R을 활용한 의학통계 방법론[1]

김충락[2], 김진미[3]

2024-02-13

[1]본 도서는 BK21 미래인재양성사업팀 “빅 데이터 통계분석”의 2차년도 (2021. 3. 1 - 2022. 2. 28) 사업비에 의해 지원 받았음.

[2]부산대학교 통계학과 명예교수

[3]부산대학교병원 의생명연구원 연구교수

차 례

그림 차례

표 차례

제1부 기초통계이론

제 1 장

의학연구를 위한 통계적 방법

이 책에서는 의학연구에서 발생하는 다양한 형태의 자료를 분석하는 통계적 방법론을 소개하고, 실제 자료를 분석하는 과정에서 R이라는 통계패키지를 활용하는 방법을 제시한다. R에 수록되어 있는 실제 자료를 분석하는데 사용되는 여러 가지 통계적 방법론을 소개한 후 분석된 예제를 R 코드와 함께 output을 제시하고 그 결과를 해석한다.

1.1 모집단과 표본

COVID19 백신을 개발하는 과정에서 반드시 필요한 것이 임상실험이다. 이 실험에서 개발된 백신이 항체를 형성할 확률을 제시해야 한다. 이 경우 항체를 형성할 확률의 정확한 값은 모든 인류를 대상으로 했을 경우에만 계산할 수 있으나 이는 불가능하다. 따라서, 전 인류의 일부분인 사람들을 대상으로 임상실험을 한다. 이처럼 관심의 대상이 되는 전체 집단을 모집단 (population)이라 하며, 이 모집단으로부터 임의로 추출한 부분 집합을 표본 (sample)이라 한다. 한편, 이런 표본을 추출하는 행위를 표본추출 (sampling)이라 하며 모집단의 특성을 나타내는 값을 흔히 모수 (parameter)라 부른다.

우리의 최종 관심은 모수가 어떤 값인지 추정하는 것이다. 이를 위해 표본을 추출하고 표본에 의거하여 여러 가지 통계적 방법을 적용한 후 모수의 값에 대해 추론하는 일련의 과정이 통계학이다. 이 과정에서 자연스럽게 발생하는 의문이 있다. (1) 과연 표본을 어떻게 추출해야 하며 (2) 표본의 크기는 얼마나 되어야 하는가? (3) 이렇게 내린 결론에 대해 어느 정도의 신뢰를

가질 수 있는가? 등이다.

또 다른 예로서 간암에 걸린 환자의 생존시간에 영향을 미치는 요인 (변수)을 밝히는 것은 매우 중요한 일이다. 간 이식, 환부의 절제 등 치료 방법뿐만 아니라 전이 여부, 암의 진행 정도, 나이, 유전적 요인 등 다양한 요인이 있을 수 있으나 생존시간에 미치는 모든 요인을 전부 고려하는 것은 불가능하다. 즉, 특정 환자의 생존 시간이 모수라면 이를 정확하게 계산하는 것이 불가능하므로 현재 주어진 자료를 이용하여 그 환자의 생존 시간을 최대한 정확하게 예측하기 위해 적절한 통계적 방법이 필요하다.

1.2 자료

1.2.1 자료의 종류

자료는 크게 양적 (quantitative, numerical)자료와 질적 (qualitative)자료로 구분한다. 먼저, 양적자료는 말 그대로 자료 그 자체가 숫자와 일대일로 대응하는 것인데 그 성질에 따라 연속형 자료와 이산형 자료로 구분된다. 연속형자료는 혈압, 몸무게 등 일정구간의 실수값을 모두 취할 수 있는 자료이고, 이산형 자료는 어느 학급의 여학생 수, 연간 결혼 건수 등 정수값을 취하는 자료이다.
반면, 질적자료는 자료 그 자체가 숫자의 개념을 가지는 것이 아니라 구분하는 개념을 가지므로 범주형 (categorical) 자료라고도 한다. 질적자료는 크게 명목형 (nominal)자료와 순서형 (ordinal)자료로 나눈다. 명목형자료는 성별, 피부색 등 단지 구분을 위한 것으로 성별의 경우 남자를 1, 여자를 0으로 변환하거나 남자를 -1, 여자를 +1로 변환할 수 있는데 이는 단지 구분을 위해 숫자를 대응시킬 뿐이다. 한편, 순서형 자료는 범주들이 순서의 개념을 가지는 것으로 예를 들어, 작업 성취도를 상, 중, 하로 평가할 때 1,2,3으로 코딩하거나 3,2,1 등으로 코딩하는 것이 바람직하며, 1,3,2 나 2,1,3 등 순서의 개념을 무시하고 코딩하면 곤란하다.

1.2.2 기호

흔히 자료는 여러 개의 변수 (variable)로 구성되어 있으며 각 변수들은 반응변수 (response) 또는 설명변수 (covariates)로 구성되어 있다. 반응변수는 흔히 Y라는 기호로 나타내고, 설명변수 (공변량 이라고도 함)는 p개가 있을 경우 $X_1, \cdots, X_p$로 나타낸다. i번째 설명변수 X_i가 n번 관측되었을 때 $x_{i1}, \cdots, x_{in}$으로 나타내며, 혼동의 여지가 없는 경우 특정 변수 X에

대한 관측치 역시 $x_1, \cdots, x_n$으로 나타낸다. 즉, 특정 변수를 지칭할 경우 대문자로 표시하고, 어떤 변수가 관측된 값을 표시할 경우 소문자를 사용하여 구분한다. 예를 들어 y_i는 변수 Y의 i번째 관측치를 의미한다.

1.3 R

1.3.1 R의 소개 및 설치

자료를 분석하기 위해서 필요한 여러 가지 계산을 도와주는 소프트웨어를 통계패키지(statistical package)라고 한다. 자료분석을 위한 통계패키지는 여러 종류가 개발되어 있으나 대표적인 몇 가지만 소개하면 다음과 같다.

SAS (Statistical Analysis System), MINITAB, BMDP (BioMeDical Package), SPSS (Statistical Package for Social Science)

이외에도 여러 종류의 통계패키지가 있으나 대부분 사용료를 지불해야만 이용할 수 있다. 그러나 이제 소개하는 R이란 통계패키지는 매우 효율적이면서 무료로 사용할 수 있으므로 본 강좌에서는 R을 사용할 것이며 이에 대한 사용법을 간략히 소개한다.

R이란 무엇인가? R은 다음과 같은 특징이 있다.

(1) open source program
(2) S (programming language) => S plus => R
(3) 그래픽 기능이 매우 뛰어남
(4) 5,000개가 넘는 패키지 (일종의 앱)가 탑재되어 있으며 계속 새로운 패키지가 탑재됨

R을 사용하기 위해서는 www.r-project.org 에 접속하여 다음의 절차로 다운받으면 된다.

