**## 숫자형 통계분석**

1. 상관관계 (인과관계 없음, 그냥 X,Y는 어느정도 관계가 있다는 전제)

키, 몸무게, 발이 있다면 각각 어떤 영향을 주는지의 관계

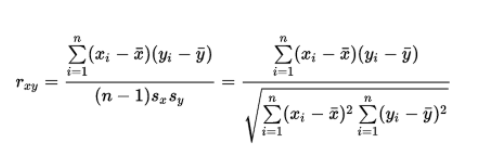
상관관계 얼마나 연관이 되어있는가를 숫자로 알고싶다면?

두 변수 간의 강도를 숫자로 표현하는 방법 -> 상관관계를 확인!

1 : 완전한 직선적인 관계

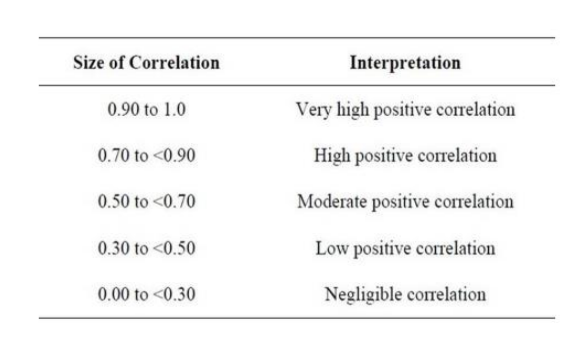
0 : Not 직선적인 관계

-1 : 완전한 반대의 직선적인 관계



공분산 : 두 변수가 같은 방향으로 움직이는 정도

상관계수 : 공분산을 표준편차로 나눈 값



0.9~1은 거의 없는 case / 0.25 이하면 독립변수이고 영향 없음 / 0.3 이상이면 영향 있음

**## 상관계수를 구하는 방법**

1. Pearson(피어슨)

2. Kendall(켄달)

3. spearman(스피어만)

https://bioinformaticsandme.tistory.com/58

**## 상관관계분석(Correlation)**

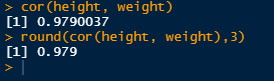
cor()함수 : 상관계수를 구할 수 있음.

height <- c(164, 175, 166, 185)

weight <- c(62, 70, 64, 86)

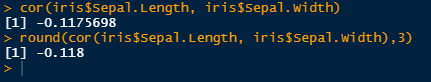
cor(height, weight) ## height, weight의 상관관계 분석 구문

round(cor(height, weight),3) ## 결과값 나온걸 소수점 셋째자리까지 반올림



cor(iris$Sepal.Length, iris$Sepal.Width)

round(cor(iris$Sepal.Length, iris$Sepal.Width),3)



**## 데이터 불러오기 패키지 (결측값이 포함된 데이터)**

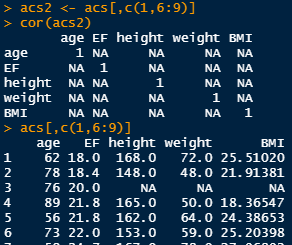
install.packages("moonBook")

library(moonBook)

data(acs)

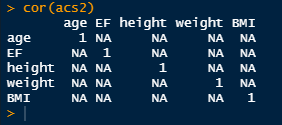
## 연속형 데이터들만 가지고 와서 상관관계를 분석하겠다는 의미 (범주형 X)

acs2 <- acs[,c(1,6:9)]



## 각 변수들의 상관관계가 아래와 같다.

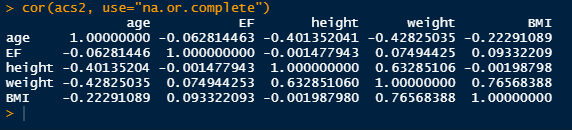
cor(acs2)



## 각 변수들에 있는 NA(결측값)을 제외하고 상관관계를 확인

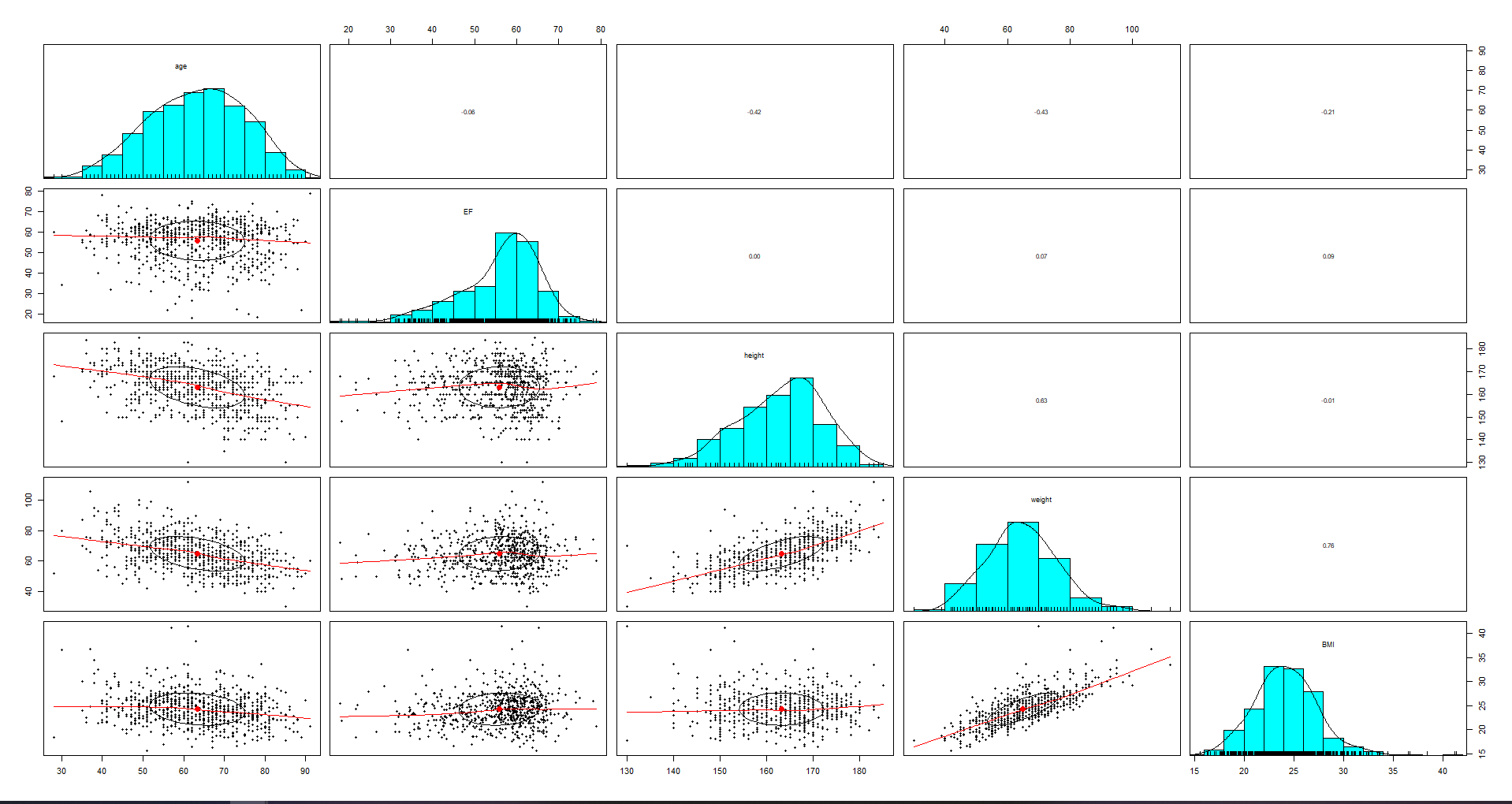
Age와 가장 상관관계가 있는 것은 weight (나이가 들면 들수록 몸무게가 줄어든다)

