32		

chi-square 검정

<실습> 교육수준과 흡연율 간의 관련성 분석

1. 파일 가져오기

```
setwd("c:/workspaces/Rwork/data")
smoke <- read.csv("smoke.csv", header=TRUE)
# 변수 보기
head(smoke) # education, smoking 변수
names(smoke)
[1] "education" "smoking"
```

● 변수 모델링

객체를 대상으로 분석할 속성(변수)을 선택하여 속성 간의 관계 설정 과정

예) smoke 객체에서 education, smoking 속성을 분석대상으로 하여 교육수준이 흡연율과 관련성이 있는가를 education -> smoking 형태로 기술한다.
education은 영향을 미치는 변수로 독립변수라 하며, 영향을 받는 smoking은 종속변수라고 한다.

chi-square 검정

2. 코딩 변경 - 변수 리코딩 <- 가독성 제공

education(독립변수) : 1:대졸, 2:고졸, 3:중졸

smoke(종속변수): 1:과다흡연, 2:보통흡연, 3:비흡연

```
table(smoke$education, smoke$smoking)
```

```
smoke$education2[smoke$education==1] <- "대졸"
```

```
smoke$education2[smoke$education==2] <- "고졸"
```

```
smoke$education2[smoke$education==3] <- "중졸"
```

```
smoke$smoking2[smoke$smoking==1] <- "과대흡연"
```

```
smoke$smoking2[smoke$smoking==2] <- "보통흡연"
```

```
smoke$smoking2[smoke$smoking==3] <- "비흡연"
```

```
smoke # 가독성을 위한 변수값 변경 결과
```

chi-square 검정

3. 교차표 작성

```
table(smoke$education2, smoke$smoking2)
```

과대흡연 보통흡연 비흡연

고졸	22	21	9
대졸	51	92	68
중졸	43	28	21

chi-square 검정

4. 독립성 검정

```
library(gmodels) # CrossTable() 함수 사용
```

```
CrossTable(smoke$education2, smoke$smoking2, chisq = TRUE)
```

Pearson's Chi-squared test

```
Chi^2 = 18.91092    d.f. = 4    p = 0.0008182573
```

chi-square 검정

2) 동질성 검정 - 교차테이블 이용

- 두 집단의 분포가 동일한가? 분포가 동일하지 않는가?를 검정하는 방법.
- 즉, 동일한 분포를 가지는 모집단에서 추출된 것인지를 검정하는 방법.
 - 예) 귀무가설 : 집단 간의 비율이 동일하다.
 - 예) 교육방법에 따른 만족도에 차이가 없다.
 - 예) 대립가설 : 집단 간의 비율이 동일하지 않다.
 - 예) 교육방법에 따른 만족도에 차이가 있다.

chi-square 검정

1. 파일 가져오기

```
setwd("c:/workspaces/Rwork/data")
data <- read.csv("homogeneity.csv", header=TRUE)
head(data) # 변수 보기
data <- subset(data, !is.na(survey), c(method, survey))
```

chi-square 검정

2. 변수리코딩 - 코딩 변경

```
# method: 1:방법1, 2:방법2, 3:방법3  
# survey: 1:매우 만족, 2:만족, 3:보통, 4: 불만족, 5: 매우 불만족
```

```
# 교육방법2 필드 추가
```

```
data$method2[data$method==1] <- "방법1"  
data$method2[data$method==2] <- "방법2"  
data$method2[data$method==3] <- "방법3"
```

```
# 만족도2 필드 추가
```

```
data$survey2[data$survey==1] <- "매우 만족"  
data$survey2[data$survey==2] <- "만족"  
data$survey2[data$survey==3] <- "보통"  
data$survey2[data$survey==4] <- "불만족"  
data$survey2[data$survey==5] <- "매우 불만족"
```

chi-square 검정

3. 교차분할표 작성

```
table(data$method2, data$survey2) # 교차표 생성 -> table(행, 열)
```

만족 매우만족 매우불만족 보통 불만족

방법1 8 5 6 15 16 -> 50

방법2 14 8 6 11 11 -> 50

방법3 7 8 9 11 15 -> 50

주의 : 반드시 각 집단별 길이(50)가 같아야 한다.

chi-square 검정

4. 동질성 검정 - 모수 특성치에 대한 추론검정

```
chisq.test(data$method2, data$survey2)
```

Pearson's Chi-squared test

```
data: data$method2 and data$survey2  
X-squared = 6.5447, df = 8, p-value = 0.5865
```

<해설>

유의수준 0.05에서 χ^2 값이 6.545, 자유도 8, 그리고 유의확률 0.586을 보이고 있다. 즉 6.545 이상의 카이제곱값이 얻어질 확률이 0.586라는 것을 보여주고 있다.

이 값은 유의수준 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 '교육방법에 따른 만족도에 차이가 없다.'라고 할 수 있다.

