Distribution & Hypothesis Test

기초 통계량 - 평균, 분산, 표준편차

- 평균
  - 전체 데이터의 합을 데이터의 개수로 나눈 값
  - 편차와 분포를 반영하지 못하는 문제

- 중앙값(median)
  - 모든 데이터를 크기 순서로 정렬했을 때 가운데 위치한 값
  - 정도의 차이는 있으나 평균과 마찬가지로 분포를 반영하지 못하는 문제

- 분산
  - 각 데이터와 평균 사이의 편차를 제곱한 값의 평균
  - 편차에 음의 값이 존재하고 편차의 평균이 0이 되므로 제곱의 평균 사용

- 표준편차
  - 분산의 제곱근을 구한 값

R의 평균, 분산, 표준편차 관련 함수

```
> score <- c(85, 90, 93, 86, 82)

# 평균
> mean(score)
[1] 87.2

# 중앙값
> median(score)
[1] 86

# 분산
> var(score)
[1] 18.7

# 표준편차
> sd(score)
[1] 4.32435
```

표본의 분산, 표준편차를 계산할 때 모집단 개수 – 1을 분모로 사용

기초 통계량

- R의 4분위수, 빈도수 관련 함수

이름	설명
다섯 수치 요약	최소값, 제1사분위수, 중앙값, 제3사분위수, 최대값
IQR (Inter-Quarter Range)	3사분위수 - 1사분위수
최빈값	발생한 데이터 중 빈도수가 가장 높은 데이터

```
fivenum(1:10)
summary(1:10) #fivenum + average
x <- 1:10
c( min(x), quantile (x, 1/4) , median (x), quantile (x, 3/4) , max (x))
IQR (1:10)
quantile (1:10 , c(1/4, 3/4))

x <- factor (c("a", "b", "c", "c", "c", "d", "d"))
x
table (x)
which.max(table(x))
which.min(table(x))
```

기초 통계량

- R의 4분위수 관련 함수 테스트 결과

```
> fivenum(1:10)
[1]  1.0  3.0  5.5  8.0 10.0
> summary(1:10)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.25    5.50    5.50    7.75   10.00
> x <- 1:10
> c( min(x), quantile (x, 1/4) , median (x), quantile (x, 3/4) , max (x))
        25%         75%
 1.00  3.25  5.50  7.75 10.00
> IQR (1:10)
[1] 4.5
> quantile (1:10 , c(1/4, 3/4))
 25%  75%
3.25 7.75
>
> x <- factor (c("a", "b", "c", "c", "c", "d", "d"))
> x
[1] a b c c c d d
Levels: a b c d
> table (x)
x
a b c d
1 1 3 2
> which.max(table(x))
c
3
> which.min(table(x))
a
1
```

¼분위수 : 중앙 값보다 작은 값의 중앙 값
¾분위수 : 중앙 값보다 큰 값의 중앙 값

¼분위수 : min + (max – min) * ¼
¾분위수 : min + (max – min) * ¼

분할표 만들기

최대값의 위치

최소값의 위치

표본 추출

- 단순 임의 추출
  - 데이터, 표본크기, 복원추출여부, 가중치를 전달인자로 사용해서 표본 추출

```
sample(1:10, 5)
sample(1:10, replace = T)
sample(1:10, 5, replace = T, prob = 1:10)
sample(1:10) # shuffle
```

```
> sample(1:10, 5)
[1]  5 10  8  1  2
> sample(1:10, replace = T)
 [1] 10  4  6  2  3  7  3  5  7  7
> sample(1:10, 5, replace = T, prob = 1:10)
[1] 9 9 6 3 8
> sample(1:10) # shuffle
 [1]  4  3  5  9  7  1 10  2  6  8
```

표본 추출

- 층화 임의 추출
  - 데이터가 중첩 없이 구분될 수 있고 각 분할의 성격이 명확하게 다른 경우 사용
  - 예) 남성 20%, 여성 80%로 구성된 집단의 평균 키 측정

```
install.packages("sampling"); library(sampling);

x <- strata(c("Species"), size=c(3, 3, 3), method="srswor", data=iris)
x
getdata(iris, x)

#종별로 다른 수의 표본 추출
strata(c("Species"), size = c(3, 1, 1),  method = "srswr", data = iris)

iris$Species2 <- rep (1:2 , 75)
strata(c("Species", "Species2"), size =c(1, 1, 1, 1, 1, 1) ,
       method ="srswr", data = iris )
```

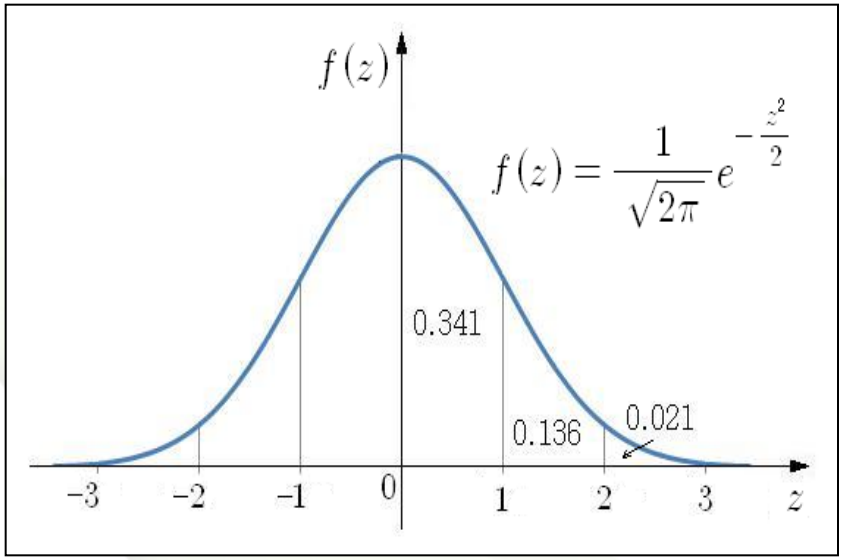
표본 추출

- 계통 추출
  - 모집단의 임의 위치에서 매 n번째 항목을 표본으로 추출
  - 데이터에 주기성이 있을 경우 편향된 표본 추출 위험

```
install.packages("doBy"); library(doBy);

#각 층마다 동일한 표본 추출
sampleBy(~Species , frac =.1 , data = iris, systematic = T)
```

표준 정규 분포

- -(무한대) ~ +(무한대) 까지의 모든 수치 데이터로 구성



- 데이터의 평균 값을 기준으로 좌/우 대칭형으로 분포 되어 있는 형태

- 상대도수는 수치에 따라 다르다
  - › 무한대의 데이터에 대해 도수 측정은 불가능하므로 상대도수를 사용
  - › 단일 데이터가 아닌 범위 데이터에 대한 상대도수(비율)을 주로 사용

- 평균값은 0, 표준편차는 1

- 표준편차의 1배 범위의 상대도수는 0.6826 (대략 70%), 2배 범위의 상대도수는 0.9544 (대략 95%) → 일반적으로 1.96배를 95%로 사용