cor(acs2, use="na.or.complete")



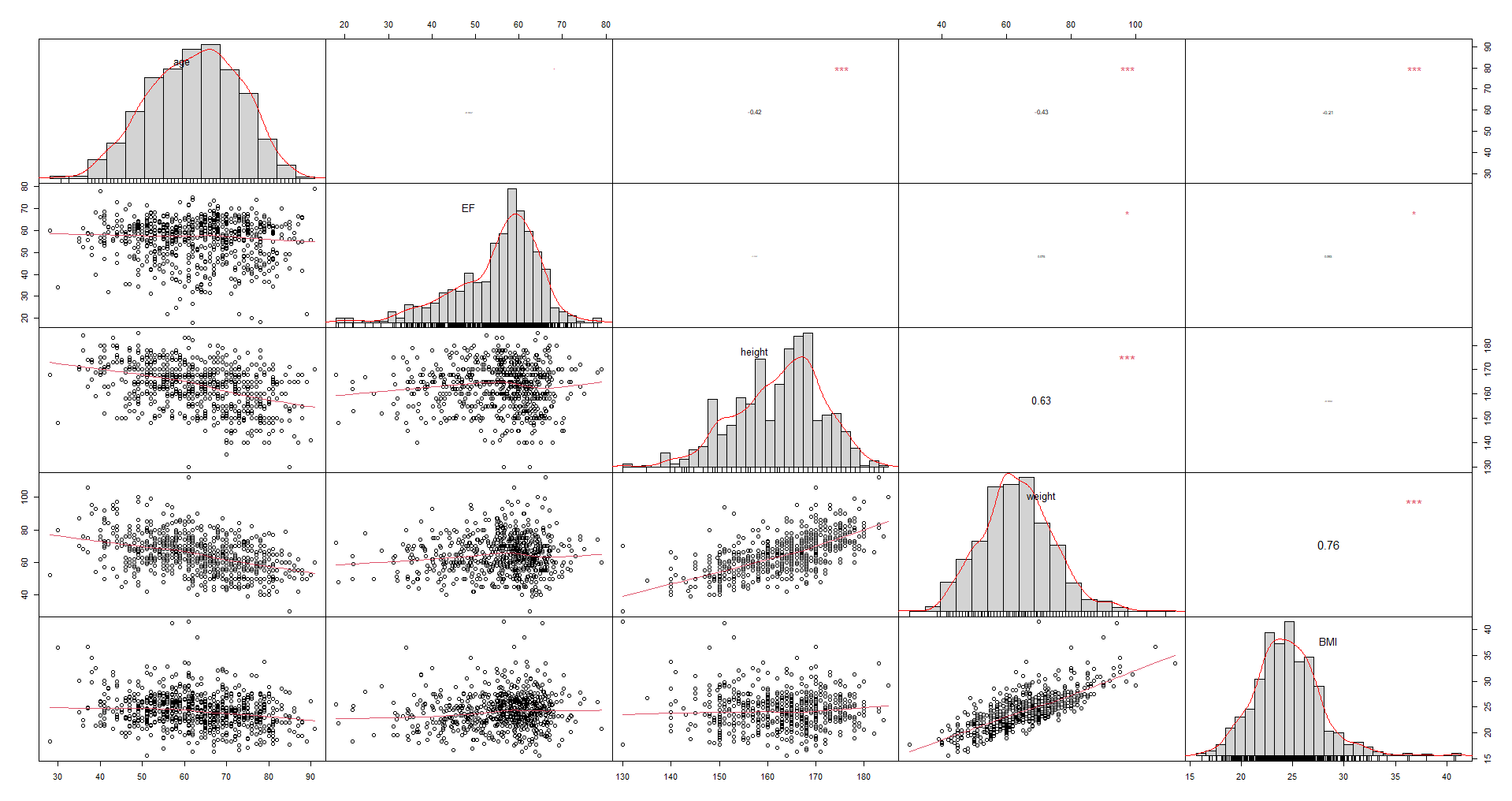
**#산점도행렬 그리기 ## 필살기 함수**

pairs.panels(acs2)



## 히스토그램으로 그리기, 크기 190, 유의확률을 볼 때 별보다는 숫자를 중요시

chart.Correlation(acs2, histogram=TRUE, pch=19)