(1) go to the site called CRAN (Comprehensive R Archive Networks)
http://www.r-project.org/
(2) execute “Download R“
(3) choose Korea http://cran.nexr.com/

(4) click “Download for R Windows“

(5) click “base“

(6) click “Download R 4.3.2 for Windows”

1.3.2 R 사용법

A. 주의할 점

(1) case sensitive

(2) commands are separated by ; or newline

(3) comments can be put anywhere starting with #

(4) subsequent commands are made by +

B. 내장 기능 (Inbuilt Facilities)

(1) help, example, demo

help(solve) 또는 ?solve # solve 라는 명령어 사용법에 대한 설명#

example(solve) # solve 라는 명령어에 대한 예제#

demo(persp) # persp 라는 명령어에 대한 예시#

(2) Data

data() # 내장되어 있는 자료파일을 불러올 수 있음#

```
- women (height, weight, n=15)
- stackloss (Air.Flow, Water.Temp, Acid.Conc., stack.loss, n=21)
- faithful (eruptions, waiting, n=272)
- sleep (extra, group, n=20)
```

(3) Libraries

- Some useful libraries in R
  - lattice : lattice graphics
  - MASS : Modern Applied Statistics using S-Plus
  - mgcv : generalized additive models
  - nlme : mixed effects models
  - nnet : neural networks and multinomial log-linear models
  - spatial : spatial statistics
  - survival : survival analysis
- To see contents of “survival” library, for example, type
  library(help=survival)

(4) Packages

You can install packages using “install packages”.

You have to open “library” to use packages.

e.g.) packages -> install packages -> choice of country -> download “lars” -> library(lars) -> ?lars

(5) data editing

To use a “bacteria” dataset in the “MASS” library, type

```
library(MASS); attach(bacteria); bacteria
```

C. Simple Manipulations : Numbers and Vectors

(1) Vectors and assignment

x <- c(10.4, 5.6, 3.1, 6.4, 21.7)

assign(“x”, c(10.4, 5.6, 3.1, 6.4, 21.7))

c(10.4, 5.6, 3.1, 6.4, 21.7) -> x

1/x

y <- c(x, 0, x); y

(2) Vector arithmetic

v <- 2*x + y + 1 ; v

15/7 : real

15%/%7 : integer part

15%%7 : remainder part

sum((x-mean(x))^2)/(length(x)-1)

var(x)

sqrt(-17) : NaN (not a number)

(3) Generating regular sequences

1:30 ; c(1:30); t <-c(1:30)

s3 <- seq(-5, 5, by=.2); s3

s4 <- seq(length=51, from=-5, by=.2) ; s4

s5 <- rep(x, times=5); s5

s6 <- rep(x, each=5); s6

1.3.3 R Studio

R을 보다 효과적이고 편리하게 사용할 수 있도록 도와주는 GUI (Graphical User Interface) 프로그램으로 www.rstudio.com 에서 다운받아 사용할 수 있다.

1.3.4 사용할 자료

이 책에서 사용하게 될 자료는 다음과 같은 4가지로서 모두 R에 수록되어 있다. 각 자료의 자료명, 출처, 자료를 구성하고 있는 각 변수에 대한 설명, 실제 자료의 일부, 그리고 관심있는 분석의 대상 및 내용이 소개되어 있으므로 자세히 읽은 후 각 자료에 대해 충분히 이해하기 바란다.

(I)피마 인디언 여자의 당뇨병 (Diabetes in Pima Indian Women)

(1) **출처** : R의 “MASS” library ; file name “Pima.te”

(2) **자료의 설명** : 미국 아리조나 주 피닉스 인근의 인디언 주거지역에 거주하는 21세 이상의 여자를 대상으로 세계보건기구 (WHO)의 기준에 따라 당뇨병에 대한 역학 조사를 실시하였다. 이 자료는 US National Institute of Diabetes and Digestive and Kidney Diseases에 의해 수집되었는데 532명에 대한 자료 중 332명에 대한 것이며, 100명의 훈련자료 (training data)는 Pima.tr에 있고, 여기에 일부 변수의 기록이 누락된 자료 (missing data) 100개를 합한 것은 Pima.tr2에 수록되어 있다. 이 자료는 모두 8개의 변수로 구성되어 있고 각 변수에 대한 코드 및 설명은 다음과 같다.

variable	label
npreg	number of pregnancies (임신 횟수)
glu	plasma glucose concentration in an oral glucose tolerance test (플라즈마 글루코스 농도 - 혈당수치)
bp	diastolic blood pressure (mm Hg) (이완기 혈압)
skin	triceps skin fold thickness (mm) (삼두근 두께)
bmi	body mass index (weight in kg/(height in m)^2) (체질량지수)
ped	diabetes pedigree function (당뇨병 가계력 산술식)
age	age in years (나이)
type	Yes or No, for diabetic according to WHO criteria (당뇨병 유무)

```
library(MASS)
data(Pima.te)
str(Pima.te)
```

```
#> 'data.frame':    332 obs. of  8 variables:
#>  $ npreg: int  6 1 1 3 2 5 0 1 3 9 ...
#>  $ glu  : int  148 85 89 78 197 166 118 103 126 119 ...
#>  $ bp   : int  72 66 66 50 70 72 84 30 88 80 ...
#>  $ skin : int  35 29 23 32 45 19 47 38 41 35 ...
#>  $ bmi  : num  33.6 26.6 28.1 31 30.5 25.8 45.8 43.3 39.3 29 ...
#>  $ ped  : num  0.627 0.351 0.167 0.248 0.158 0.587 0.551 0.183 0.704 0.263 ...
#>  $ age  : int  50 31 21 26 53 51 31 33 27 29 ...
#>  $ type : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 1 1 2 ...
```

(3) **분석할 내용**

- 당뇨병에 걸린 집단과 그렇지 않은 집단 간에 임신횟수, 혈당수치, 이완기 혈압, 삼두근 두께, 체질량 지수, 당뇨병 가계력 산술식, 나이 등은 차이가 있는가? (two-sample t-test)
- 혈당수치가 100이상인 집단과 100미만인 집단간에 당뇨병 유병율은 같은가? (가설검정, p-value, 신뢰구간, 2x2 분할표 검정 등)
- 체질량 지수에 영향을 주는 변수들은 어떤 것들이 있으며 그 관계식은 무엇인가? (선형회귀, training data and test data)
- 당뇨병 유무에 대한 모형의 구축 (로지스틱 회귀, SVM, k-NN)
- 당뇨병 유뮤에 영향을 미치는 변수들을 찾아서 만든 로지스틱 회귀식을 이용하여 어떤 사람이 당뇨병에 걸릴 것인가에 대한 예측을 할 경우 예측력은 얼마나 정확한가? (false positive and false negative, sensitivity and specificity, ROC curve and AUC)

(II) 폐암자료 (NCCTG Lung Cancer Data)

(1) **출처** : R의 “survival” library; file name “lung”

(2) **자료의 설명** : 미국 NCCTG (North Central Cancer Treatment Group)에서 수집된 228명의 폐암환자의 생존 시간 및 일상생활에 대한 자료로서 모두 8개의 변수로 구성되어 있고 각 변수에 대한 코드 및 설명은 다음과 같다.