집단 간 차이 분석

집단 간 차이 분석

Ttest_Anova 수업내용

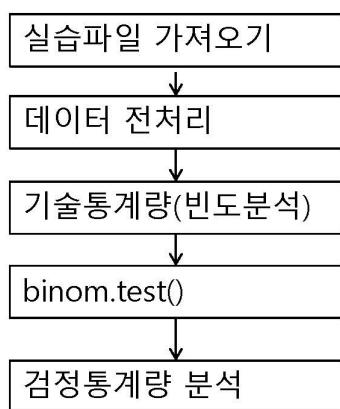
- 1) 단일 집단 분석
- 2) 두 집단 분석
- 3) 세 집단 분석(분산 분석)

단일집단 비율 검정

```
#####
# 추론통계학 분석 - 1-1. 단일집단 비율검정
#####
# 방법 : 1개 집단의 비율과 기존 집단과의 비율 차이 분석
# 작업절차
#   1. 실습데이터 가져오기
#   2. 빈도수와 비율계산
#   3. binom.test() 이용
#####
```

단일집단 비율 검정

- 분석절차



단일집단 비율 검정

<연구가설>

- 연구가설(H_1) : 기존 2020년도 고객 불만율과 2021년도 CS교육 후 불만율에 차이가 있다.
- 귀무가설(H_0) : 기존 2020년도 고객 불만율과 2021년도 CS교육 후 불만율에 차이가 없다.

<연구환경>

2020년도 114 전화번호 안내고객을 대상으로 불만을 갖는 고객은 20%였다. 이를 개선하기 위해서 2021년도 CS교육을 실시한 후 150명 고객을 대상으로 조사한 결과 14명이 불만을 갖고 있었다. 기존 20% 보다 불만율이 낮아졌다고 할 수 있는가?

대상 파일 : c:/workspaces/Rwork/data/one_sample.csv
해당 변수 : survey(만족도)
변수 척도 : 명목척도(y/n)
가정 : 기존 불만율과 CS교육 후 불만율 분석

단일집단 비율 검정

1. 실습데이터 가져오기

```
getwd()  
  
setwd("c:/workspaces/Rwork/data")  
  
data <- read.csv("one_sample.csv", header=TRUE)  
  
head(data)  
  
x <- data$survey # 만족도 변수
```

단일집단 비율 검정

2. 빈도수와 비율 계산

```
summary(x) # 결측치 없음  
length(x) # 150개  
table(x)  
#   x  
# 0 1  
# 14 136 -> 0:불만족(14), 1: 만족(136)
```

```
#table(x, useNA="ifany") # 시리얼 데이터와 NA 개수 출력 시
```

```
install.packages("prettyR")  
library(prettyR) # freq() 함수 사용  
freq(x)  
# Frequencies for x  
# 1 0 NA  
# 136 14 0 <- 빈도수  
#% 90.7 9.3 0 <- 비율 제공
```

단일집단 비율 검정

1) 만족율 기준 검정

양측검정

```
binom.test(c(136,14), p=0.8) # 기준 80% 만족율 기준 검증 실시  
binom.test(c(136,14), p=0.8, alternative="two.sided", conf.level=0.95)  
# alternative="two.sided" : 양측검정-> p-value = 0.0006735  
# 해설 : 기존 만족율(80%)과 차이가 있다. -> 연구가설 채택
```

단측검정

```
binom.test(c(136,14), p=0.8, alternative="greater", conf.level=0.95)  
# alternative="greater" : 단측검정-> 방향성 # p-value = 0.0003179  
# 해설 : CS교육을 통해서 기존 만족율(80%) 이상의 효과를 얻을 수 있다고  
# 볼 수 있다. 따라서 기존 20% 보다 불만율이 낮아졌다고 할 수 있다.
```

단일집단 비율 검정

2) 불만족율 기준 검정

양측검정

```
binom.test(c(14,136), p=0.2) # 기준 20% 불만족율 기준 검증 실시  
binom.test(c(14,136), p=0.2, alternative="two.sided", conf.level=0.95)  
# alternative="two.sided" : 양측검정-> p-value = 0.0006735  
# 해설 : 기존 불만족율(20%)과 차이가 있다. -> 연구가설 채택
```

단측검정

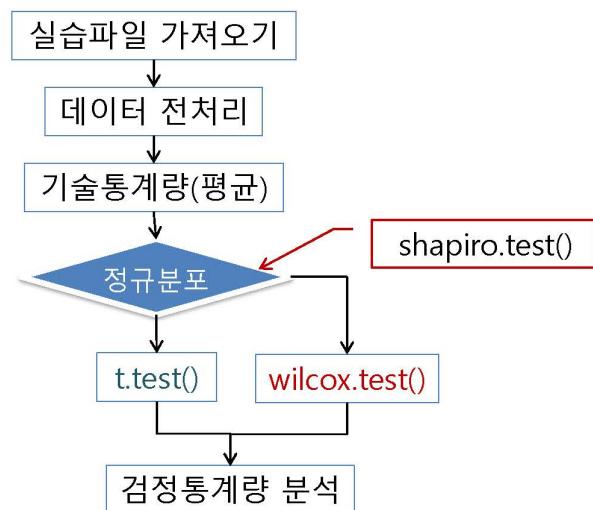
```
binom.test(c(14,136), p=0.2, alternative="greater", conf.level=0.95)  
# alternative="greater" : 단측검정-> 방향성 # p-value = 0.9999  
# 불만족율 20% 보다 크지 않다.  
binom.test(c(14,136), p=0.2, alternative="less", conf.level=0.95)  
# p-value = 0.0003179 -> 불만족율 20% 보다 적다.
```