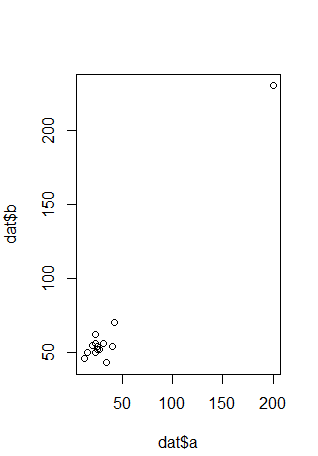
**#킹콩 data의 추가 -> 데이터 하나의 큰 영향**

dat<-data.frame(

a=c(15,20,25,27,31,25,23,23,42,12,34,23,40),

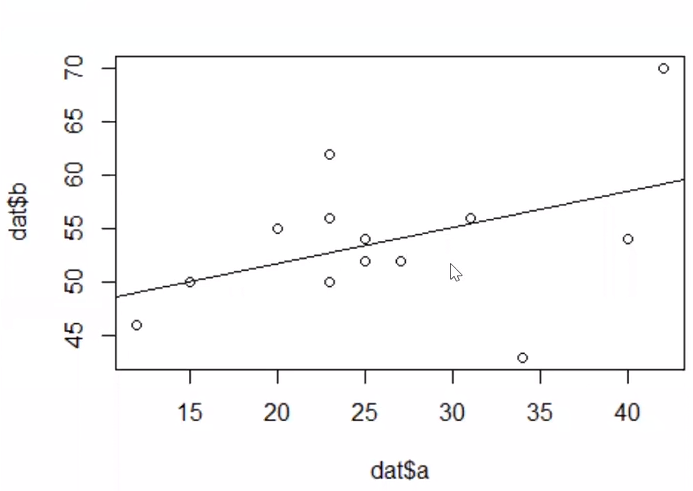
b=c(50,55,52,52,56,54,62,56,70,46,43,50,54)

plot(dat$a,dat$b) **## 산점도**

****

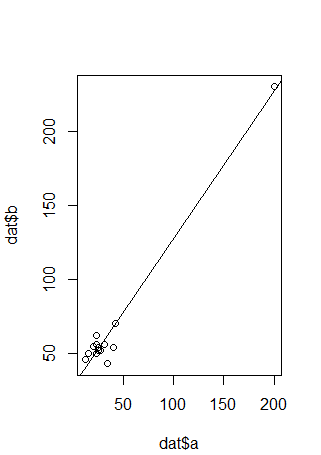
abline(lm(dat$b~dat$a)) **## 추세선 lm(y~x) 순서**

오차를 최소로 하는 선을 긋는 lm()의 결과값은 **절편**과 **기울기**가 생김

****

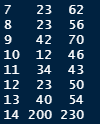
cor(dat$a,dat$b) **## 상관계수 (괜찮은 수치)**

****

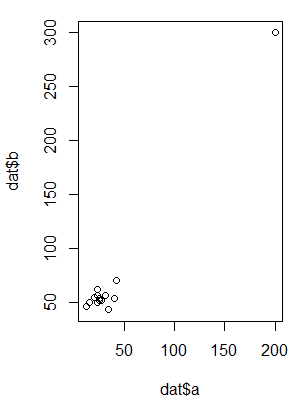
****

**#outlier 추가**

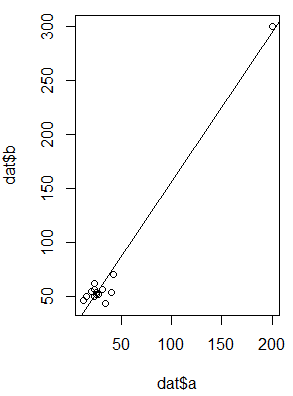
dat[14,]<-c(200,230) ## 아웃라이어가 나타난다. ## 상관계수 시 아웃라이어를 고민필요



plot(dat$a,dat$b)



abline(lm(dat$b~dat$a))



cor(dat$a,dat$b) **## 아웃라이어 선 하나에 의해 상관관계가 높아짐**

**## 분산을 기반으로 통계검정을 했기 때문에**

**상관관계가 아웃라이어에 민감하다. (평균 개념)**



**##heatmap expression**

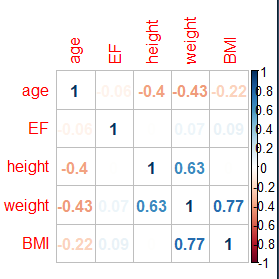
install.packages("corrplot")

library(corrplot)

corrplot(cor(acs2, use="na.or.complete"))

corrplot(cor(acs2, use="na.or.complete"), method="number") **## 히트맵으로 출력**

method : square, ellipse, number, shade, color, pie



**#cor의 비모수적인 표현들 (pearson이 default)**

1. spearman : 양, 음의 개수를 통해서 나오는 방법론

단위가 다르거나, 값 자체에 대해서 랭킹으로 상관관계를 메겨야 할 때 사용

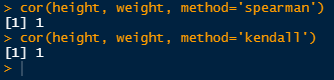
2. kendall's tau

cor(height,weight) **## pearson**

 ## 상관관계는 **pearson을 가장 많이 사용**

cor(height,weight,method="spearman")

cor(height,weight,method="kendall")



**############### 연습문제 ###############**

#1. iris에서 연속형 데이터를 갖고 상관관계를 구하고 Sepal.Length와 가장 상관있는 변수는 무엇인가?

#(2가지 이상의 시각화를 그려보시오)

pairs.panels(iris[,1:4])

chart.Correlation(iris[,1:4], histogram=TRUE, pch=19)

**#mpg에서 qesc까지의 변수를 갖고 상관관계를 구하시오**

pairs.panels(mtcars[,1:7], show.points=10) **## 데이터 글자 크기 조정**

chart.Correlation(mtcars[,1:7], histogram=TRUE, pch=19)

2. T 검정 (A,B반이 수학시험을 봤는데 두 반의 차이가 있어? 2개의 집단의 평균비교 시 사용

1) 단일표본 T 검정 - 특정 집단의 평균이 어떤 숫자와 같은지 다른지를 비교 (잘 사용안함)

1.1) 귀무가설 : 두 집단간의 평균 차이가 없다 (=)

1.2) 대립가설 : 두 집단간의 평균 차이가 있다 (!=)

2) 독립표본 T 검정 - 서로 다른 두 개의 그룹 간의 평균 비교

2.1) 독립성 검정 (서로 다른 그룹인가?)

2.2) 정규성 검정 (정규분포의 형태를 띄고 있는가?)

1,50, 99와 49,50,51은 등분산이 아닌 이분산이다.

**## 정규성을 갖는지 확인하는 검정 (TIP) ## 많이 쓰이지는 않는다.**

– 데이터 크기가 30 넘으면 생략도 가능

- **유의확률이 0.05보다 커야됩니다**(주의!)

- 귀무가설 : 정규성을 갖는다

- 대립가설 : 정규성을 갖지 않는다.