variable	label
inst	Institution code (병원 코드)
time	Survival time in days (생존시간 (단위: 일))
status	censoring status 1=censored, 2=dead (중도절단 여부)
age	Age in years (나이)
sex	Male=1 Female=2 (성별)
ph.ecog	performance score - ECOG (Eastern Cooperative Oncology Group) 이 개발한 암환자의 현 상황을 0부터 5까지의 정수값으로 나타낸 것으로 다음과 같이 정의한다. 0 : 정상인처럼 아무런 제약없이 활동함 1 : 병의 징후는 있으나 다른 도움 없이 일상적 일을 할 수 있음 2 : 스스로 움직일 수 있으나 일상적 일을 할 수 없음 3 : 제한적으로 움직일 수 있으나 침상에서 반이상의 시간을 보내야 함 4 : 스스로 움직일 수 없으며 침상이나 의자에서만 생활이 가능함 5 : 사망
ph.karno	Karnofsky performance score (bad=0-good=100) rated by physician (의사가 판단하는 카노프스키 점수)
pat.karno	Karnofsky performance score as rated by patient (환자가 판단하는 카노프스키 점수)
meal.cal	Calories consumed at meals (식사 칼로리 양)
wt.loss	Weight loss in last six months (지난 6개월간 체중감소량)

```
library(survival)
data(cancer, package="survival")
str(lung)
#> 'data.frame':    228 obs. of  10 variables:
#>  $ inst     : num  3 3 3 5 1 12 7 11 1 7 ...
#>  $ time     : num  306 455 1010 210 883 ...
#>  $ status   : num  2 2 1 2 2 1 2 2 2 2 ...
#>  $ age      : num  74 68 56 57 60 74 68 71 53 61 ...
#>  $ sex      : num  1 1 1 1 1 1 2 2 1 1 ...
#>  $ ph.ecog  : num  1 0 0 1 0 1 2 2 1 2 ...
#>  $ ph.karno : num  90 90 90 90 100 50 70 60 70 70 ...
#>  $ pat.karno: num  100 90 90 60 90 80 60 80 80 70 ...
#>  $ meal.cal : num  1175 1225 NA 1150 NA ...
#>  $ wt.loss  : num  NA 15 15 11 0 0 10 1 16 34 ...
```

(3) **분석할 내용**

- 폐암환자의 연령은 어떤 분포를 나타내는가? (히스토그램, stem-and-leaf plot, 커널 밀도함수, box plot 등)
- 의사가 판단하는 카노프스키 점수와 환자가 판단하는 카노프스키 점수간에 상관관계는 어느 정도인가? (피어슨 상관계수 및 검정, 스피어만 상관계수 및 검정)
- 폐암환자의 생존함수에 대한 추정 (카플란-마이어)
- 폐암환자의 생존시간이 성별에 따라 다른가? (로그-순위 검정)
- 폐암환자의 생존시간에 영향을 미치는 변수는 무엇인가? (칵스회귀모형, 변수의 선택)
- performance score (단계)를 반응변수로 한 회귀모형의 구축 (proportional odds model)
- 폐암환자의 생존시간 예측을 위한 머신러닝 모형 (칵스회귀, SVR, random forest, boosting, concordance index의 소개 및 의미)

(III) 덴마크 악성 흑색종 (Malignant Melanoma in Denmark)

(1) **출처**: R의 “MASS” library ; file name “Melanoma”

(2) **자료의 설명** : 덴마크에서 발생한 205명의 악성 흑색종 환자의 생존시간에 대한 자료로서 7개의 변수로 구성되어 있으며 각 변수에 대한 코드 및 설명은 다음과 같다.

variable	label
time	survival time in days, possibly censored
status	1 died from melanoma, 2 alive, 3 dead from other causes
sex	1 = male, 0 = female
age	age in years
year	of operation
thickness	tumour thickness in mm
ulcer	1 = presence, 0 = absence

```
library(MASS)
data("Melanoma")
str(Melanoma)
#> 'data.frame':    205 obs. of  7 variables:
#>  $ time     : int  10 30 35 99 185 204 210 232 232 279 ...
#>  $ status   : int  3 3 2 3 1 1 1 3 1 1 ...
#>  $ sex      : int  1 1 1 0 1 1 1 0 1 0 ...
#>  $ age      : int  76 56 41 71 52 28 77 60 49 68 ...
#>  $ year     : int  1972 1968 1977 1968 1965 1971 1972 1974 1968 1971 ...
#>  $ thickness: num  6.76 0.65 1.34 2.9 12.08 ...
#>  $ ulcer    : int  1 0 0 0 1 1 1 1 1 1 ...
```

(3) **분석할 내용**

- 흑색종 환자의 생존함수 추정 (카플란-마이어 추정치)

- 흑색종 환자의 생존시간 예측을 위한 머신러닝 모형 (칵스회귀, SVR, random forest, boosting, concordance index의 소개 및 의미)
- 궤양 (ulcer) 여부에 따른 생존함수의 차이에 대한 검정 (로그-순위 검정)

(IV) 전립선암 환자자료 (Prostate Cancer Data)

(1) **출처**: R의 "lasso2" library ; file name "Prostate"

(2) **자료의 설명** : 근치 전립선 절제술을받을 예정인 남성 97명의 전립선 특이 항원 수준과 여러 임상 측정치 사이의 상관 관계를 조사한 연구로서 9개의 변수로 구성되어 있으며 각 변수에 대한 코드 및 설명은 다음과 같다.

variable	label
lcavol	log(cancer volume), 암부피의 로그값
lweight	log(prostate weight), 전립선 무게의 로그값
age	age (나이)
lbph	log(benign prostatic hyperplasia amount), 양성 전립선 세포의 증식량의 로그값
svi	seminal vesicle invasion (정낭 침입)
lcp	log(capsular penetration), 정낭 침범정도의 로그값
gleason	Gleason score (글리슨 수치)
pgg45	percentage Gleason scores 4 or 5 (글리슨 수치가 4 또는 5의 퍼센트)
lpsa	log(prostate specific antigen), 전립선 항원 수준의 로그값

```
library(lasso2)
data(Prostate)
str(Prostate)
#> 'data.frame':    97 obs. of  9 variables:
#>  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
#>  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
#>  $ age    : num  50 58 74 58 62 50 64 58 47 63 ...
#>  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
```

```
#>  $ svi    : num  0 0 0 0 0 0 0 0 0 0 ...
#>  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
#>  $ gleason: num  6 6 7 6 6 6 6 6 6 6 ...
#>  $ pgg45  : num  0 0 20 0 0 0 0 0 0 0 ...
#>  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

(3) **분석할 내용**

- lcavol을 반응변수, 나머지 변수들을 설명변수로 취급하여 선형회귀모형을 실시한다. (multiple linear regression, rigde regression, lasso regression, variable selection)

제 2 장

기술통계

의학연구의 최종 목적은 흔히 여러 변수들 간의 관련성을 탐색해 내는 것이다. 이러한 목적을 달성하기 위해 적절한 통계 분석 도구를 바로 적용하기 전에 각 변수의 특색 (평균, 표준편차, 분포, 이상치의 존재 여부 등)을 살피는 예비분석 (preliminary analysis)이 반드시 필요하다. 이러한 과정은 흔히 기술통계학 (descriptive statistic)이라 불리며 다음과 같은 다양한 도구들이 있다.

2.1 중심 및 산포 측도

2.1.1 중심 측도

n개의 관측치 $x_1, \cdots, x_n$에 대한 중심의 측도 (measure of center)로는 다음과 같은 것들이 있다. 먼저, 표본평균 (sample mean)은

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

으로 나타낸다. 표본 중간값 (sample median)은 흔히 Q_2로 나타내며 이는 자료를 작은 값부터 큰 값으로 나열했을 때 가운데 위치한 값이다. 만약 자료의 개수가 홀수이면 $(n+1)/2$번째 값으로 주어지지만 자료의 개수가 짝수인 경우는 $n/2$번째와 $(n/2)+1$번째 자료의

평균을 사용한다. 자료를 작은 값부터 큰 값으로 나열했을 때 하위 25%에 위치한 값을 제1사분위수 (1st quartile)라하고 흔히 Q_1으로 나타내며, 하위 75% (상위 25%)에 위치한 값을 제3사분위수 (3rd quartile)라하고 흔히 Q_3으로 나타낸다. 같은 논리로 중간값은 제2사분위수 (2nd quartile)과 같다. 일반적으로 하위 $100 \times p\%, (0 < p < 1)$에 해당되는 값을 $100 \times p\%$백분율 (percentile)이라 부른다.

중심의 측도로 평균을 많이 사용하는데 다음과 같은 주의를 요한다. 예를 들어, 11개의 관측치로 구성된 자료가 1, 3, 4, 6, 6, 7, 8, 8, 9, 10, 15으로 주어져 있을 때 평균은 (1+3+4+6+6+7+8+8+9+10+15)/11 = 7 이 되고, 중간값은 11개 관측치의 6번째 값인 7이 된다. 만약 관측치 중에서 가장 큰 값인 15가 150으로 바뀌게 되면 중간값은 변하지 않지만, 평균은 19.27로 커지게 된다. 즉, 평균은 이상치 (outlier)에 대해 민감 (sensitive)하지만 중간값은 이상치에 거의 영향을 받지 않게 된다 (robust). 따라서, 자료에 이상치가 있을 경우 자료의 중심을 나타내는 값으로 평균보다 중간값이 더 좋은 측도라 할 수 있다.

2.1.2 산포 측도

n개의 관측치 $x_1, \cdots, x_n$에 대한 산포의 측도 (measure of dispersion)는 자료가 중심에서 얼마나 많이 퍼져 있는지를 나타내는 측도로서 다음과 같은 것들이 있다. 먼저, 표본분산 (sample variance)은

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

으로 나타낸다. 표본표준편차 (sample standard deviation)는 표본분산의 제곱근인 $s = \sqrt{s^2}$이다. 그 외에 산포에 대한 측도로는 표본범위 (sample range)가 있는데

$$R = max - min$$

으로 주어지고 사분위수범위 (interquartile range)는

$$IQR = Q_3 – Q_1$$

으로 정의되며 이는 제3사분위수에서 제1사분위수를 뺀 값이다.

예제 : 피마 인디언 자료 (Pima.te)에서 이완기 혈압 (bp)에 대한 중심 및 산포를 구해보자.

변수명 X에 있는 자료에 대한 중심 및 산포에 대한 R 명령어는 `summary(X)` 또는 `fivenum(X)` 를 이용하며 `fivenum(X)`은 자료의 min, Q_1, Q_2, Q_3, max의 5가지 값을 나타낸다. `summary(X)`는 `fivenum(X)`에서 표본평균이 하나 더 추가된다. `DescTools` 패키지의 `Desc(X)`는 다양한 기술통계량과 그래프를 제공한다.