단일집단 평균 검정

```
#####
# 주론통계학 분석 - 1-2. 단일집단 평균 검정(단일표본 T검정)
#####
# 방법 : 1개 집단의 평균과 어떤 특정한 값과 차이가 있는지 검증
# 작업절차
#   1. 실습파일 가져오기
#   2. 데이터 분포 및 결측치 제거(데이터 정제)
#   3. 정규분포 검정 : 모집단의 특성 반영 유무
#   4. 가설검정(모수/비모수) -> t.test()/wilcox.test()
#####
```

단일집단 평균 검정

● 분석절차



단일집단 평균 검정

<연구가설>

- 연구가설(H_1) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다.
- 귀무가설(H_0) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.

<연구환경>

국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A 회사에서 생산된 노트북 평균 사용시간과 차이가 있는지를 검정하기 위해 A회사 노트북 150대를 랜덤으로 선정하여 검정을 실시한다.

```
# 대상 파일 : c:/workspaces/Rwork/data/one_sample.csv  
# 해당 변수 : time  
# 변수 척도 : 비율척도(직접 입력한 수치 데이터)  
# 가정 : 기존 노트북 평균 사용시간 vs A회사 노트북 평균 사용시간  
# 검정 : 노트북 평균 사용시간 수집 -> 평균 -> 정규성 검정 -> T검정
```

단일집단 평균 검정

1. 실습파일 가져오기

```
setwd("c:/workspaces/Rwork/data")  
  
data <- read.csv("one_sample.csv", header=TRUE)  
  
head(data)  
  
x <- data$time # 노트북 사용 시간  
  
head(x)
```

단일집단 평균 검정

2. 데이터 분포 /결측치 제거

```
summary(x) # NA-41개  
mean(x) # error  
mean(x, na.rm=T) # NA 제외 평균(방법1)  
# 데이터 정제 -> 5.556881  
x1 <- na.omit(x) # NA 제외 평균(방법2)  
x1  
  
# 평균(mean) 특징  
# 평균 모양 : 양측에 대한 균형  
# 대상 : 수치 데이터 -> 비율(ratio)  
# 적용 : 평균 차이 검정으로 의사결정  
# 평균 검정 : 평균에 의미가 있는가 검정, 평균을 중심으로 종 모양 형성  
# 왜도 : 한쪽으로 치우쳐진 정도
```

단일집단 평균 검정

3. 정규분포 검정

```
# 정규분포(바른 분포) : 평균에 대한 검정  
# 정규분포 검정 귀무가설 : 정규분포와 차이가 없다.  
# shapiro학자가 만든 함수 이용 : shapiro.test()  
shapiro.test(x1) # x1 데이터에 대한 정규분포를 검정하는 함수  
# W = 0.9914, p-value = 0.7242 <- 정규분포  
# 검정결과 분석 : 0.05보다 작으면 정규분포가 아닌 것으로 판단  
# 명목척도 -> 보기 항목으로 정규분포가 그려지기 때문에 의미 없음  
# 비율척도, 수치 기반 척도(평균에 의미 있는 척도) -> 정규분포 검정 필요  
  
# 정규분포(모수검정) -> t.test()  
# 비정규분포(비모수검정) -> wilcox.test()  
  
hist(x1) # 정규분포 형태
```

단일집단 평균 검정

4. 가설검정 - 모수/비모수

```
# t.test()  
# - 모집단의 평균값을 검정하는 함수  
# - 예) 기존평균사용시간 5.2시간 기준으로 검정(같다 vs 차이)  
help(t.test)  
# t -> student에서 t
```

1) 양측검정

```
t.test(x1, mu=5.2) # mu(그리스 로마 - 평균) : 기존 5.2시간 기준 검정  
# x1 : 표본집단 평균, mu=5.2, 모집단의 평균값  
  
# 정제 데이터와 5.2시간 비교  
t.test(x1, mu=5.2, alter="two.side", conf.level=0.95)  
# p-value = 0.0001417  
# 해설 : 평균 사용시간 5.2시간과 차이가 있다.(귀무가설 기각)
```