3) 등분산성 검정 (등분산의 형태를 띄고 있는가?)

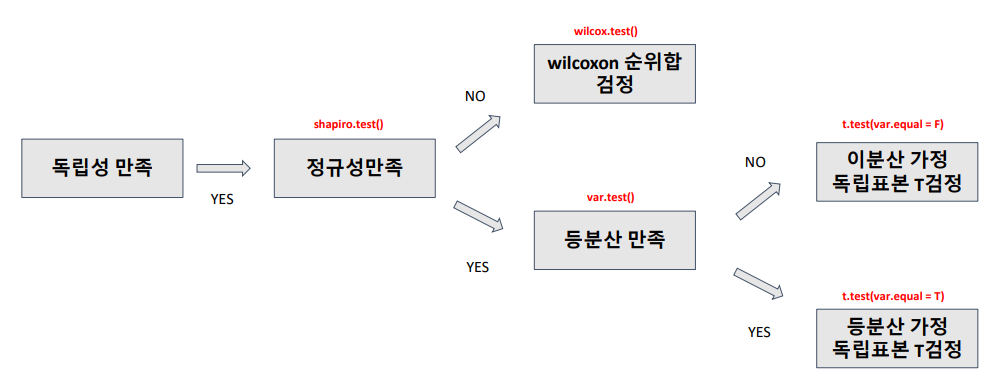
**## 분산이 같은지 확인하는 검정 (TIP) ## 실무에서 많이 쓰인다. (꼭 필요)**

- var.test() - **유의확률이 0.05보다 커야됩니다(주의!)**

- 귀무가설 : 등분산이다

- 대립가설 : 등분산이 아니다.

아래 독립성만족 > 정규성만족 > 등분산 만족의 흐름이 중요하다.



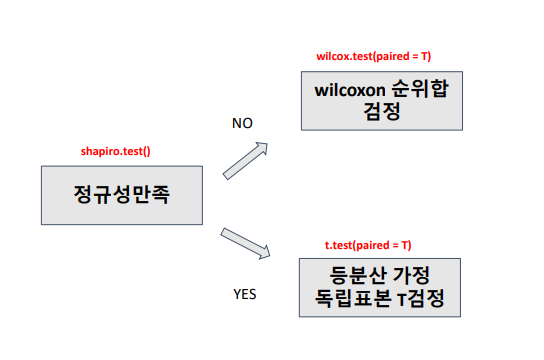
**## 전과 후의 차이가 있는지 없는지의 여부가** 대응표본 T 검정이다.

귀무가설 : 치료 전과 후의 차이의 평균은 0이다(효과 x)

대립가설 : 치료 전과 후의 차이의 평균은 0이 아니다 (효과 0)

기본 가정 1가지

1. **정규성 검정** (1가지만 진행)



같은 집단이기 때문에 어느정도 등분산은 서로 만족한다는 가정하에 진행을 함

대부분 t.test로 진행

t.test(paird=T)가 맞으면 대응표본이고, 그게 아닐 경우에는 독립표본이다.

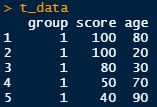
**## 2. 집단에대한 평균비교 t-test (2개의 집단)**

**t\_data<-data.frame(**

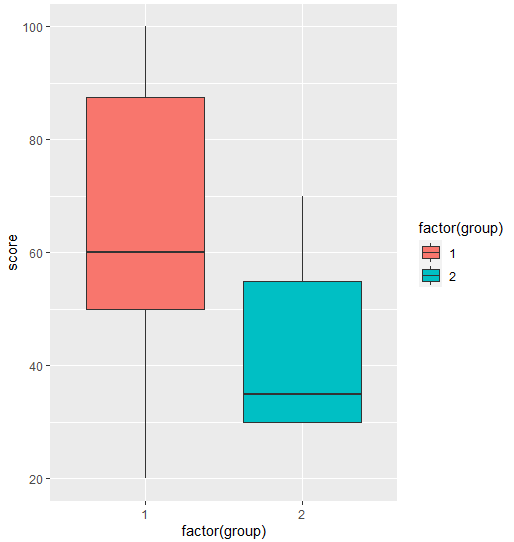
group=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2),

score=c(100,100,80,50,40,90,20,50,50,70,30,40,30,70,30,40,30,60,30,60),

age=c(80,20,30,70,90,20,30,60,50,50,20,30,20,20,25,10,13,12,11,10))



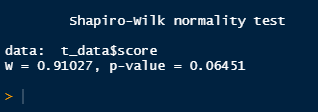
ggplot(t\_data,aes(x=factor(group),y=score,fill=factor(group))) + geom\_boxplot()



**#정규성검정**

#정규성검정 - shapiro

shapiro.test(t\_data$score)



## p-value가 0.05보다 높으면 귀무가설을 띈다. ## 정규성에는 문제가 없다.

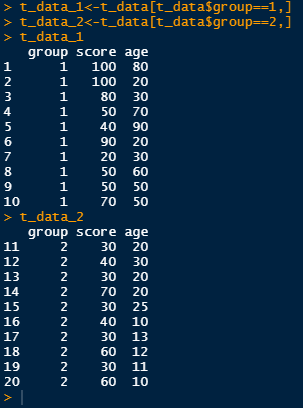
## 0.06451 (유의확률)이 0.05보다 높기 때문에 귀무확률 선택 ( 정규성을 뜬다. -> 정규분포를 띈다.)