```
fivenum(Pima.te$bp)
#> [1]  24  64  72  80 110
summary(Pima.te$bp)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   24.00   64.00   72.00   71.65   80.00  110.00
DescTools::Desc(Pima.te$bp)
#> ------------------------------------------------------------------------------
#> Pima.te$bp (integer)
#>
#>   length       n    NAs  unique     0s   mean  meanCI'
#>      332     332      0      36      0  71.65   70.27
#>           100.0%   0.0%           0.0%          73.04
#>
#>      .05     .10    .25  median    .75    .90     .95
#>    50.00   56.20  64.00   72.00  80.00  88.00   90.00
#>
#>    range      sd  vcoef     mad    IQR   skew    kurt
#>    86.00   12.80   0.18   11.86  16.00  -0.08    0.81
#>
#> lowest : 24, 30 (2), 44 (4), 46 (2), 50 (10)
#> highest: 100 (3), 104, 106, 108, 110 (2)
#>
#> heap(?): remarkable frequency (8.1%) for the mode(s) (= 70)
#>
#> ' 95%-CI (classic)
```

2.2 자료의 분포

자료의 분포는 주로 그림이나 그래프로 나타내며, 히스토그램 (histogram), 커널밀도함수 (kernel density function), 줄기-잎 그림 (stem-and-leaf-plot), 상자그림 (box plot) 등이 있다. 앞에서 사용했던 피마 인디언 자료 (Pima.te)에서 이완기 혈압 (bp)에 대한 분포를 살펴보자.

히스토그램 (histogram)

Pima.te$bp에 있는 332개 관측치에 대한 히스토그램

```
par(mfrow=c(2,2))
hist(Pima.te$bp)
hist(Pima.te$bp, breaks=5)
hist(Pima.te$bp, probability=T)
hist(Pima.te$bp, probability=T, breaks=c(20,40,60,80,120))
```

커널밀도함수 (kernel density function)

Pima.te$bp에 있는 332개 관측치에 대한 커널밀도함수. Bandwidth는 밀도함수의 평활정도 (smoothness)를 나타내는 값으로 사용자가 지정하지 않으면 default로 주어진다.