일반 정규 분포

- 표준정규분포의 모든 데이터에 일정한 수(표준편차)를 곱한 후 일정한 수(평균)를 더한 데이터 분포

표준편차가 σ, 평균이 μ인 **정규분포는 → σ * 표준정규분포데이터 + μ**

- 일반정규분포 데이터의 분포 특성

$(\mu - 1 * \sigma) \sim (\mu + 1 * \sigma)$ 범위의 상대도수는 0.6826
$(\mu - 2 * \sigma) \sim (\mu + 2 * \sigma)$ 범위의 상대도수는 0.9544
$(\mu - 1.96 * \sigma) \sim (\mu + 1.96 * \sigma)$ 범위의 상대도수는 95%

- 일반정규분포 데이터를 표준정규분포로 변환

표준편차가 σ, 평균이 μ인 **정규분포 데이터에 대해**

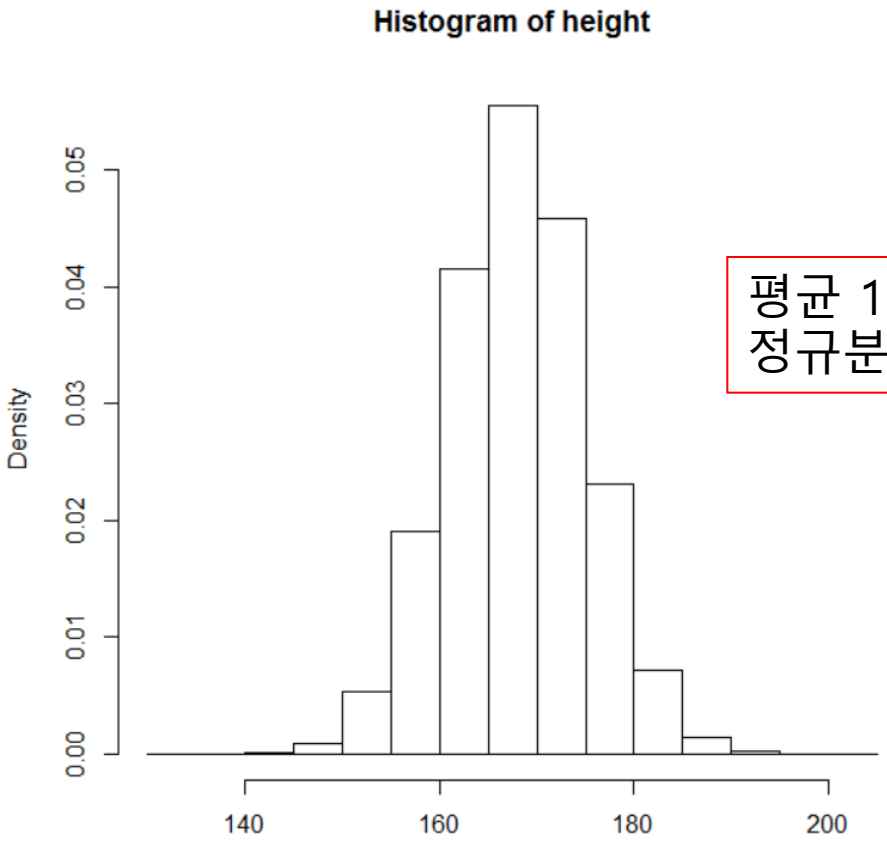
$$Z = \frac{X - \mu}{\sigma}$$

를 따르는 Z는 표준정규분포 데이터

정규 분포

- R에서 정규 분포 데이터 사용

```
> height <- rnorm(n = 1000000, mean = 168, sd = 7)
> hist(height, breaks = 10, probability = T)
```

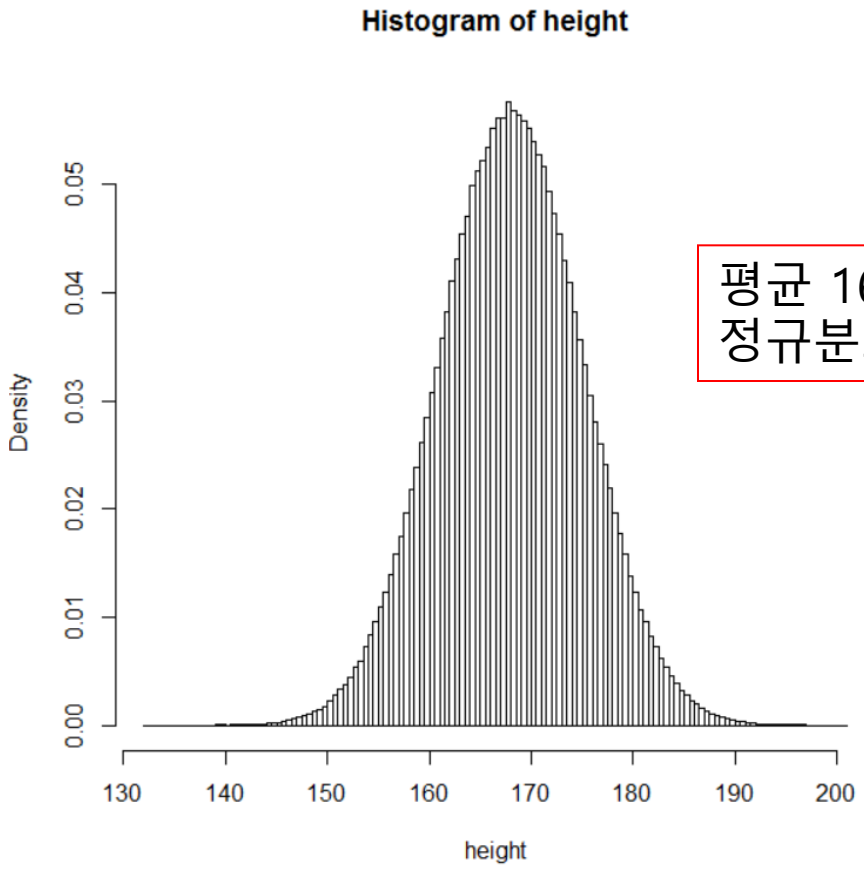


평균 168, 표준편차 7인 백만건의
정규분포 데이터 추출

정규 분포

- R에서 정규 분포 데이터 사용

```
> height <- rnorm(n = 1000000, mean = 168, sd = 7)
> hist(height, breaks = 100)
```