단일집단 평균 검정

● 점추정 vs 구간추정

```
#alternative hypothesis: true mean is not equal to 5.2  
#95 percent confidence interval:  
# 5.377613 5.736148 -> 구간추정(95% 신뢰구간 추정)  
#sample estimates:  
# mean of x  
# 5.556881 -> 점추정 : mean값과 직접 비교하여 추정  
  
# 점추정(point) vs 구간추정(interval estimation)  
# 점추정 : 모수를 하나의 값으로 추정(평균이나 중위수 사용)  
# 구간추정 : 모수가 포함될 것이라고 제시하는 구간추정(신뢰구간)
```

단일집단 평균 검정

2) 단측검정

```
t.test(x1, mu=5.2, alter="greater", conf.level=0.95)  
# p-value = 7.083e-05 = 0.00007083  
# 해설 : A회사 노트북의 평균 사용시간은 5.2시간 보다 더 길다.
```

```
# 검정 결과를 변수에 저장하여 특정 변수 확인하기  
result <- t.test(x1, mu=5.2, alter="greater", conf.level=0.95)  
names(result)  
str(result)  
result$p.value # 7.083346e-05 -> 세밀한 정보 제공
```

단일집단 평균 검정

【단일표본 t-검정 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다. 귀무가설(H0) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.
2) 연구환경	국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A회사에서 생산된 노트북 사용시간과 차이가 있는지를 검정하기 위해서 A회사 노트북 150대를 랜덤으로 선정하여 검정을 실시한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	단일표본 T검정
5) 검정통계량	$t = 3.9461, df = 108$
6) 유의확률	$P = 0.0001417$
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이를 보인다고 할 수 있다. 즉, 국내에서 생산된 노트북의 평균 사용 시간은 5.2이며, A회사에서 생산된 노트북의 평균 사용 시간은 5.56으로 국내 평균 사용 시간 보다 더 길다고 할 수 있다.

두 집단 비율 검정

```
#####
# 추론통계학 분석 - 2-1. 두 집단 비율 검정
#####
# 방법 : 두 집단 간 비율 차이에 관한 분석
# 작업절차
#   1. 실습파일 가져오기
#   2. 두 집단 subset 작성(데이터 정제, 전처리)
#       -> 데이터 정체, 전처리
#       -> 기술통계량 - 빈도수
#       -> 두 변수(집단)에 대한 교차분석
#   3. 두 집단 비율차이 검정
#       -> prop.test()
#####
```

두 집단 비율 검정

● 분석절차



두 집단 비율 검정

<연구가설>

- 연구가설(H_1) : 두 가지 교육방법에 따라 교육생의 만족율에 차이가 있다.
- 귀무가설(H_0) : 두 가지 교육방법에 따라 교육생의 만족율에 차이가 없다.

<연구환경>

IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 교육을 실시하였다. 2가지 교육방법 중 더 효과적인 교육방법을 조사하기 위해서 교육생 300명을 대상으로 설문을 실시하였다. 조사한 결과는 다음 표와 같다.

대상 파일 : c:/workspaces/Rwork/data/two_sample.csv
해당 변수 : method(명목척도), survey(명목척도)
변수 척도 : 명목척도 : 빈도수(기술통계량)

두 집단 비율 검정

<설문조사 교차표>

교육방법	만족	불만족	참가자
PT교육	110	40	150
코딩교육	135	15	150
합계	245	55	300

두 집단 비율 검정

1. 실습데이터 가져오기

```
getwd()  
setwd("c:/workspaces/Rwork/data")  
data <- read.csv("two_sample.csv", header=TRUE)  
data  
head(data) # 변수명 확인
```

두 집단 비율 검정

2. 두 집단 subset 작성

```
data$method # 1, 2 -> 노이즈 없음  
data$survey # 1(만족), 0(불만족)  
# 데이터 정체/전처리  
x<- data$method # 교육방법(1, 2) -> 노이즈 없음  
y<- data$survey # 만족도(1: 만족, 0:불만족)  
x;y
```

두 집단 비율 검정

1) 데이터 확인

```
# 교육방법 1과 2 모두 150명 참여  
table(x) # 1 : 150, 2 : 150  
# 교육방법 만족/불만족  
table(y) # 0 : 55, 1 : 245
```

2) data 전처리 & 기술통계량 -> 빈도수 -> 정규성 검정 필요 없음

```
# 두 변수에 대한 교차분석  
table(x, y, useNA="ifany") # 결측치 까지 출력  
#####  
# y  
#x 0 1  
# 1 40 110 -> 방법A - 110 만족  
# 2 15 135 -> 방법B - 135 만족  
#####
```

두 집단 비율 검정

3. 두 집단 비율차이검증 - prop.test()