```
par(mfrow=c(2,2))
plot(density(Pima.te$bp))
plot(density(Pima.te$bp, bw=1))
plot(density(Pima.te$bp, bw=5))
plot(density(Pima.te$bp, bw=10))
```

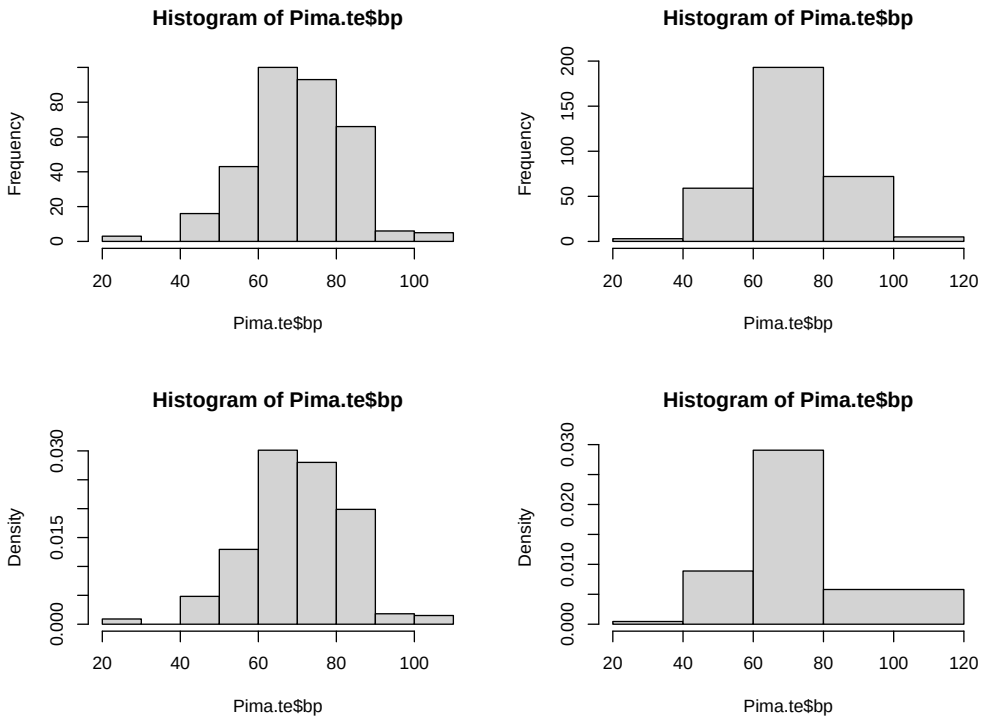

Histogram of Pima.te$bp
Frequency
0
20
40
60
80
20
40
60
80
100
Pima.te$bp
Histogram of Pima.te$bp
Frequency
0
50
100
150
200
20
40
60
80
100
120
Pima.te$bp
Histogram of Pima.te$bp
Density
0.000
0.015
0.030
20
40
60
80
100
Pima.te$bp
Histogram of Pima.te$bp
Density
0.000
0.010
0.020
0.030
20
40
60
80
100
120
Pima.te$bp

그림 2.1: 히스토그램

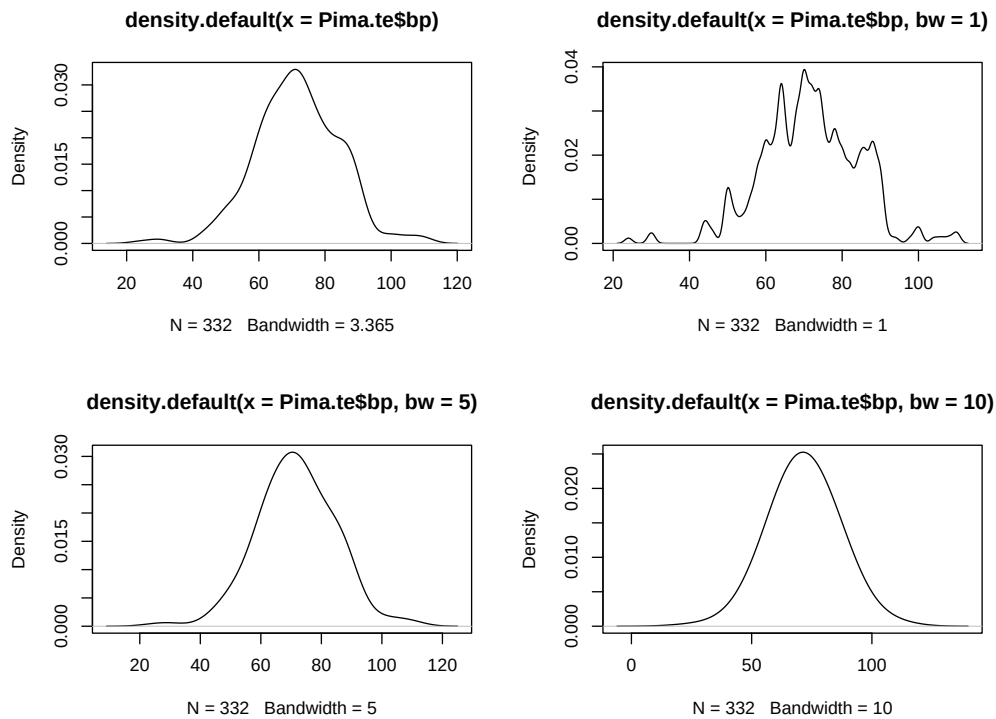

density.default(x = Pima.te$bp)
Density
0.000
0.015
0.030
20
40
60
80
100
120
N = 332 Bandwidth = 3.365
density.default(x = Pima.te$bp, bw = 1)
Density
0.00
0.02
0.04
20
40
60
80
100
N = 332 Bandwidth = 1
density.default(x = Pima.te$bp, bw = 5)
Density
0.000
0.015
0.030
20
40
60
80
100
120
N = 332 Bandwidth = 5
density.default(x = Pima.te$bp, bw = 10)
Density
0.000
0.010
0.020
0
50
100
N = 332 Bandwidth = 10

그림 2.2: 커널밀도함수

상자그림 (box plot)

Pima.te$bp에 있는 332개 관측치에 대한 상자그림. 수평으로 된 5개의 바(bar)는 아래로부터 자료의 min, Q_1, Q_2, Q_3, max의 5가지 값을 나타낸다. 단, 작은 원으로 표시된 것은 이상치(outlier)를 나타낸 것으로 이상치가 있는 경우 자료로 부터 이상치를 제외한 후 min, Q_1, Q_2, Q_3, max의 5가지 값을 나타낸다.

```
par(mfrow=c(1,2))
boxplot(Pima.te$bp)
boxplot(Pima.te$bp, outline=F)
```

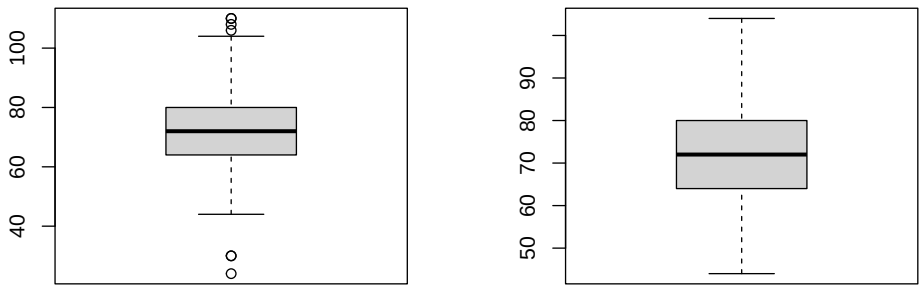


그림 2.3: 상자그림

줄기-잎 그림 (stem-and-leaf plot)

Pima.te$bp에 있는 332개 관측치에 대한 줄기-잎 그림. 첫번째 열은 10단위 수(줄기), 오른쪽 열(잎)은 1단위 수를 나타낸다. 즉, 24, 30, 30, ⋯, 110을 나타낸다. 줄기-잎 그림은 모든 관측치의 값을 정확하게 나타내고, 자료의 분포를 알 수 있는 것이 특징이다.

```
stem(Pima.te$bp)
#>
#>   The decimal point is 1 digit(s) to the right of the |
#>
#>    2 | 4
#>    2 |
#>    3 | 00
#>    3 |
#>    4 | 4444
#>    4 | 66
#>    5 | 000000000022224444
#>    5 | 6666666888888888888
#>    6 | 00000000000000022222222222224444444444444444444444444
#>    6 | 55566666666666888888888888888888
#>    7 | 000000000000000000000000022222222222222222222224444444444444444
#>    7 | 55666666666666688888888888888888
#>    8 | 00000000000000222222222224444444444
#>    8 | 55556666666666668888888888888888
#>    9 | 00000000000024
#>    9 | 8
#>   10 | 0004
#>   10 | 68
#>   11 | 00

stem(Pima.te$bp, scale=0.5)
#>
#>   The decimal point is 1 digit(s) to the right of the |
#>
#>    2 | 4
#>    3 | 00
#>    4 | 444466
#>    5 | 000000000022224444666666888888888888
#>    6 | 00000000000000022222222222224444444444444444444444455566666666666+9
#>    7 | 000000000000000000000000022222222222222222222224444444444444444+26
#>    8 | 000000000000002222222222244444444455556666666666666888888888888888
#>    9 | 000000000000248
#>   10 | 000468
#>   11 | 00
```

2.3 이변량 자료와 상관계수

하나의 변수에 대한 자료를 일변량 자료 (univariate data)라고 하는데 이를 확장하여 여러 개의 변수에 대한 자료를 다변량 자료 (multivariate data)라고 한다. 여기서는 두 개의 변수에 대한 자료, 즉 이변량 자료 (bivariate data)를 요약하는 몇 가지 방법을 소개한다.

2.3.1 질적 자료

두 변수가 모두 질적 자료라고 가정하고 첫 번째 자료는 r개의 범주를 가지고, 두 번째 자료는 c개의 범주를 가진다고 가정하자. 이러한 자료를 행렬의 형태로 요약한 표를 $r \times c$ 분할표 (contingency table)라 부른다. 예를 들어, 폐암자료 (lung)에서 성별 (sex)에 따른 수행점수 (ph.ecog)는 2×6 분할표로서 다음과 같다.