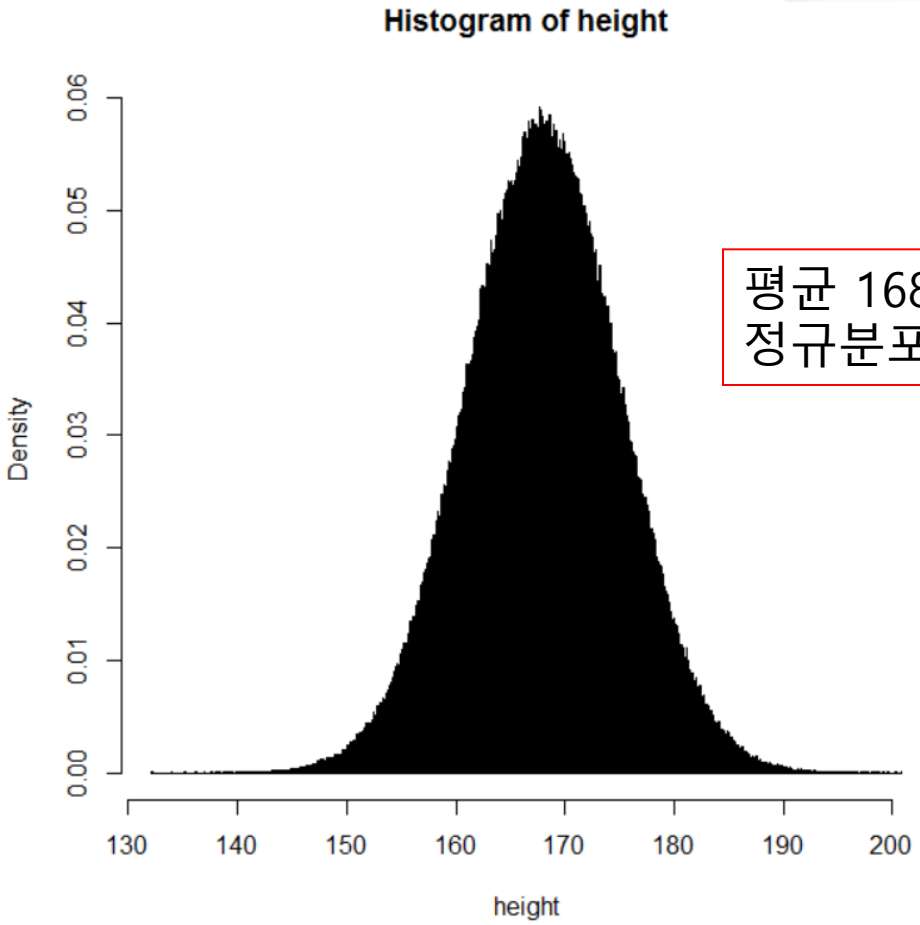


평균 168, 표준편차 7인 백만건의
정규분포 데이터 추출

정규 분포

- R에서 정규 분포 데이터 사용

```
> height <- rnorm(n = 1000000, mean = 168, sd = 7)
> hist(height, breaks = 1000)
```



평균 168, 표준편차 7인 백만건의
정규분포 데이터 추출

모집단과 표본

- 모집단
  - 조사 대상 데이터 전체
  - 유한 모집단 / 무한 모집단으로 구분 (일반적으로 무한대의 데이터 가정)

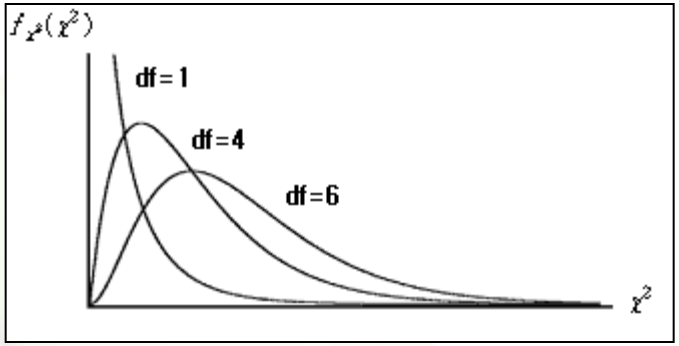
- 모평균, 모분산, 모표준편차
  - 모집단 데이터의 평균(μ), 분산(σ^2), 표준편차(σ)
  - 무한모집단과 같이 계산이 불가능한 경우 → 표본의 데이터로 추측

- 표본
  - 모집단에서 추출된 데이터
  - 표본의 크기가 커지면 표본 통계량이 모집단 통계량에 근접할 확률 증가
  - 표본평균($\bar{x}$), 표본분산(s^2), 표본표준편차(s)

카이제곱분포

- 표준정규분포 모집단에서 얻은 n개의 데이터에 대해 각각 제곱한 합으로 계산되는 통계량 V는 자유도 n인 카이제곱분포를 한다

$$V = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$$



- 자유도에 따른 각 데이터와 상대도수를 작성한 카이제곱표 제공

v \ α	0.995	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	0.00004	0.0002	0.001	0.004	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76