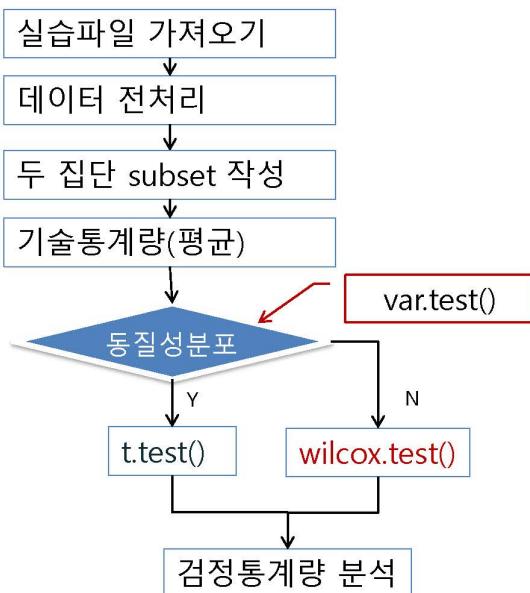
```
help(prop.test) # prop.test(x,n,p, alternative, conf.level, correct)  
  
# 양측검정  
prop.test(c(110,135),c(150,150)) # 방법A 만족도와 방법B 만족도 차이 검정  
# p-value = 0.0003422  
#sample estimates: 집단 간 비율  
# prop 1 prop 2  
#0.7333333 0.9000000  
prop.test(c(110,135),c(150,150), alternative="two.sided", conf.level=0.95)  
# 해설) p-value = 0.0003422 - 두 집단간의 만족도에 차이가 있다.  
  
# 단측검정  
prop.test(c(110,135),c(150,150), alter="greater", conf.level=0.95)  
# 해설) p-value=0.9998 : 방법A가 방법B에 비해 만족도가 낮은 것으로 파악
```

두 집단 평균 검정 (독립표본 T검정)

```
#####
# 추론통계학 분석 - 2-2. 두 집단 평균 검정(독립표본 T검정)
#####
# 방법 : 두 집단 간 평균 차이에 관한 분석
# 작업절차
#   1. 실습파일 가져오기
#   2. 두 집단 subset 작성(데이터 정제, 전처리)
#   3. 두 집단 간 동질성 검증(정규분포 검정)
#       -> var.test()
#   4. 두 집단 평균 차이 검정
#       -> t.test() or wilcox.test()
#####
```

두 집단 평균 검정

● 분석절차



두 집단 평균 검정

<연구가설>

- 연구가설(H_1) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
- 귀무가설(H_0) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 없다.

<연구환경>

IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육 방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기 시험을 실시하였다. 두 집단간 실기시험의 평균에 차이가 있는가 검정한다.

```
# 대상 파일 : c:/workspaces/Rwork/data/two_sample.csv  
# 해당 변수 : method(명목척도), score(비율척도)  
# 대상 변수 : 교육방법, 시험성적  
# 모형(모델) : 교육방법(A/B) -> 시험성적(비율-성적)
```

두 집단 평균 검정

1. 실습파일 가져오기

```
data <- read.csv("c:/workspaces/Rwork/data/two_sample.csv", header=TRUE)  
data  
print(data)  
head(data) #4개 변수 확인  
summary(data) # score - NA's : 73개
```

2. 두 집단 subset 작성(데이터 정제, 전처리)

```
result <- subset(data, !is.na(score), c(method, score))  
# c(method, score) : data의 전체 변수 중 두 변수만 추출  
# !is.na(score) : na가 아닌 것만 추출  
# 위에서 정제된 데이터를 대상으로 subset 생성  
result # 방법1과 방법2 혼합됨  
length(result$score) # 227
```

두 집단 평균 검정

```
# 데이터 분리  
1) 교육방법 별로 분리  
a <- subset(result,method==1)  
b <- subset(result,method==2)  
  
2) 교육방법에서 점수 추출  
a1 <- a$score  
b1 <- b$score  
  
# 기술통계량 -> 평균값 적용 -> 정규성 검정 필요  
length(a1); # 109  
length(b1); # 118
```

두 집단 평균 검정

3. 분포모양 검정 : 두 집단의 분포모양 일치 여부 검정
 - # 귀무가설 : 두 집단 간 분포의 모양이 동질적이다.
 - # 두 집단간 동질성 비교(분포모양 분석)
var.test(a1, b1) # p-value = 0.3002 -> 차이가 없다.
 - # 동질성 분포 : t.test()
 - # 비동질성 분포 : wilcox.test()
4. 가설검정 – 두 집단 평균 차이검정
 - t.test(a1, b1)
 - t.test(a1, b1, alter="two.sided", conf.int=TRUE, conf.level=0.95)
p-value = 0.0411 - 두 집단간 평균에 차이가 있다.
 - t.test(a1, b1, alter="greater", conf.int=TRUE, conf.level=0.95)
p-value = 0.9794 : a1을 기준으로 비교 -> a1이 b1보다 크지 않다.
 - t.test(a1, b1, alter="less", conf.int=TRUE, conf.level=0.95)
p-value = 0.02055 : a1이 b1보다 작다.