**#등분산성 검정**

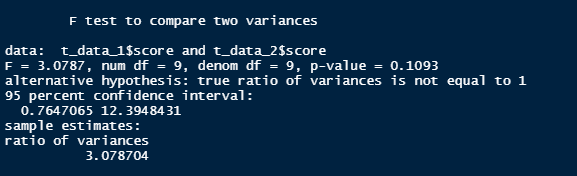
## 1,2번에 대한 데이터

t\_data\_1<-t\_data[t\_data$group==1,]

t\_data\_2<-t\_data[t\_data$group==2,]

****

var.test(t\_data\_1$score,t\_data\_2$score)



* p-value = 0.1093 (귀무가설 채택 -> 등분산이다. )
* 등분산성을 확인하기 위해서 v.test 사용

**#t\_test(등분산 확인)방법 2가지 존재**

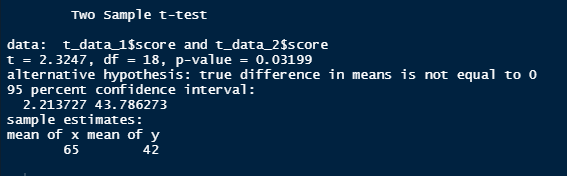
**TYPE1**

t.test(t\_data\_1$score, t\_data\_2$score, var.equal=T) **##var.equal=T (디폴트 값) (아니면 F 써줘야 함)**

**TYPE2**

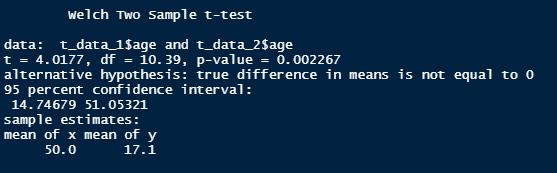
t.test(score~group,data=t\_data,var.equal=T)??? ##score를구할껀데 group에 따라서 진행

**문법: t.test(관측치~집단 구분 기준, 데이터프레임, t-test 유형, 신뢰범위)**

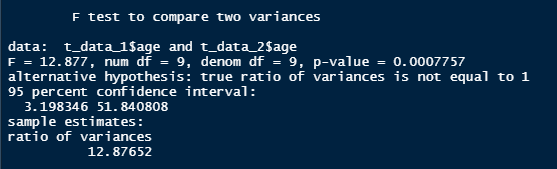
\* p-value = 0.03199, 0.05보다 작기 때문에 대립가설 채택이고, 차이가 존재한다는 의미

**#등분산이 아닐경우**

var.test(t\_data\_1$age,t\_data\_2$age) ## p-value = 0.0007757, 대립가설을 채택하는게 맞는데?



t.test(t\_data\_1$age,t\_data\_2$age,var.equal=F) ## 등분산이 아니기 때문에 var.equal=F 진행

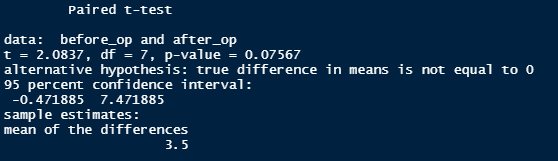


**# 대응 T-test의 수행(전/후비교) - paried=T를 붙여줌**

before\_op = c(137,119,117,122,132,110,114,117) **## 약 먹기 전후로 가정**

after\_op = c(126,111,117,116,135,110,113,112)

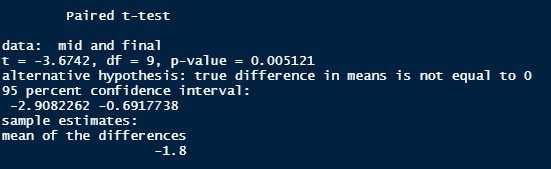
t.test(before\_op,after\_op,paired=T)



\* p-value = 0.07567, 0.05 보다 높으니까 귀무가설을 채택한다.

mid = c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28) **## 중간 기말 시험결과 차이**

final = c(19, 20, 24, 24, 25, 25, 26, 26, 28, 32)



* 귀무가설 p-value = 0.005121, 0.05 보다 낮아 대립가설을 채택하고 차이는 존재한다.

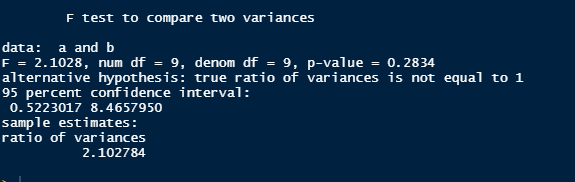
**################## T검정 연습해보기 ###################**

# 1 (독립표본 T 검정)

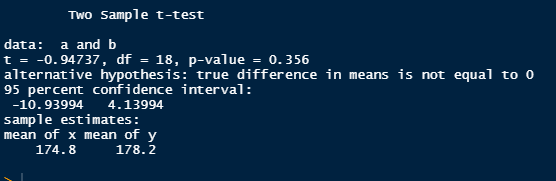
a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)

b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)

var.test(a,b)



t.test(a,b, var.equal = T)



## 귀무가설 : a집단과 b집단의 차이가 없다.

## 대립가설 : a집단과 b집단의 차이가 있다.

## 유의확률은 0.356이므로 유의수준 0.356보다 크므로 귀무가설 채택하며,

즉, a집단과 b집단의 차이가 없다.

**# 2. am 변수에 따라 mpg가 차이가 있는지 확인하시오 (1,1 / 0,1 이유를 모르겠음)**

var.test(mtcars[mtcars$am==1,1],mtcars[mtcars$am==0,1])

t.test(mtcars[mtcars$am==1,1],mtcars[mtcars$am==0,1],var.equal = T)

<https://studymaps.tistory.com/20>

3. 분산분석 (T검정이 2개이면, 분산분석은 3개 이상이라고 이해)

**3개 이상** 집단의 **평균비교** 할 때 사용된다.

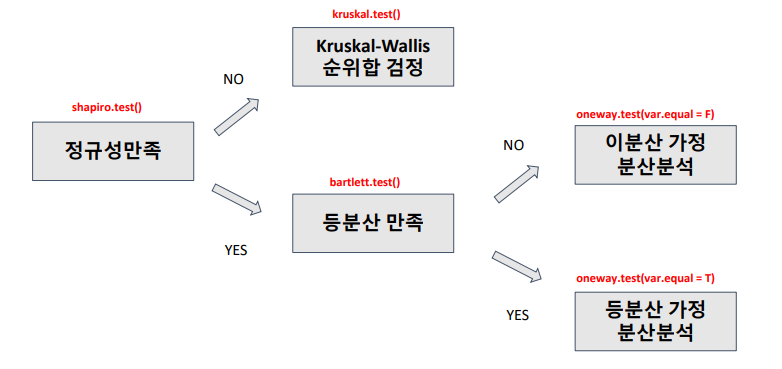
귀무가설 : 모든 집단들의 평균은 같다.

대립가설 : **적어도** 하나의 집단에서 평균에 차이가 난다.