t 분포

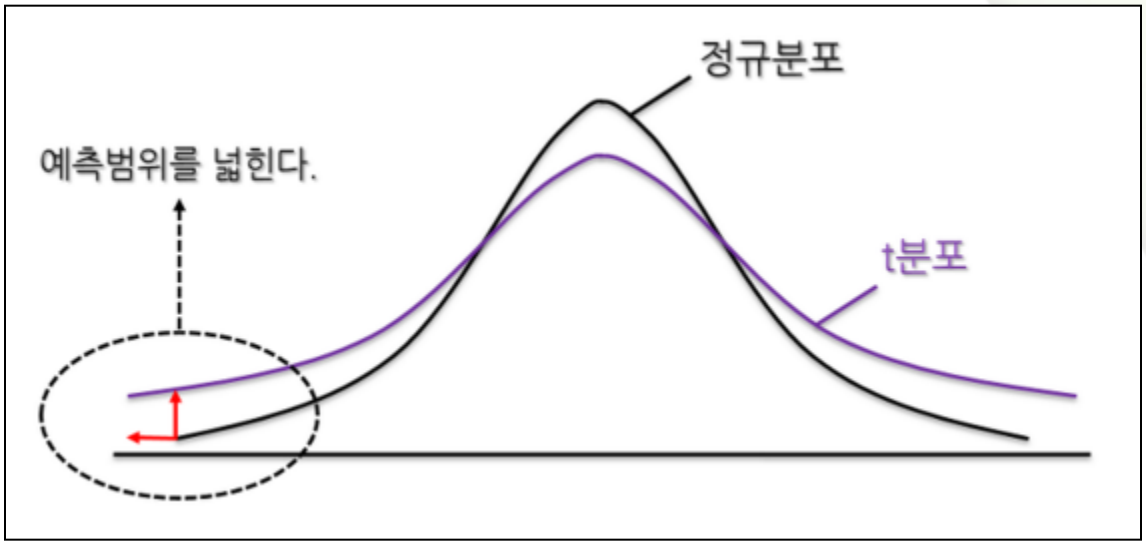
- 표본평균 , 표본표준편차 를 이용한 통계량 t 도출

$$T = \frac{(\bar{x} - \mu)\sqrt{n-1}}{s}$$

→ 자유도 (n - 1)인 t분포

→ $\frac{(\text{표준정규분포}) * \sqrt{\text{자유도}}}{\sqrt{\text{카이제곱분포}}}$

- T분포 히스토그램



- 정규모집단에서 추출된 표본 데이터의 통계량으로 모평균 추정 가능

상관 관계 분석

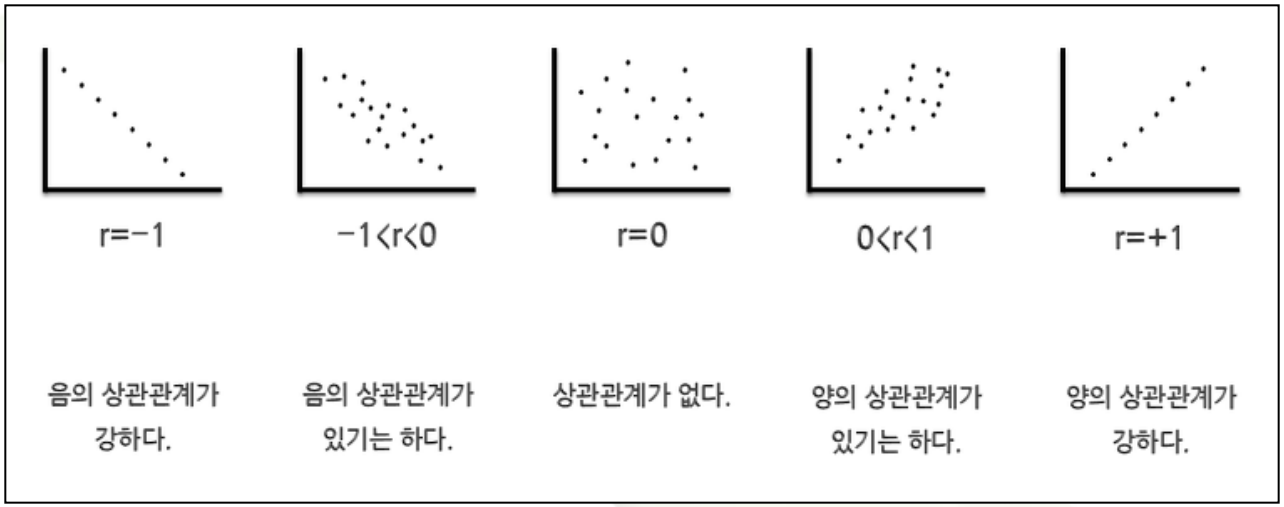
- 두 변수간에 선형적 관계가 있는 지 분석하는 방법.

- 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계(Correlation coefficient)라 한다.

- 상관분석에서는 상관관계의 정도를 나타내는 단위로 상관계수 r을 사용한다
  - 상관관계의 정도를 파악하는 상관계수(Correlation coefficient)는 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다.
  - 두 변수간에 원인과 결과의 인과관계가 있는지에 대한 것은 회귀분석을 통해 인과관계의 방향, 정도와 수학적 모델을 확인해 볼 수 있다.
  - $0 < r \leq +1$ 이면 양의 상관, $-1 \leq r < 0$ 이면 음의 상관, $r = 0$이면 무상관 이라고 한다.
  - 하지만 0인 경우 상관이 없다는 것이 아니라 선형의 상관관계가 아니라는 것이다.

R 상관 관계 분석

- 상관관계



- 상관 계수 공식

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

- 상관 관계 분석 실행

```
> x <- c(3, 5, 8, 11, 13)  # 제품 판매 개수
> y <- c(1, 2, 3, 4, 5)    # DM 발송 회수
> cor(x, y)
[1] 0.9970545
```

상관계수

- 두 확률 변수 사이의 관계를 파악하는 방식
  - › 일반적으로 피어슨 상관계수를 의미

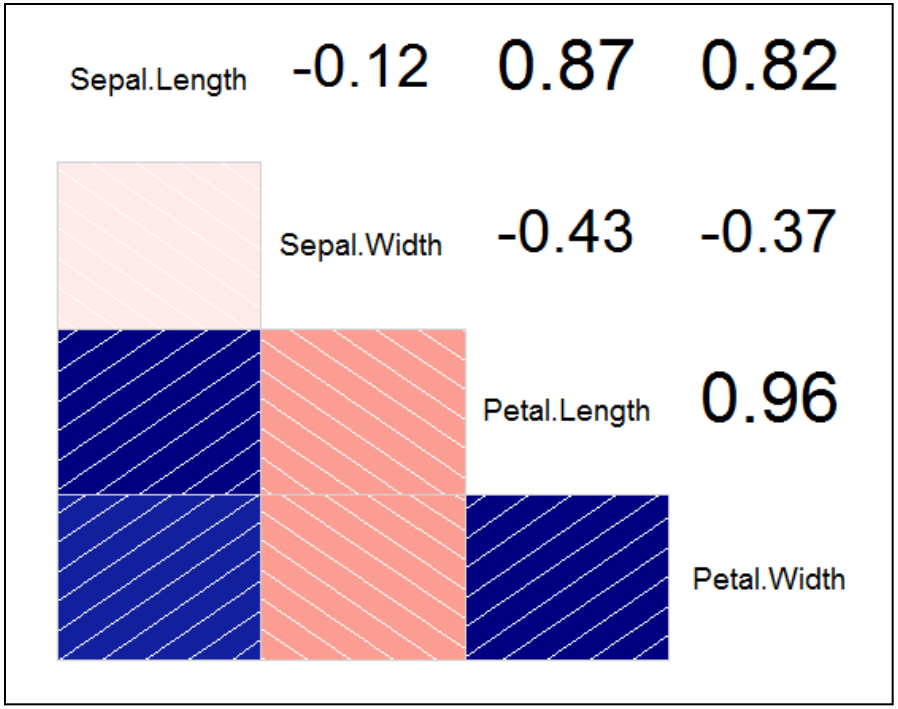
```
cor(iris$Sepal.Width, iris$Sepal.Length)

cor(iris[, 1:4])
```

```
> cor(iris$Sepal.Width, iris$Sepal.Length)
[1] -0.1175698
> cor(iris[, 1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

상관계수

- corrgram 패키지
  - 상관계수를 시각화하는데 유용한 패키지

```
install.packages("corrgram")
library(corrgram)

corrgram(cor(iris[, 1:4]), type = "corr", upper.panel = panel.conf)
```



상관계수

- 스피어만 상관계수
  - 상관계수를 계산할 두 데이터의 실제 값 대신 두 값의 순위를 사용해 상관계수 비교
  - 이산형 데이터 및 순서형 데이터 적용 가능
  - 예)국어점수-영어점수 관계는 피어슨 / 국어석차-영어석차는 스피어만