두 집단 평균 검정

【독립표본 t-검정 결과 정리 및 기술】

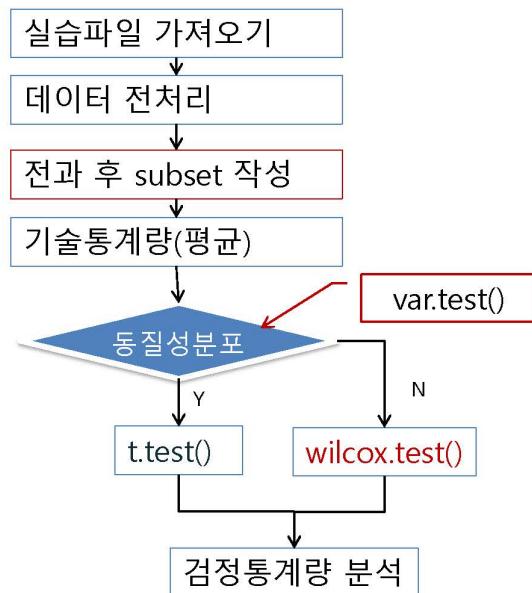
1) 가설 설정	연구가설(H1) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다. 귀무가설(H0) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 없다.
2) 연구환경	IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기시험을 실시하였다. 두 집단간 실기시험의 평균에 차이가 있는가 검정한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	독립표본 T검정
5) 검정통계량	$t = -2.0547, df = 218.192$
6) 유의확률	$P = 0.0411$
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 “교육 방법에 따른 두 집단 간 실기 시험의 평균에 차이가 있다”라고 말할 수 있다. 단측 검정을 실시한 결과 교육 방법1이 교육 방법2 보다 크지 않은 것으로 나타났다. 즉, 실시간 코딩 교육 방법이 교육 효과가 더 높은 것으로 분석된다.

대응 두 집단 평균 검정 (대응표본 T검정)

```
#####
# 추론통계학 분석 – 2-3. 대응 두 집단 평균 검정(대응표본 T검정)
#####
# 방법 : 동일한 표본을 대상으로 측정된 두 변수의 평균 차이를 검정하는
#         분석.
# 작업절차
#   1. 실습파일 가져오기
#   2. 두 집단 subset 작성(데이터 정제, 전처리)
#   3. 두 집단 간 동질성 검증(정규분포 검정)
#       -> var.test(x,y paired=TRUE)
#   4. 두 집단 평균 차이 검정
#       -> t.test(x,y, paired=TRUE)
#       -> wilcox.test(x,y, paired=TRUE)
#####
```

대응 두 집단 평균 검정

● 분석절차



대응 두 집단 평균 검정

<연구가설>

- 연구가설(H_1) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다.
- 귀무가설(H_0) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 없다.

<연구환경>

A교육센터에서 교육생 100명을 대상으로 교수법 프로그램 적용 전에 실기 시험을 실시한 후 1개월 동안 동일한 교육생에게 교수법 프로그램을 적용한 후 실기시험을 실시한 점수와 평균에 차이가 있는가 검정한다.

```
# 대상 파일 : c:/workspaces/Rwork/data/paired_sample.csv  
# 해당 변수 : before, after  
# 대상 변수 : 교수법 프로그램을 적용하기 전 / 후  
# 모형(모델) : 교수법 전/후 -> 시험성적(비율-성적)
```

대응 두 집단 평균 검정

1. 실습파일 가져오기

```
getwd()  
setwd("c:/workspaces/Rwork/data/")  
data <- read.csv("paired_sample.csv", header=TRUE)
```

2. 두 집단 subset 작성

1) 데이터 정제

```
# subset(x, subset, select, ..) -> subset은 반드시 논리적이어야 함  
result <- subset(data, !is.na(after), c(before,after))  
# data 테이블을 대상으로 after 결측치 제거하여 subset 생성  
result # 결측 데이터 4개
```

대응 두 집단 평균 검정

2) 동일한 사람에게 두 번 질문

```
x <- result$before # 교수법 적용 전 점수  
y <- result$after # 교수법 적용 후 점수  
x;y # 대응포인 경우 표본수가 같아야 한다. -> 짹을 이루어야 되기 때문에  
length(x) # 96 -> 4개 결측치 제거  
length(y) # 96  
mean(x) # 5.16875  
mean(y) # 6.220833 -> 1.052 정도 증가
```

3. 분포모양 검정 : 두 집단의 분포모양 일치 여부 검정

```
var.test(x, y, paired=TRUE) # p-value = 0.7361 -> 차이가 없다.  
# 동질성 분포 : t.test()  
# 비동질성 분포 : wilcox.test()
```

대응 두 집단 평균 검정

4. 가설검정

```
t.test(x, y, paired=TRUE) # p-value < 2.2e-16  
# 단측검정 - 방향성 검정  
t.test(x, y, paired=TRUE, alter="greater", conf.int=TRUE, conf.level=0.95)  
#p-value = 1 -> x을 기준으로 비교 : x가 y보다 크지 않다.  
t.test(x, y, paired=TRUE, alter="less", conf.int=TRUE, conf.level=0.95)  
# p-value < 2.2e-16 -> x을 기준으로 비교 : x가 y보다 적다.
```

<해설>

교수법 프로그램을 적용하기 전 시험성적과 교수법 프로그램을 적용한 후 시험성적을 비교한 결과 교수법을 적용한 후 시험성적이 약 1.052 점수가 향상된 것으로 나타났다.