기본 가정 2가지

1. 정규성 검정

2. 등분산성 검정

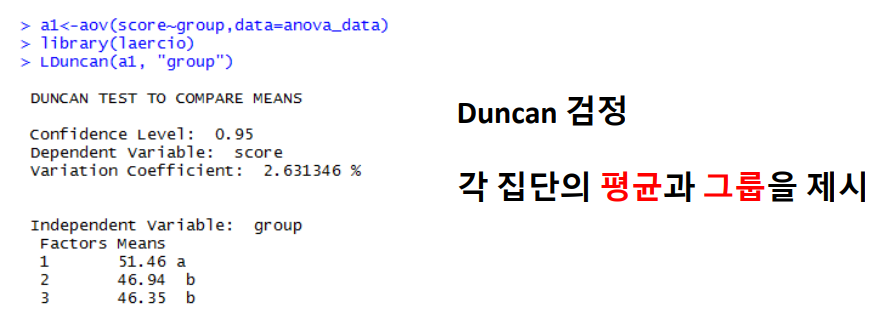


**## 사후분석**

2개의 집단이면 다르다고 판단 가능! 3개 이상의 집단이므로 **어디가** 차이나는지 확인이 필요!

LDuncan(a1, “group”)에서 a1 = 1번 factor, “group”이 2,3번 factor이다.

차이가 있다 없다는 제외하고 사후검정까지 해주는게 BEST이다.



성별, 혈액형이 한 판에 나오면 성별과 혈액형이 독립이 맞는지 아닌지에 대한 검정이 카이제곱 검정이고, x,y는 범주형이 그 범주형에 대한 빈도(수치)가 나오면 그 확률에 대해서 카이제곱 검정을 진행한다.

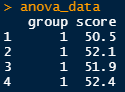
#3개이상의 평균비교 시 분산분석 - Anova(Analysis of Variance)

install.packages("laercio")

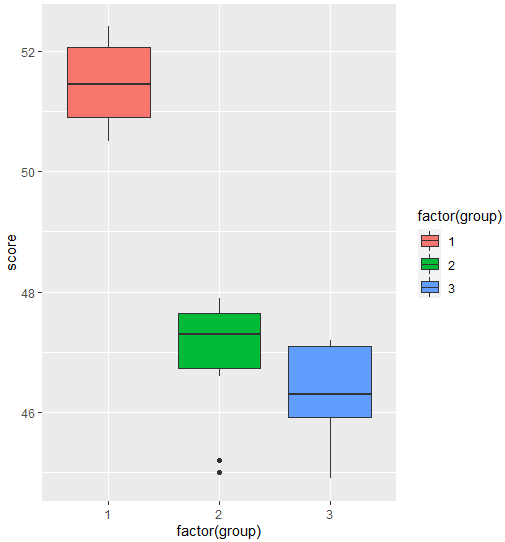
anova\_data<-data.frame(

group=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3),

score=c(50.5, 52.1, 51.9, 52.4, 50.6, 51.4, 51.2, 52.2, 51.5, 50.8,47.5, 47.7, 46.6, 47.1, 47.2, 47.8, 45.2, 47.4, 45.0, 47.9,46.0, 47.1, 45.6, 47.1, 47.2, 46.4, 45.9, 47.1, 44.9, 46.2))

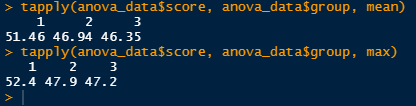


ggplot(anova\_data,aes(x=factor(group),y=score,fill=factor(group))) + geom\_boxplot()



tapply(anova\_data$score, anova\_data$group, mean)

tapply(anova\_data$score, anova\_data$group, max)



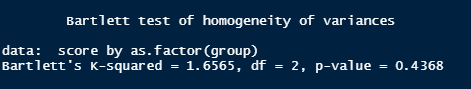
tapply() : 그룹마다의 mean, max를 구한다.

tapply(벡터, 요인, 함수)

#등분산성 test

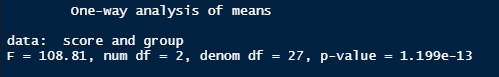
bartlett.test(score~as.factor(group),data=anova\_data)

3그룹 이상임으로 하나하나 작성하는 것이 아니라 그룹값에 따라 스코어 값이 차이가 나는지에 대해서 구한다를 위와 같이 작성한다.

****

0.05보다 높음으로 등분산이다. (t검정이랑 동일한데 사후검정이 있냐 없냐의 차이임)

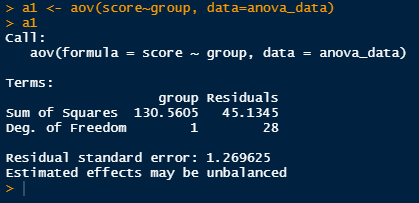
oneway.test(score~group,data=anova\_data,var.equal = T)



p-value가 지수형태로 나오는건 유의확률(0.05)보다 무조건 낮은 것을 의미함.

귀무가설 기각, 대립가설 채택(적어도 하나에는 차이가 있다.)

**## aov는 가설검정에 특화되어 있는 함수이다.**

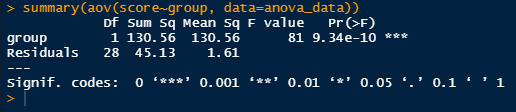


위와는 다르게 바로 분산분석표 형태를 볼 수 있다.

aov group에 따른 score를 aov로 넣을 것이다 라는 의미

summary(aov(score~group,data=anova\_data))

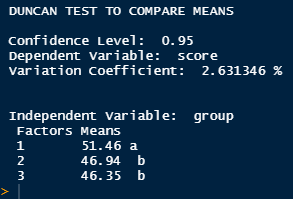
그룹내 그룹별로의 분산 분석표를 보고 싶으면 summary(aov(….)를 사용한다.

p-value(유의확률도 확인가능)

**#사후분석**

library(laercio)

LDuncan(a1, "group")



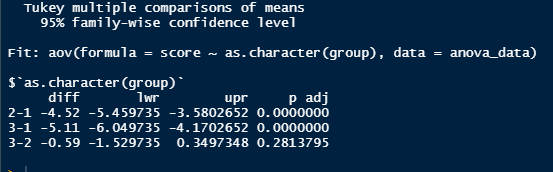
a1은 위에서 aov에서 검정한 부분이고 그걸 그룹별로 사후분석을 하겠다는 의미

LDuncan은 laercio 패키지 안에 포함된 함수이다.

**#group에 해당하는 부분이 문자형 이어야함 (**TukeyHSD)

TukeyHSD도 대표적인 사후검정의 함수이고, 그 이외에도 여러가지가 있다.