```
install.packages("Hmisc")
library(Hmisc)

x <- c(3, 4, 5, 3, 2, 1, 7, 5)
rank(sort(x))

m <- matrix(c(1:10, (1:10)^2), ncol = 2)
m

rcorr(m, type = "pearson")$r
rcorr(m, type = "spearman")$r
```

```
> rcorr(m, type = "pearson")$r
          [,1]      [,2]
[1,] 1.0000000 0.9745587
[2,] 0.9745587 1.0000000
> rcorr(m, type = "spearman")$r
     [,1] [,2]
[1,]    1    1
[2,]    1    1
```

상관계수

- 켄달의 순위 상관계수
  - (X, Y) 형태의 순서쌍 데이터에 대해 x1 < x2 에 대해 y1 < y2가 성립하면 concordant, 성립하지 않으면 discordant라고 정의

```
install.packages("Kendall")
library(Kendall)

Kendall(c(1,2,3, 4, 5), c(1, 0, 3, 4, 5))
```

```
> Kendall(c(1,2,3, 4, 5), c(1, 0, 3, 4, 5))
tau = 0.8, 2-sided pvalue =0.086411
```

귀무 가설과 대립 가설

- 귀무 가설 (null hypothesis) 또는 영가설
  - 처음부터 버릴 것을 예상하는 가설
  - 차이가 없거나 의미 있는 차이가 없는 경우의 가설
  - 모집단의 모수(평균, 표준편차 등)가 귀무 가설에서의 값과 같다는 가설

- 대립 가설
  - 모집단의 모수가 귀무 가설에서의 값보다 크거나 작거나 혹은 다르다는 가설
  - 대립 가설은 단측 또는 양측
    - 단측 : 모집단 모수가 귀무 가설에서의 값과 특정 방향으로 다른지 여부를 확인하기 위한 가설
    - 양측 : 모집단 모수가 귀무 가설에서의 값보다 크거나 작은지 여부를 확인하기 위한 가설

귀무 가설과 대립 가설

- 대립 가설은 단측 또는 양측
  - 양측
    - 모집단 모수가 귀무 가설에서의 값과 특정 방향으로 다른지 여부를 확인하기 위한 가설

> 한 연구자가 국가 시험을 치른 특정 고등학교 학생들의 시험 결과를 갖고 있습니다. 연구자는 해당 학교의 점수가 국가 평균 850점과 다른지 여부를 확인하려고 합니다. 점수가 국가 평균과 다른지 여부를 확인하려고 하기 때문에 양측 대립 가설(방향성이 없는 가설이라고도 함)이 적절합니다. (H_0: $\mu = 850$ vs. H_1: $\mu \neq 850$)

  - 단측
    - 모집단 모수가 귀무 가설에서의 값보다 크거나 작은지 여부를 확인하기 위한 가설

> 한 연구자가 국가 시험을 준비하는 교육 과정에 참여한 학생들의 시험 결과를 갖고 있습니다. 연구자는 교육을 받은 학생들의 점수가 국가 평균인 850점보다 높은지 여부를 확인하려고 합니다. 교육을 받은 학생들의 점수가 국가 평균보다 높다는 특정한 가설을 세웠기 때문에 단측 대립 가설(방향 가설이라고도 함)을 사용할 수 있습니다. (H_0: $\mu = 850$ vs. H_1: $\mu > 850$)

상관계수 검정

- corr.test() 함수를 사용해서 상관 계수의 유의성을 판단할 수 있다

- 귀무 가설은 상관계수가 0인 가설

```
cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method = "pearson")

cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method = "spearman")

cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method = "kendall")
```

결과는 다음 페이지에

상관계수

- 상관계수 검정

```
> cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method = "pearson")

        Pearson's product-moment correlation

data:  c(1, 2, 3, 4, 5) and c(1, 0, 3, 4, 5)
t = 3.9279, df = 3, p-value = 0.02937
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1697938 0.9944622
sample estimates:
      cor
0.9149914

> cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method = "spearman")

        Spearman's rank correlation rho

data:  c(1, 2, 3, 4, 5) and c(1, 0, 3, 4, 5)
S = 2, p-value = 0.08333
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9

> cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method = "kendall")

        Kendall's rank correlation tau

data:  c(1, 2, 3, 4, 5) and c(1, 0, 3, 4, 5)
T = 9, p-value = 0.08333
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.8
```

교차 분석

- 명목 및 서열척도의 범주형 변수들을 분석하기 위한 것으로써 두 변수간의 독립성과 관련성을 분석.

- 한 변수의 범주를 다른 변수의 범주에 따라 빈도를 교차 분류하는 두 개 이상의 행과 열을 갖는 교차표(crosstabs table)를 작성

- 단순히 교차 빈도를 집계할 뿐만 아니라 두 변수 간의 독립성 여부를 판정하는 독립성 검정(검정 ; Chi-square test) 수행 가능.

분할표

- 명목형 또는 순서형 자료의 도수를 표 형태로 기록한 것

- 분할표가 작성되면 Chi Square Test로 확률 변수간의 의존관계를 살표보는 독립성 검정 및 도수가 특정 분포를 따르는지에 대한 적합도 검정 수행 가능

분할표

- table() 함수를 사용해서 분할표 작성

```
table (c("a", "b", "b", "b", "c", "c", "d"))
```

```
> table (c("a", "b", "b", "b", "c", "c", "d"))

a b c d
1 3 2 1
```

- xtabs() 함수를 사용해서 분할표 작성
  - › Formula 사용해서 데이터 지정 가능 (도수 ~ 분류1 + 분류2 + … + 분류n)

```
d <- data.frame (x=c("1", "2", "2", "1"),
                 y=c("A", "B", "A", "B"),
                 num=c(3, 5, 8, 7))
d

xt <- xtabs(num ~ x + y, data = d)
xt
```

```
> d <- data.frame (x=c("1", "2", "2", "1"),
+                  y=c("A", "B", "A", "B"),
+                  num=c(3, 5, 8, 7))
> d
  x y num
1 1 A   3
2 2 B   5
3 2 A   8
4 1 B   7
> xt <- xtabs(num ~ x + y, data = d)
> xt
   y
x    A B
  1 3 7
  2 8 5
```

분할표

- margin.table() 함수를 사용해서 행/열의 빈도수 계산

```
margin.table(xt, 1)
margin.table(xt, 2)
```

```
> xt
   y
x   A B
  1 3 7
  2 8 5
> margin.table(xt, 1)
x
 1  2
10 13
> margin.table(xt, 2)
y
 A  B
11 12
```

- prop.table() 함수를 사용해서 각 셀의 비율 계산

```
prop.table(xt, 1)
prop.table(xt, 2)
prop.table(xt)
```

```
> prop.table(xt, 1)
   y
x         A         B
  1 0.3000000 0.7000000
  2 0.6153846 0.3846154
> prop.table(xt, 2)
   y
x         A         B
  1 0.2727273 0.5833333
  2 0.7272727 0.4166667
> prop.table(xt)
   y
x         A         B
  1 0.1304348 0.3043478
  2 0.3478261 0.2173913
```

검정

- 독립성 검정
  - 데이터 사이의 독립성 여부 검정
  - 분할표를 작성한 경우 행에 나열된 속성과 열에 나열된 속성이 독립인지 검정

- 적합도 검정
  - 데이터가 특정 분포를 따른다는 가정을 검정

독립성 검정 사례 1

- gmodels 패키지 설치