대응표본 t-검정

【대응표본 t-검정 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다. 귀무가설(H0) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 없다.
2) 연구환경	A교육센터에서 교육생 100명을 대상으로 교수법 프로그램 적용 전에 실기시험을 실시한 후 1개월 동안 동일한 교육생에게 교수법 프로그램을 적용한 후 실기시험을 실시한 점수와 평균에 차이가 있는가 검정한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	대응표본 T검정
5) 검정통계량	$t = -13.6424, df = 95$
6) 유의확률	$P = < 2.2e-16$
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 “교수법 프로그램 적용 전과 적용 후의 두 집단 간 학습력의 평균에 차이가 있다.”라고 말할 수 있다. 또한 단측 검정을 실시한 결과 교수법 프로그램 적용 전 학습력이 교수법 프로그램 적용 후 학습력 보다 크지 않은 것으로 나타났다. 즉, 교수법 프로그램이 학습력에 효과가 있는 것으로 분석된다.

세 집단 비율 검정

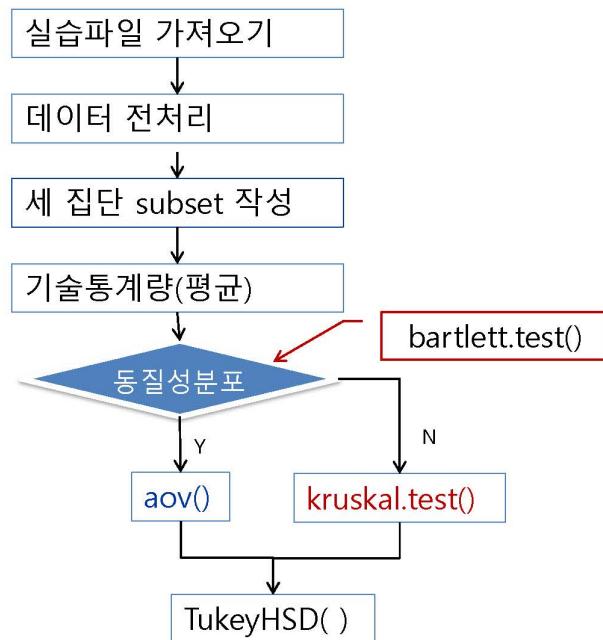
```
#####
# 추론통계학 분석 - 3-1. 세 집단 비율 검정
#####
# 방법 : 세 집단(이상)간 빈도수에 대한 비율 차이를 검정하는 분석
# 작업절차
#   1. 파일 가져오기
#   2. 데이터 정제/전처리 - NA, outline 제거
#   3. 세집단 subset 작성
#     -> 코딩 변경
#     -> 기술통계량(빈도수)
#     -> 교차표 작성
#   4. 세 집단 비율 차이 검정 : prop.test()
#   5. 검정통계량 분석
#####
```

세 집단 평균 검정

```
#####
# 추론통계학 분석 - 3-2. 세 집단 평균 검정(분산 분석)
#####
# 방법 : 세 집단(이상)간 평균 차이에 관한 분석
# 작업절차
#   1. 파일 가져오기
#   2. 데이터 정제/전처리 - NA, outline 제거
#   3. 세집단 subset 작성
#     -> 코딩 변경
#     -> 기술통계량(빈도수)
#     -> 교차표 작성
#   4. 세집단 동질성 검정 : bartlett.test()
#   5. 분산검정 : aov() or kruskal.test()
#   6. 사후검정 : TukeyHSD()
#####
```

세 집단 평균 검정

● 분석절차



세 집단 평균 검정

<연구가설>

- 연구가설(H_1) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.
- 귀무가설(H_0) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.

<연구환경>

세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 50명씩을 대상으로 실기시험을 실시하였다. 세 집단간 실기시험의 평균에 차이가 있는가 검정한다.

```
# 대상 파일 : c:/workspaces/Rwork/data/three_sample.csv  
# 해당 변수 : method(명목척도), score(비율척도)  
# 대상 변수 : 교육방법, 시험성적  
# 모형(모델) : 교육방법(A/B) -> 시험성적(비율-성적)
```

세 집단 평균 검정

1. 파일 가져오기

```
data <- read.csv("c:/workspaces/Rwork/data/three_sample.csv", header=TRUE)
```

2. 데이터 정제/전처리 - NA, outline 제거

```
data <- subset(data, !is.na(score), c(method, score))  
data # method, score
```

```
# 차트이용 - outline 보기(데이터 분포 현황 분석)  
plot(data$score) # 차트로 outline 확인 : 50이상과 음수값  
barplot(data$score) # 바 차트  
boxplot(data$score) # 박스 차트  
mean(data$score) # 14.45
```