```
sex <- factor(lung$sex, 1:2, c("Male","Female"))
ph.ecog <- factor(lung$ph.ecog, 0:5, 0:5)
tab <- table(sex, ph.ecog); tab
#>         ph.ecog
#> sex       0  1  2  3  4  5
#>   Male   36 71 29  1  0  0
#>   Female 27 42 21  0  0  0

addmargins(tab)
#>         ph.ecog
#> sex        0   1   2   3   4   5 Sum
#>   Male    36  71  29   1   0   0 137
#>   Female  27  42  21   0   0   0  90
#>   Sum     63 113  50   1   0   0 227
round(prop.table(tab, 1)*100, 1) #option: 1=row, 2=column, null=total
#>         ph.ecog
#> sex         0    1    2    3    4    5
#>   Male   26.3 51.8 21.2  0.7  0.0  0.0
#>   Female 30.0 46.7 23.3  0.0  0.0  0.0
```

2.3.2 양적 자료

n개의 이변량 양적자료는 $(x_1, y_1), \cdots, (x_n, y_n)$으로 표현할 수 있다.

(1) 산점도 (scatter plot)

이변량 자료는 이차원 평면에 각 변수의 값에 해당되는 점을 찍은 그림 (산점도)으로 쉽게 나타낼 수 있다. 흔히 산점도는 하나의 변수 값이 증가할 때, 다른 변수의 값이 증가 (또는 감소)하는 추세에 있거나 별 다른 함수관계를 보이지 않는 경우도 있다. 즉, 두 변수의 함수관계를 짐작할 수 있는데 이를 좀 더 명확히 표현할 수 있는 측도를 소개한다.

(2) 표본 상관계수 (sample correlation coefficient)

두 변수의 선형적 함수 관계를 나타내는 측도로서 피어슨 표본 상관계수 (Pearson's sample correlation coefficient)는

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

로 정의되며 여기서

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,\ S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

을 나타낸다. 상관계수는 언제나 -1과 1 사이의 값으로서 -1에 가까우면 음의 상관관계, 1에 가까우면 양의 상관관계, 그리고 0에 가까우면 상관관계가 없음을 나타낸다. 여기서 한 가지 주의할 점이 있다. 위에서 정의된 상관계수는 두 변수X와 Y의 선형적인 관계 (linear relationship)에 대한 것이다. 즉, 상관계수 r이 0에 가까우면 선형적인 상관관계는 없으나 비선형적인 상관관계 (nonlinear relationship)는 존재할 수 있으므로 반드시 산점도를 통해 확인할 필요가 있다.

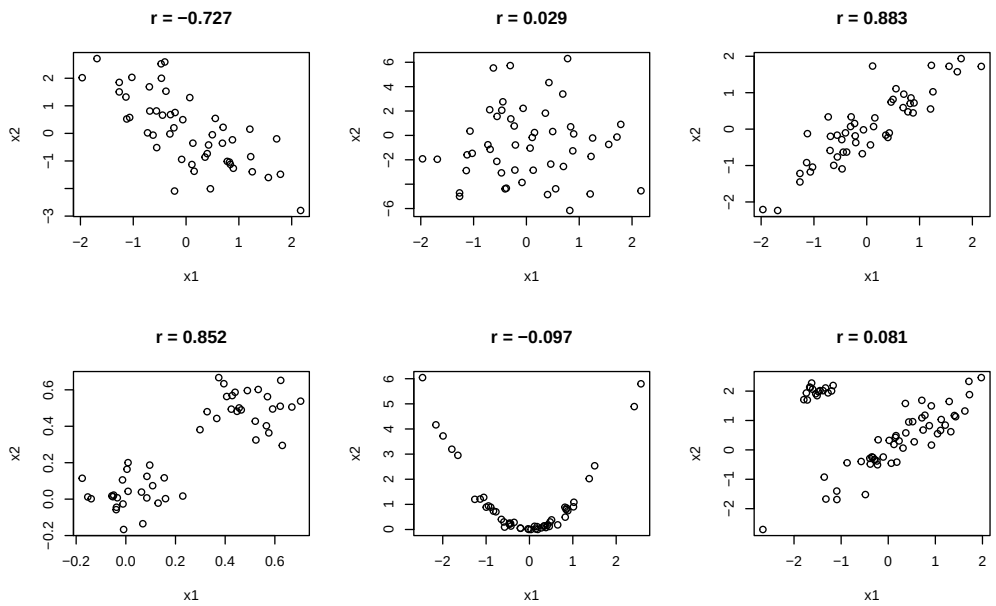


그림 2.4: 산점도와 상관계수

예제 : 폐암자료 (lung)에서 의사가 판단하는 카노프스키 점수 (ph.karno)와 환자가 판단하는 카노프스키 점수 (pat.karno)간의 산점도와 상관계수를 구하자.

```
plot(lung$ph.karno, lung$pat.karno)
```

```
## correlation coefficient
x1 <- lung$ph.karno; x2 <- lung$pat.karno
d <- na.omit(data.frame(x1, x2))