```
> install.packages("gmodels")
> library(gmodels)
```

- 데이터 읽기 및 변수 저장

```
> study <- read.csv("data-files/pass_cross.csv", header = T)
> study
 X      X.1  X.2    X.3
1  공부함 공부안함 합격 불합격
2       1        0    1      0
3       1        0    1      0
4       0        1    0      1
5       0        1    0      1
6       1        0    1      0
7       1        0    1      0
8       0        1    1      0
9       0        1    1      0
10      0        1    1      0
.......... (이하 생략)
```

독립성 검정 사례 1

- 교차분석 실행 및 결과

```
> CrossTable(study$공부함, study$합격, chisq = T)


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

독립성 검정 사례 1

- 교차분석 실행 및 결과

Total Observations in Table: 50

study$공부함	study$합격 0	study$합격 1	Row Total
0	13	12	25
	0.900	0.600	
	0.520	0.480	0.500
	0.650	0.400	
	0.260	0.240	
1	7	18	25
	0.900	0.600	
	0.280	0.720	0.500
	0.350	0.600	
	0.140	0.360	
Column Total	20	30	50
	0.400	0.600	

독립성 검정 사례 1

- 교차분석 실행 및 결과

```
Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3     d.f. =  1     p =  0.08326452

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  2.083333     d.f. =  1     p =  0.1489147
```

독립성 검정 사례 2

- 독립성 검정에 Chi-Squared Test 사용

```
library(MASS)
data(survey)
str(survey)
head(survey)
head(survey[,c("Sex", "Exer")])

xtabs(~ Sex + Exer, data = survey)

chisq.test(xtabs(~Sex + Exer, data = survey))
```

```
> chisq.test(xtabs(~Sex + Exer, data = survey))

        Pearson's Chi-squared test

data:  xtabs(~Sex + Exer, data = survey)
X-squared = 5.7184, df = 2, p-value = 0.05731
```

독립성 검정

- 샘플 수가 작은 경우 chisq.test()는 경고 메시지 표시
  - › 예) 기대 빈도가 5보다 작은 셀이 전체의 20% 이상인 경우 등

```
xtabs(~W.Hnd + Clap, data = survey)
chisq.test(xtabs(~W.Hnd + Clap, data = survey))
```

```
> chisq.test(xtabs(~W.Hnd + Clap, data = survey))

        Pearson's Chi-squared test

data:  xtabs(~W.Hnd + Clap, data = survey)
X-squared = 19.252, df = 2, p-value = 6.598e-05

Warning message:
In chisq.test(xtabs(~W.Hnd + Clap, data = survey)) :
  카이제곱 approximation은 정확하지 않을수도 있습니다
```

독립성 검정

- 피셔의 정확 검정
  - › 카이 제곱 검정이 빈도수 부족 등을 이유로 문제 되는 경우 사용

```
xtabs(~W.Hnd + Clap, data = survey)
fisher.test(xtabs(~ W.Hnd + Clap, data = survey))
```

```
> fisher.test(xtabs(~ W.Hnd + Clap, data = survey))

        Fisher's Exact Test for Count Data

data:  xtabs(~W.Hnd + Clap, data = survey)
p-value = 0.0001413
alternative hypothesis: two.sided
```

독립성 검정

- 맥니마 검정
  - 응답자의 성향이 사건 전, 후에 어떻게 변화하는지 알아보는 등의 경우에 사용

```
help(mcnemar.test)

Performance <-
  matrix(c(794, 86, 150, 570),
         nrow = 2,
         dimnames = list("1st Survey" = c("Approve", "Disapprove"),
                         "2nd Survey" = c("Approve", "Disapprove")))
Performance
mcnemar.test(Performance)
```

```
> mcnemar.test(Performance)

        McNemar's Chi-squared test with continuity correction

data:  Performance
McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05
```

적합도 검정

- Chi Square Test를 사용해서 데이터가 특정 분포를 따르는지 검정
  - › 독립성 검정에 비해 검정 기준으로 데이터의 빈도수가 아닌 분포를 사용

```
table(survey$W.Hnd)
chisq.test(table(survey$W.Hnd), p = c(.3, .7))
```

```
> chisq.test(table(survey$W.Hnd), p = c(.3, .7))

        Chi-squared test for given probabilities

data:  table(survey$W.Hnd)
X-squared = 56.252, df = 1, p-value = 6.376e-14
```

적합도 검정

- binomial test
  - › binom.test(성공횟수, 전체횟수, 확률)

```
binom.test(86, 86 + 150, 0.5)
```

```
> binom.test(86, 86 + 150, 0.5)

        Exact binomial test

data:  86 and 86 + 150
number of successes = 86, number of trials = 236, p-value = 3.716e-05
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3029404 0.4293268
sample estimates:
probability of success
             0.3644068
```

적합도 검정

- Shapiro-Wilk Test
  - › 표본이 정규분포로부터 추출된 것인지 테스트하기 위한 방법

```
shapiro.test(rnorm(1000))
```

```
> shapiro.test(rnorm(1000))

        Shapiro-Wilk normality test

data:  rnorm(1000)
W = 0.99874, p-value = 0.7138
```

T 분포 검정

- 연속 확률 분포 및 표본 분포 → 정규 분포와 유사한 분포
- 정규분포는 표본의 데이터 수가 많아야 신뢰도가 향상되는 단점
- 예측 범위가 넓은 분포를 사용해서 정규 분포의 문제점에 대응한 것이 T 분포
  - 주로 30개 미만의 표본에 대해 적용

추정 및 검정 - 일표본 평균

- T 검정을 사용해서 일표본 평균 추정 및 검정

```
x <- rnorm(30)

t.test(x)
```

```
One Sample t-test

data:  x
t = -0.71532, df = 29, p-value = 0.4801
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.5874137  0.2829896
sample estimates:
mean of x
-0.152212
```

귀무가설 채택

- 데이터가 정규분포를 따르는지 불명확한 경우 Shapiro-Wilk Test 또는 Q-Q Plot 등을 사용해서 분포에 대한 적합도 검정 필요

추정 및 검정 - 일표본 평균

- T 검정을 사용해서 일표본 평균 추정 및 검정

```
> score1 <- read.csv( " data-files/tdata.csv", header = T)
> score1
   번호 성적
1     1   77
2     2   85
3     3   63
...........(이하 생략)

> result <- t.test(score1$성적, alternative = c("greater"), mu = 75)
> result

        One Sample t-test

data:  score1$성적
t = 0.68948, df = 9, p-value = 0.254
alternative hypothesis: true mean is greater than 75
95 percent confidence interval:
 71.51677      Inf
sample estimates:
mean of x
     77.1
```