세 집단 평균 검정

```
# outline 제거 - 평균(14) 이상 제거  
length(data$score) # 91  
data2 <- subset(data, score <= 14) # 14이상 제거  
length(data2$score) # 88(3개 제거)  
  
##### 정제된 데이터 보기 #####  
x <- data2$score  
boxplot(x)  
plot(x)  
bp <- boxplot(data2$score) # 차트 결과 저장
```

세 집단 평균 검정

3. 세 집단 subset 작성

```
# 코딩 변경 - 변수 리코딩 -> method: 1:방법1, 2:방법2, 3:방법3
data2$method2[data2$method==1] <- "방법1"
data2$method2[data2$method==2] <- "방법2"
data2$method2[data2$method==3] <- "방법3"
table(data2$method2) # 교육방법 별 빈도수
#방법1 방법2 방법3
# 31 27 30

x <- table(data2$method2)
#교육방법에 따른 시험성적 평균 구하기
y <- tapply(data2$score, data2$method2, mean)
# 방법1 방법2 방법3
# 4.187097 6.800000 5.610000
out <- data.frame(교육방법=x, 시험성적=y)
out # 교육방법에 따른 시험성적 평균 교차표
# 교육방법.Var1 교육방법.Freq 시험성적
#방법1 방법1 31 4.187097
#방법2 방법2 27 6.800000
#방법3 방법3 30 5.610000
```

세 집단 평균 검정

4. 동질성 검정 - 정규성 검정

```
# bartlett.test(종속변수 ~ 독립변수) # 독립변수 - 세 집단
bartlett.test(score ~ method, data=data2)
#Bartlett's K-squared = 3.3157, df = 2, p-value = 0.1905

# data2의 테이블을 대상으로
# 3집단 이상인 경우 : (종속변수 ~ 독립변수) 분석식으로 표현
# ~ : 틸드 -> 집단별로 subset를 만들지 않고 사용하도록 편의성 제공

# 귀무가설 : 세 집단 간 분포의 모양이 동질적이다.
# 해설 : 유의수준 크기 때문에 귀무가설을 기각할 수 없다.

# 동질한 경우 aov() 사용 : aov - Analysis of Variance(분산분석)
# 동질하지 않은 경우 - kruskal.test()
```

세 집단 평균 검정

5. 분산검정

```
help(aov)
# 분산분석 결과를 result에 저장
# 귀무가설 : 세 집단의 평균에 차이가 없다.

data2$method2 <- factor(data2$method2)
# factor() : method가 집단 구성변수라는 것을 명시

# aov(종속변수 ~ 독립변수, data=data set)
result <- aov(score ~ method2, data=data2)
names(result)

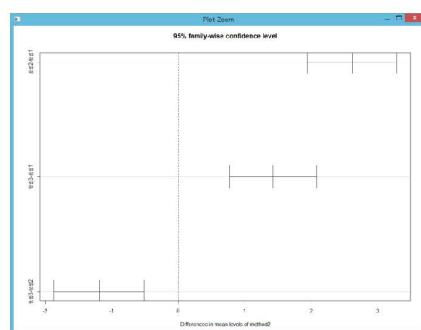
# aov()의 결과값은 summary()함수를 사용해야 p-value 확인
summary(result) # Pr(>F) : 9.39e-14 -> 귀무가설 기각
# 해석 : 0.05보다 현저하게 작음
# 교육방법에 따라서 시험성적 평균에 차이가 있다.
```

세 집단 평균 검정

6. 사후검정

```
# 집단간 차이 상세보기 -> A!=B!=C, A==B!=C, A!=B==C
TukeyHSD(result) # 분산분석의 결과로 사후검정
# $method2
#          diff      lwr      upr      p adj
#방법2-방법1 2.612903 1.9424342 3.2833723 0.0000000
#방법3-방법1 1.422903 0.7705979 2.0752085 0.0000040
#방법3-방법2 -1.190000 -1.8656509 -0.5143491 0.0001911
# 교육방법 간 비교 -> p값(tapply 차이 검정) -> 4.187097 6.800000 5.610000

# 해석) A B C 집단간 모두 차이가 있다.
plot(TukeyHSD(result))
# 그래프 보기(lwr~upr변수 이용)
```



세 집단 평균 검정

【분산분석 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다. 귀무가설(H0) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.
2) 연구환경	세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 50명씩을 대상으로 실기시험을 실시하였다. 세 집단간 실기시험의 평균에 차이가 있는가 검정한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	ANOVA 검정
5) 검정통계량	F = 43.58, Df = 2, Sum Sq=99.37, Mean Sq = 49.68
6) 유의확률	P = 9.39e-14 ***
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있는 것으로 나타났다. 또한 사후검정 방법인 Tukey 분석을 실시한 결과 '방법2-방법1'의 평균 점수의 차이가 가장 높은 것으로 나타났다.