cov(d$x1, d$x2)/sqrt(var(d$x1)*var(d$x2))
#> [1] 0.5202974
cor(d$x1, d$x2)
#> [1] 0.5202974
```

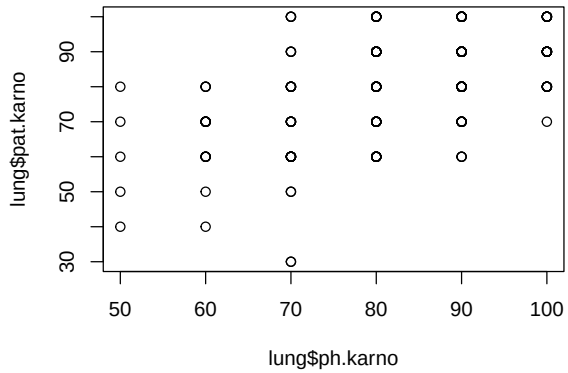


그림 2.5: 산점도

(3) 허위상관과 잠복변수 (spurious correlation and lurking variable)

두 변수의 선형적 상관관계를 파악하기 위해 상관계수를 계산하여 그 값이 1또는 -1에 가까울 경우 두 변수간의 관련성에 대해 결론을 내릴 경우 한 가지 유의해야할 점이 있다. 예를 들어, X는 어느 도시의 연간 강력범죄 건수, Y는 그 도시의 교회 수라고 할 경우 두 변수 간의 상관계수는 1에 가까운 값을 갖게 된다. 이를 바탕으로 "강력범죄를 줄이려면 교회를 없애면 된다"라는 결론을 내리는 것이 타당할까? 실제로 그 도시의 인구 (Z)가 많으면 범죄 건수도 많고, 또한 교회 수도 많게 되므로 범죄 건수와 교회 수가 직접적인 상관이 있는 것이 아니다. 이를 허위상관 (spurious correlation)이라 하며, 이 때 고려해야할 "인구"라는 숨은 변수를 잠복변수 (lurking variable)라 부른다. 즉, 상관계수를 이용한 결과를 해석할 때 이러한 허위상관 여부를 잘 판단해야 하며 허위상관이 의심될 때 그에 해당되는 잠복변수를 찾아내는 것이 매우 중요하다.

제 3 장

통계적 추정

μ와 σ^2은 모집단의 평균과 분산을 나타내는데 이는 모수로서 알 수 없는 값이다. 우리는 앞에서 모집단으로부터 추출한 표본에 대한 평균과 분산을 소개한 적이 있다. 즉, μ와 σ^2을 알 수 없기 때문에 표본을 추출하여 표본평균과 표본분산을 계산하는 것이다.

모집단	표본
모평균 $\mu = E(X)$	표본평균 $\bar{X} = \frac{1}{n}\sum X_i$
모분산 $\sigma^2 = E[(X-\mu)^2]$	표본분산 $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$

모집단으로부터 표본을 추출하여 모집단의 특성을 나타내는 모수에 대한 여러 가지 정보를 얻기 위한 일련의 과정을 통계적 추론 (statistical inference)이라 한다. 통계적 추론은 추정 (estimation)과 가설검정 (test of hypotheses)으로 구성되고, 추정은 점 추정 (point estimation)과 구간 추정 (interval estimation)을 의미한다.

3.1 여러 가지 분포

통계적 추론은 추출된 표본, 즉 자료에 의해 이루어진다. 자료는 어떤 특정 형태의 분포를 따르는 경우가 많다. 자료는 크게 이산형 자료와 연속형 자료로 구분되므로 이산형 자료가

따르는 이산형 분포와 연속형 자료가 따르는 연속형 분포가 있다. 여기서는 의학자료 분석에 자주 언급되는 대표적인 분포 몇 가지를 소개한다.

3.1.1 이산형 분포

(1) 베르누이 시행과 이항분포

다음의 세 가지 조건을 만족하는 시행을 베르누이 시행 (Bernoulli trials)이라 한다.

(i) 시행의 결과는 두 가지 (성공과 실패) 중 하나로 나타난다.
(ii) 각 시행에서 성공의 확률 $p = P(S)$은 동일해야 한다.
(iii) 각 시행은 서로 독립이어야 한다.

이제 각 베르누이 시행에서 성공의 확률이 p인 n번의 시행에서 성공의 횟수를 X라 하면 X는 이항분포 (binomial distribution)를 따른다고 하고 확률밀도함수는

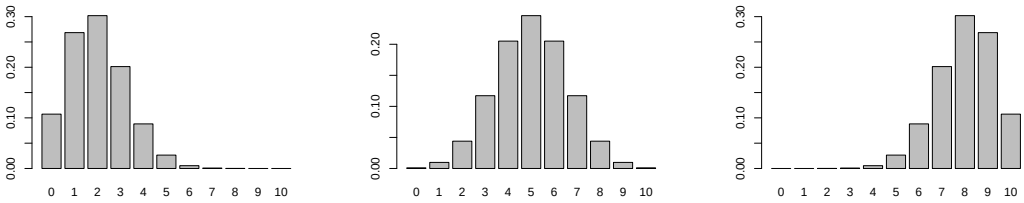
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \;\; x = 0, 1, \cdots, n$$

으로 주어지며 $X \sim B(n, p)$로 표현한다. 이항분포의 평균과 분산은 $ ☒= np ,~~ ☒^2 = np(1-p)$이 됨을 쉽게 보일 수 있다. 참고로 확률변수 Y가 베르누이 시행을 따르면 $Y \sim B(1, p)$로 표현할 수 있다. 또한, $Y_1, \cdots, Y_n$이 서로 독립인 베르누이 시행이라 하고 $X = \sum_{i=1}^{n} Y_i$라 하면 $X \sim B(n, p)$이 된다.

```
n = 10; p = 0.2
x <- 0:n
y <- dbinom(x, n, p) #B(n, p)
barplot(y, names.arg=x)
```

(2) 포아송 분포

확률변수 X를 단위시간 또는 단위공간에서 발생되는 사건의 수라 정의하고, m을 단위시간 또는 단위공간에서 발생되는 사건의 수의 평균, 즉 $m = E(X)$라 하면, 확률변수 X는 평균



(a) B(10, 0.2) (b) B(10, 0.5) (c) B(10, 0.8)

그림 3.1: 이항분포

m인 포아송 분포 (Poisson distribution)를 따른다고 하며, 확률밀도함수는

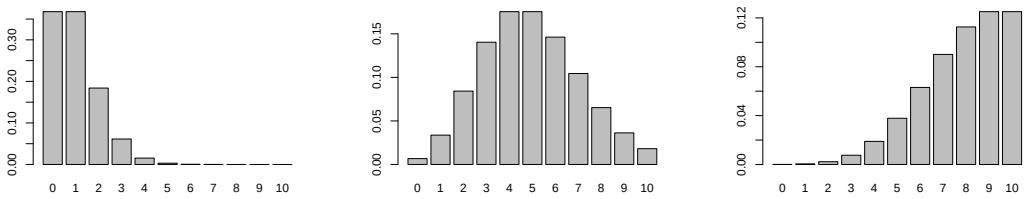
$$f(x) = \frac{e^{-m}m^x}{x!}, \quad x = 0, 1, 2, \cdots$$

으로 주어지고, $X \sim P(m)$으로 표현한다. 포아송 분포의 평균과 분산은 $\mu = \sigma^2 = m$ 이 됨을 보일 수 있다.

어떤 사건이 포아송 분포를 따르기 위한 조건으로 다음의 세 가지 성질을 만족해야 한다.

(i) 독립성 (independence): 단위시간 또는 단위공간에서 발생되는 사건의 수는 또 다른 단위시간 또는 단위공간에서 발생되는 사건의 수와 무관하다.
(ii) 단일성 (lack of clustering): 동시에 두 개 이상의 사건이 발생할 확률은 0에 가깝다.
(iii) 등발성 (constant rate): 평균 m은 모든 단위시간 또는 단위공간에서 일정하다.

```
x <- 0:10; y <- dpois(x, 1) #Poi(1)
barplot(y, names.arg=x)
```



(a) P(1) (b) P(5) (c) P(10)

그림 3.2: 포아송분포

3.1.2 연속형 분포

(1) 정규분포

정규분포 (Normal Distribution)는 독일의 수학자이자 천문학자인 가우스 (Johann Carl Friedrich Gauss, 1777-1855)에 의해 제안되어 일명 가우시안 분포 (Gaussian distribution)이라고도 하며 확률밀도함수는

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

으로 주어지고 $X \sim N(\mu, \sigma^2)$으로 나타낸다.

(2) 표준정규분포

확률변수 X가 평균 μ, 분산 σ^2을 가질 때

$$Z = \frac{X - E(X)}{\sqrt{Var(X)}} = \frac{X-\mu}{\sigma}$$

를 표준화된 변수 (standardized random variable)라 부른다. 표준화된 변수는 항상 평균 0, 분산 1을 가지게 된다. 예를 들어, $X \sim B(n,p)$일 때 $Z = \frac{X-np}{\sqrt{np(1-p)}}$의 평균은 0, 분산은 1이 된다. 정규분포에 적용하면

$$X \sim N(\mu, \sigma^2) \;\Rightarrow\; Z = \frac{X-\mu}{\sigma}$$

의 평균은 0, 분산은 1이 된다. 여기서, 매우 중요한 사실은 정규분포의 경우 표준화를 시키면 역시 정규분포가 된다는 사실이다. 즉, $Z \sim N(0,1)$이 된다. 이러한 사실은 다른 분포에서는 적용되지 않는다. 즉, 이항분포를 표준화시키면 더 이상 이항분포를 따르지 않는다는 사실이다. 표준화된 변수의 또 다른 특징은 분포에 관계없이 -3 과 +3 사이의 값을 가지는 경우가 대부분이다. 표준정규분포의 오른쪽 꼬리의 면적이 α되는 지점은 흔히 z_α로 나타낸다. 즉,

$$\alpha = P(Z > z_\alpha)$$

의 관계를 가진다.