귀무가설 채택

추정 및 검정 - 일표본 평균

- T 검정을 사용해서 일표본 평균 추정 및 검정

```
> score2 <- read.csv("data-files/tdata2.csv", header = T)
> score2
   번호 성적
1     1   85
2     2   95
3     3   75
...........(이하 생략)

> result2 <- t.test(score2$성적, alternative = c("greater"), mu = 75)
> result2

        One Sample t-test

data:  score2$성적
t = 3.9386, df = 9, p-value = 0.001707
alternative hypothesis: true mean is greater than 75
95 percent confidence interval:
 80.88039      Inf
sample estimates:
mean of x
       86
```

대립가설 채택

추정 및 검정 - 독립 이표본 평균

- 수면제별 수면 시간 증가량 비교

```
> sleep
# 결과 생략

> sleep2 <- sleep[, -3]
> sleep2
#결과 생략

> tapply(sleep2$extra, sleep2$group, mean)
   1    2
0.75 2.33

> library(doBy)
> summaryBy(extra ~ group, sleep2)
  group extra.mean
1     1       0.75
2     2       2.33
```

추정 및 검정 - 독립 이표본 평균

- 두 그룹(수면제별)의 모분산이 동일한지 검정 (등분산성 검정)

```
> var.test(extra ~ group, sleep2)

        F test to compare two variances

data:  extra by group
F = 0.79834, num df = 9, denom df = 9, p-value = 0.7427
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.198297 3.214123
sample estimates:
ratio of variances
         0.7983426
```

귀무가설 채택

› 귀무가설 채택으로 두 그룹의 등분산성 확인

추정 및 검정 - 독립 이표본 평균

- 두 그룹(수면제별) 수면 시간 증가량 평균에 차이가 있는지 T 검정 함수를 사용해서 확인

```
> t.test(extra ~ group, data = sleep2, paired = FALSE, var.equal = TRUE)

        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

› 수면 시간 증가량에 차이가 없음을 확인

추정 및 검정 - 독립 이표본 평균

- 두 그룹(수면제별) 수면 시간 증가량 평균에 차이가 있는지 F 검정 함수를 사용해서 확인

```
> oneway.test(extra ~ group, data = sleep2, var.equal = TRUE)

        One-way analysis of means

data:  extra and group
F = 3.4626, num df = 1, denom df = 18, p-value = 0.07919
```

› 수면시간 증가량 평균에 차이가 없음을 확인

추정 및 검정 – 짝지은 이표본 평균

- 두 개의 표본이 순서짱처럼 구해진 경우
  - 예를 들어 약물 섭취 전과 후의 데이터를 측정해 (x, y) 형태로 구성되는 경우

```
> sleep

> sleep$extra[sleep$group == 1]
 [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0

> sleep$extra[sleep$group == 2]
 [1]  1.9  0.8  1.1  0.1 -0.1  4.4  5.5  1.6  4.6  3.4
```

추정 및 검정 - 짝지은 이표본 평균

```
> with(sleep, t.test(extra[group == 1], extra[group == 2], paired =
TRUE))

        Paired t-test

data:  extra[group == 1] and extra[group == 2]
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

대립가설 채택

- 두 수면제의 수면시간 연장 정도가 다르다는 결론 도출

추정 및 검정 - 일표본 비율

- 비율은 prop.test 함수를 사용

```
> prop.test(42, 100)

        1-sample proportions test with continuity correction

data:  42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3233236 0.5228954
sample estimates:
   p
0.42
```

› 동전의 앞면이 나올 확률이 50%라는 귀무가설을 기각 할 수 없음

추정 및 검정 - 이표본 비율

- 비율은 prop.test 함수를 사용
  - 두 개의 동전을 각각 100회, 90회 던졌을 경우 앞면이 45회, 55회 나온 경우 두 동전의 앞면이 나올 확률이 동일한지 검정

```
> prop.test(c(45, 55), c(100, 90))

        2-sample test for equality of proportions with continuity
correction

data:  c(45, 55) out of c(100, 90)
X-squared = 4.3067, df = 1, p-value = 0.03796
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.31185005 -0.01037217
sample estimates:
   prop 1    prop 2
0.4500000 0.6111111
```

  - 두 동전의 앞면이 나올 확률은 같지 않음을 확인

추정 및 검정 - 일원배치 분산 분석

- 설명변수가 하나이고 2개 이상의 그룹으로 구분되는 경우 oneway.test, aov, anova 함수 등을 사용해서 검정 (특히 3개 이상)

- t 검정은 2개 그룹의 비교에만 사용 가능

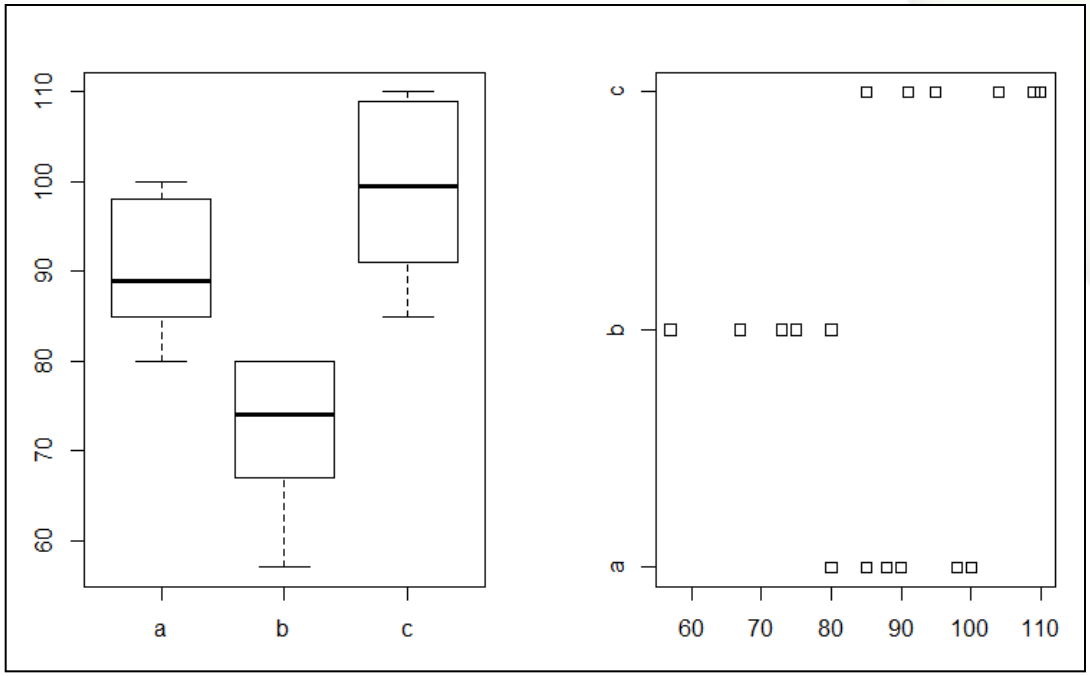
- 사례
  - 데이터 준비 (세 종류의 건전지의 수명 데이터를 가정)

```
a <- c(100, 98, 85, 90, 88, 80)
b <- c(73, 80, 80, 75, 67, 57)
c <- c(110, 104, 91, 109, 85, 95)

life <- data.frame(a, b, c)
life
#결과 생략
b.life <- stack(life)
b.life
#결과 생략
```