(각각의 차이가 그렇게 크지 않다)



Group 같은 경우에는 factor면 character형으로 바꿔서 사용해줘야 한다.

2-1, 3-1, 3-2로 각각의 p-value가 나온다.

2번과 1번은 차이가 **있네**, 3번과 1번은 차이가 **있네**, 3번과 2번은 차이가 **없네**.

**#### 등분산이 아닐경우 ####**

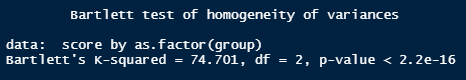
anova\_data2<-data.frame(

group=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3),

score=c(70, 30, 20.3, 85.3, 50.6, 51.4, 51.2, 52.2, 51.5, 50.8,47.5, 47.7, 46.6, 47.1, 47.2, 47.8, 45.2, 47.4, 45.0, 47.9,46.0, 47.1, 45.6, 47.1, 47.2, 46.4, 45.9, 47.1, 44.9, 46.2))

**#등분산성 test**

bartlett.test(score~as.factor(group),data=anova\_data2)



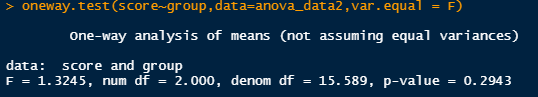
p-value가 지수형태임으로 등분산 형태가 아니다.

귀무가설이 기각이 안되면 사후분석은 사실 할 필요가 없다.

**#oneway.test**

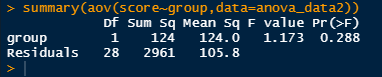
oneway.test(score~group,data=anova\_data2,var.equal = F)

등분산 형태가 아님으로 var.equal = F 적용



a2<-aov(score~group,data=anova\_data2)

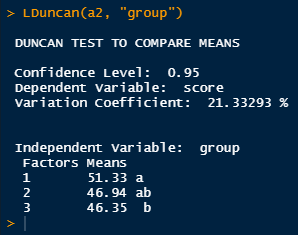
summary(aov(score~group,data=anova\_data2))



#사후분석

library(laercio)

LDuncan(a2, "group)



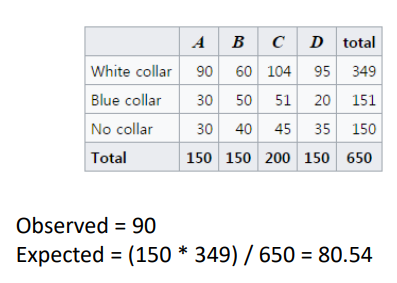
a와 b를 통계적으로 같이 보는 것 같음.

**## 범주형 통계분석**

**1. 카이제곱 (빈도 기반으로 검정)**

- N x M 교차표로 검정 가능

- **chisq.test**로 쉽게 적용가능



실제값 90, 기대값 80.54, 실제값과 기대값의 차이

## 카이제곱 2가지 검정

1) 적합도 검정 (Goodness of Fit Test) - 특정 확률을 따르고 있는지 확인

2) 독립성 검정 - 서로 연관이 있는지 없는지를 검정 (자주 사용)

두 요인이 연관이 있는지 독립인지 검정

귀무가설 : 두 범주형 변수 X,Y 는 독립 - 연관성이 없다.

대립가설 : 두 범주형 변수 X,Y는 독립이 아니다 - 연관성이 있다.

<통계의 대부분의 검정 종류>

X검정, F검정, 카이검정

**## Exact Test**

기대도수가 5 이하인 cell이 25% 이상일 때 사용 (2x2라면 1개라도 있을 때)

예) 55, 40, 30 ,**3** 일경우에는 카이제곱이 아니라 **fisher.test** 사용

카이제곱 통계량이 불안정할 수 있으므로 Fisher의 정확 검정(Fisher’s Exact test)를 사용해 연관성 검정

**fisher.test**로 쉽게 적용 가능

**## Trend Test**

독립변수가 **순위**가 있는 경우 사용

독립변수 순위가 증가함에 따라 종속변수의 비율 증가 및 감소에 대한 **경향성**을 확인하여 검증

- 귀무가설 : 종속변수의 비율이 동일하다(일정)

- 대립가설 : 종속변수의 비율이 동일하지 않다 (증가 / 감소 추세가 있다.)

**prop.trend.test()**로 쉽게 적용가능

예) ABCD / 빨주노초 (ABCD -> 1,2,3,4학년)

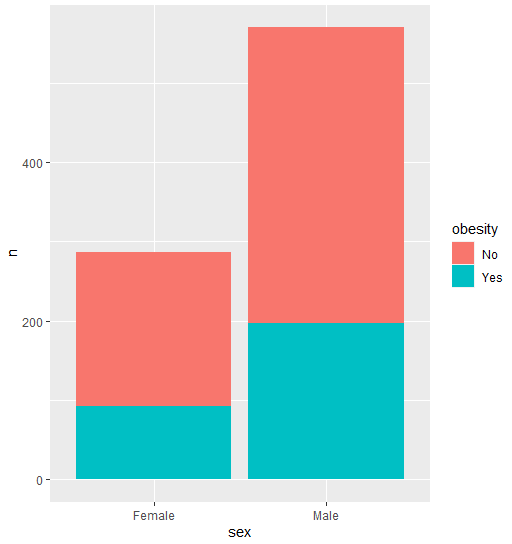
**#문자형 데이터분석**

**# 성별과 비만은 연관이 있을까?**

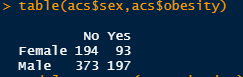
acs %>%

dplyr::count(sex, obesity) %>%

ggplot(aes(x=sex, y=n, fill=obesity)) + geom\_bar(stat='identity', postition = 'dodge')



table(acs$sex,acs$obesity)

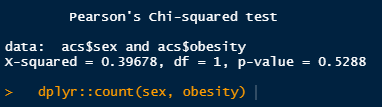


chisq.test(acs$sex,acs$obesity,correct = F)

귀무가설 : 두 범주형 변수 X,Y 는 독립 - 연관성이 없다.

0.5288은 귀무가설 -> 독립, 연관성이 없다.

Correct=T는 default임.