추정 및 검정 - 일원배치 분산 분석

- 사례 계속
  - › 데이터 시각화

```
op = par(mfrow = c(1, 2))
boxplot(values ~ ind, data = b.life)
stripchart(life)
par(op)
```



추정 및 검정 - 일원배치 분산 분석

- 사례 계속
  - › oneway.test 함수를 사용한 검정

```
> oneway.test(values ~ ind, data = b.life, var.equal = TRUE)

        One-way analysis of means

data:  values and ind
F = 14.18, num df = 2, denom df = 15, p-value = 0.0003488
```

  - › aov 함수를 사용한 검정

```
> b.life.aov <- aov(values ~ ind, data = b.life)
> summary(b.life.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
ind          2   2274  1137.1   14.18 0.000349 ***
Residuals   15   1203    80.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

추정 및 검정 - 일원배치 분산 분석

- 사례 계속
  - anova 함수를 사용한 검정

```
> b.life.lm <- lm(values ~ ind, data = b.life)
> anova(b.life.lm)
Analysis of Variance Table

Response: values
          Df Sum Sq Mean Sq F value    Pr(>F)
ind        2 2274.1 1137.06   14.18 0.0003488 ***
Residuals 15 1202.8   80.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  - 그룹이 3개 이상인 경우 T 검정은 실패

```
> t.test(values ~ ind, data = b.life, paired = FALSE, var.equal = TRUE)
Error in t.test.formula(values ~ ind, data = b.life, paired = FALSE,
var.equal = TRUE) :
  grouping factor는 반드시 2개의 수준을 가지고 있어야만 합니다
```

추정 및 검정 – 일원배치 분산 분석

- 사례 계속
  - › TukeyHSD 함수를 사용해서 그룹간 차이를 확인

```
> b.life.tukey <- TukeyHSD(b.life.aov)
> life.tukey
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = y ~ ty)

$ty
         diff       lwr      upr     p adj
a-b 18.166667  4.737562 31.59577 0.0082805
c-b 27.000000 13.570896 40.42910 0.0002867
c-a  8.833333 -4.595771 22.26244 0.2344084
```

  - › c – a 그룹간에는 차이가 없고 a – b, c – b 그룹간의 차이에 의해 전체 그룹간 차이가 발생

추정 및 검정 - 이원분산분석

- 설명 변수가 두 개인 경우의 분석

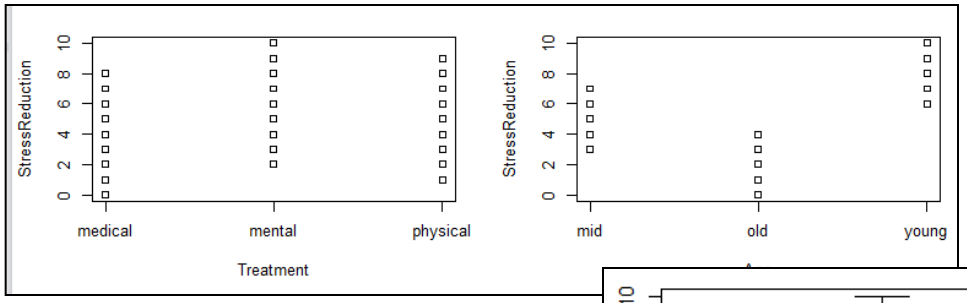
- 사례
  - 데이터 준비 (나이, 처치 방법과 스트레스 감소량)

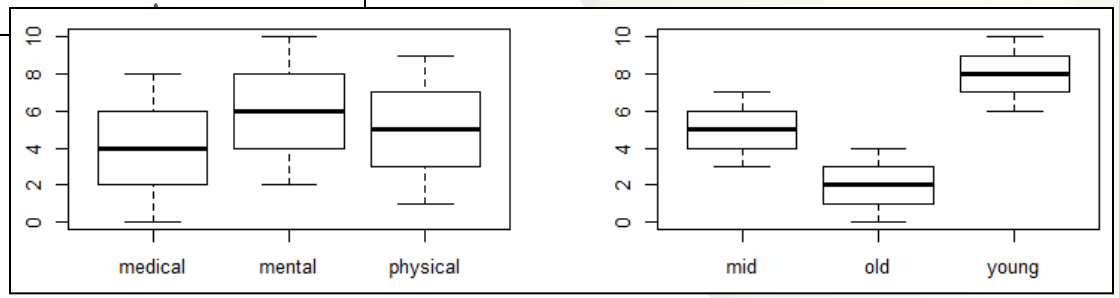
```
twoway.comparisons.data <- read.csv('data-files/twoway-
comparisons.csv')

print(twoway.comparisons.data)

# 결과 생략
```

추정 및 검정 - 이원분산분석

- 사례 계속
  - › 데이터 시각화

```
op <- par(mfrow = c(2, 2))
with(twoway.comparisons.data, boxplot(StressReduction ~ Treatment))
with(twoway.comparisons.data, boxplot(StressReduction ~ Age))

with(twoway.comparisons.data, stripchart(StressReduction ~ Treatment, vertical =
TRUE, xlab = 'Treatment'))
with(twoway.comparisons.data, stripchart(StressReduction ~ Age, vertical = TRUE,
xlab = 'Age'))
par(op)
```





추정 및 검정 - 이원분산분석

- 사례 계속
  - aov 함수를 사용해서 검정

```
> towway.comparisons.data.aov <- aov(StressReduction ~ Treatment * Age,
twoway.comparisons.data)
> summary(towway.comparisons.data.aov)
              Df Sum Sq Mean Sq F value  Pr(>F)
Treatment      2     18       9       9 0.00195 **
Age            2    162      81      81   1e-09 ***
Treatment:Age  4      0       0       0 1.00000
Residuals     18     18       1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

교호작용평가

  - Age와 Treatment 를 기준으로 하는 그룹간에 차이가 존재하는 것으로 평가

추정 및 검정 - 이원분산분석

- 사례 계속
  - lm과 anova 함수를 사용해서 검정

```
> twoway.comparisons.data.lm <- lm(StressReduction ~ Treatment * Age,
twoway.comparisons.data)
> anova(twoway.comparisons.data.lm)
Analysis of Variance Table

Response: StressReduction
               Df Sum Sq Mean Sq F value   Pr(>F)
Treatment       2     18       9       9 0.001953 **
Age             2    162      81      81    1e-09 ***
Treatment:Age   4      0       0       0 1.000000
Residuals      18     18       1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

교호작용평가

  - Age와 Treatment 를 기준으로 하는 그룹간에 차이가 존재하는 것으로 평가