****

**## 어떤 통계학과가 범주형 데이터는 연속형이 아니라는 이의를 제기했고, 값에 -1/2를 넣고 소수점까지 넣고 연속형으로 만들자고 해서 correct=F이 생겨남**

**## correct=F 개념이 결과에 크게 영향을 주거나 하진 않는데 최근부터는 적용하기 시작했다.**

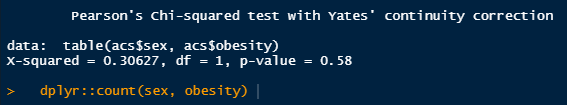
**# correct?**

**# 비 연속적 이항분포에서 확률이나 비율을 알기 위하여 연속적 분포인**

**# 카이제곱 분포를 이용할 떄는 연속성을 가지도록 비연속성을 교정해야할 필요하 있을 떄 사용하는 방법**

**# 보통 2X2 행렬에서 자주 사용함**

chisq.test(table(acs$sex,acs$obesity))

****

install.packages("gmodels")

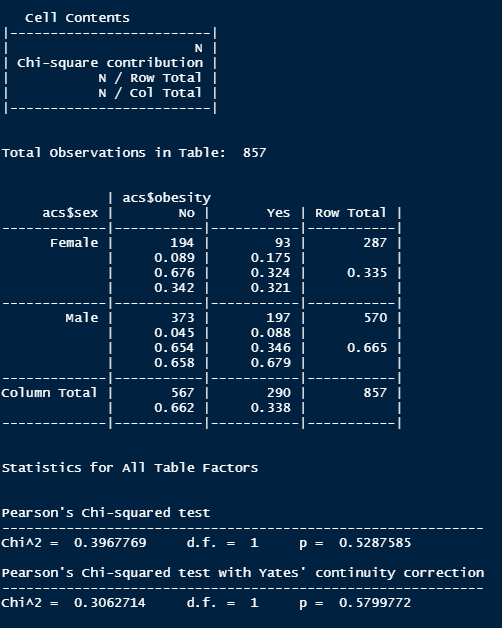
library(gmodels)

CrossTable(acs$sex,acs$obesity,chisq=T,prop.t=F)

CrossTable(table(acs$sex,acs$obesity))

chisq : 카이 제곱 검정을 수행할 때 True 사용 (chisq=T,F일 때 확인가능)

prop.t=F : 테이블 비율(proportion table)을 미포함



**# 일반횟수**

**# 카이 제곱 ( 기대치 비율 )**

**# 행을 기준으로 비율 값 ( 가로로 읽는다. )**

**# 컬럼을 기준으로 비율 값 ( 세로로 읽는다. )**

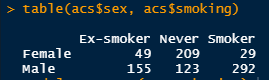
**# 전체를 기준으로 비율 값**

0.089 + 0.175 + 0.045 + 0.088 (카이제곱 검정값)



**# 성별과 비만은 연관이 있을까?**

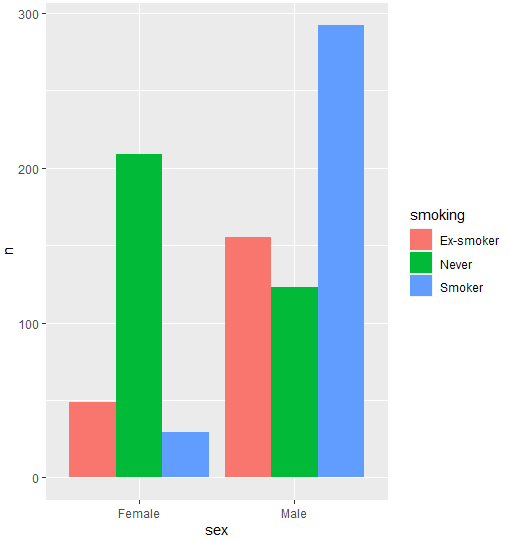
table(acs$sex, acs$smoking)



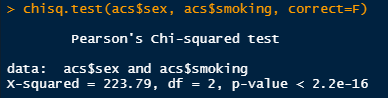
acs %>%

dplyr::count(sex,smoking) %>%

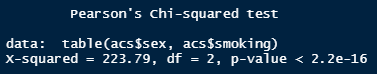
ggplot(aes(x=sex,y=n,fill=smoking)) + geom\_bar(stat="identity",position = "dodge")



chisq.test(acs$sex, acs$smoking, correct=F)



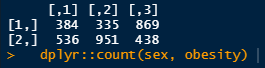
chisq.test(table(acs$sex,acs$smoking),correct = F)



둘다 지수 형태이기 때문에 성별에 따른 흡연이 **영향을 준다**

**##**

xtab <- matrix(c(384, 536, 335,951, 869, 438),nrow=2)

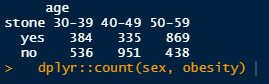


dimnames(xtab) <- list(

stone = c("yes", "no"),

age = c("30-39", "40-49", "50-59")

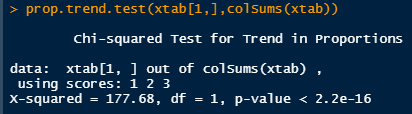
)



colSums(xtab)

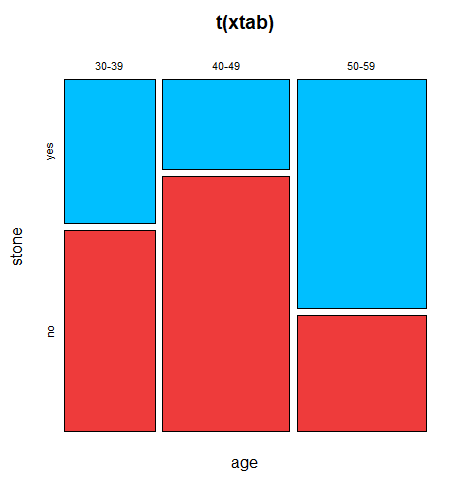


prop.trend.test(xtab[1,],colSums(xtab))



지수형태임으로 연관성이 있다. 영향을 끼친다.

mosaicplot(t(xtab),col=c("deepskyblue", "brown2"))



모자이크 플롯(mosaic plot)은 무엇일까요?

2원 3원 교차표의 시각화입니다. 따라서 전체 정사각 도형을 교차표의 행 빈도에 비례하는 직사각 도형으로 나누어 줍니다. 그 다음에, 다시 각도형을 행 내 열의 빈도에 해당하는 직사각 도형으로 나누어 줍니다. 위 그림을 참조해서 이해해보세요.