```
x <- seq(-3,3,length=100)
y <- dnorm(x, 0, 0.5)
plot(x, y, type="l")
```

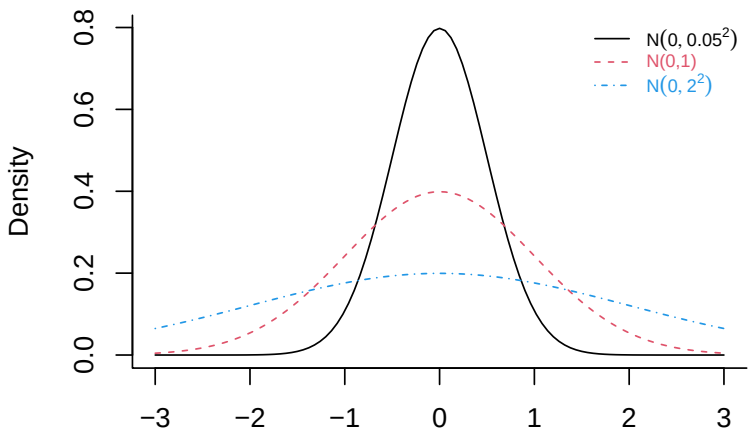
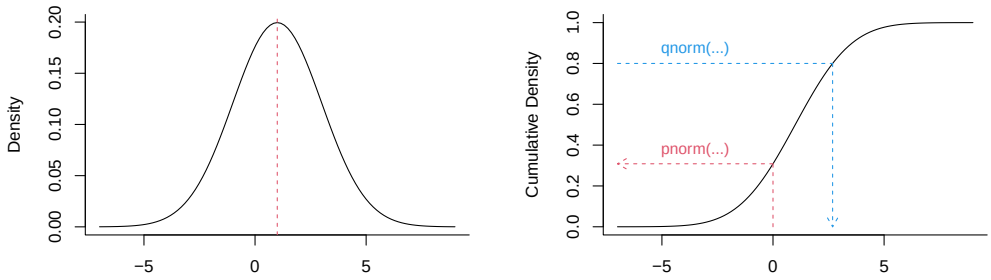


그림 3.3: 정규분포

R을 이용하여 정규분포 $N(\mu, \sigma^2)$와 관련된 여러 가지 유용한 계산을 할 수 있다. 이를 위해 밀도함수는 $f(x)$, 누적분포함수(cumulative distribution function)는 $F(x) = \int_{-\infty}^{x} f(t)dt$, 분위수함수(quantile function)는 $Q(p) = F^{-1}(p),\ 0 < p < 1$로 표현하자. 정규분포

(normal distribution)를 의미하는 것은 `norm`으로 나타내고 그 앞에 어떤 글자를 덧붙이는가에 따라 용도가 달라진다. 밀도함수의 값은 `d`, 누적분포함수의 값은 `p`, 분위수함수의 값은 `q`를 붙인다. 한편, 정규분포로 부터 난수(random numbers)를 발생하고 싶을 때는 `r`을 붙이면 된다. 예를 들어, `dnorm(1.6, 1, 2)`는 $N(1,2^2)$에서 $f(1.6)$을 계산하는 것이고, `qnorm(0.1, 1,2, lower.tail=T)`은 $N(1,2^2)$에서 $Q(0.1)$를 계산하는 것이다. 여기서, `lower.tail`은 누적분포의 왼쪽 꼬리부분 면적이 0.1이 된다는 것을 의미한다. 만약 오른쪽 꼬리부분의 면적이 0.1되는 분위수함수 값은 `qnorm(0.1, 1,2, lower.tail=F)` 또는 `qnorm(0.9, 1,2, lower.tail=T)`로 표현하면 된다.

```
par(bty="l")
x <- seq(-5,7,length=100)
y <- dnorm(x, 1, 2); plot(x, y, type="l")
y <- pnorm(x, 1, 2); plot(x, y, type="l")
```



(a) 확률밀도함수 $N(1,2^2)$ (b) 누적분포함수 $pnorm, qnorm$

그림 3.4: 확률밀도함수와 누적분포함수

```
dnorm(1.6, 1, 2); pnorm(1.6, 1, 2); qnorm(0.1, 1, 2); qnorm(0.9, 1, 2)
#> [1] 0.1906939
#> [1] 0.6179114
#> [1] -1.563103
#> [1] 3.563103
```

3.2 표본의 분포

3.2.1 모수와 통계량

앞에서 우리는 모집단의 특성을 나타내는 값을 모수 (parameter)라 하였으며, 이 모수는 흔히 알 수 없는 상수라고 가정한다. 이렇게 알 수 없는 상수인 모수의 값을 추정하기 위해 모집단으로부터 크기 n인 표본을 추출하는데 흔히 $X_1, X_2, \cdots, X_n$으로 나타낸다. 예를 들어, 표본평균 $\bar{X} = \frac{1}{n}\sum X_i$ 나 표본분산 $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$은 표본들만의 함수로 표현되는데 이를 통계량 (statistic)이라 부른다. 즉, 통계량은 특정 모수를 추정하기 위해 표본을 이용하여 만든 함수이다.

여기서, 매우 중요한 개념은 통계량 자체가 변동성을 띈다는 것이다. 즉, 통계량은 상수가 아니라 함수라는 것이다. 왜냐하면 모집단으로부터 추출한 표본 자체가 함수이기 때문이다. 크기 n인 표본은 누가 추출하느냐에 따라 달라지고 이를 바탕으로 구성된 통계량 또한 표본이 어떻게 추출되었느냐에 따라 달라지기 때문이다. 따라서, 통계량은 분포를 가지게 되며 이를 표본분포 (sampling distribution)이라 부른다.

3.2.2 표본평균의 분포

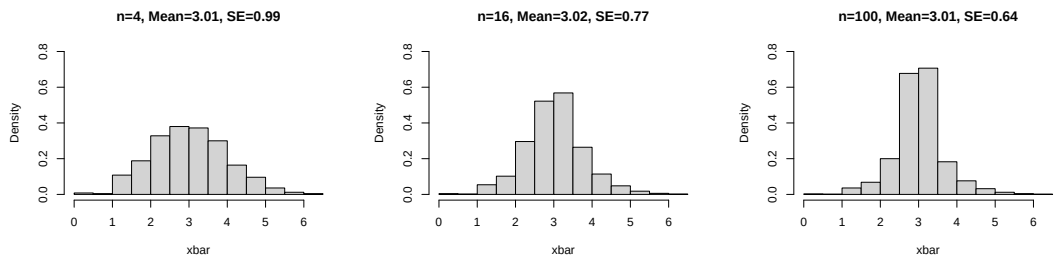
크기 n인 표본 $X_1, X_2, \cdots, X_n$ 이 동일한 분포 (동일한 모집단)를 가지며, 서로 독립일 때 $X_1, X_2, \cdots, X_n$을 임의표본 (random sample)이라 부른다. 모평균이 μ이고, 모분산이 σ^2인 모집단으로부터 크기 n인 임의표본 $X_1, X_2, \cdots, X_n$ 을 추출하여 구한 표본평균 $\bar{X} = \frac{1}{n}\sum X_i$ 의 기댓값과 분산은 각각

$$E(\bar{X}) = \mu, \;\; Var(\bar{X}) = \sigma^2/n$$

이 된다. 즉, $\bar{X}$의 표준편차는 $s.d.(\bar{X}) = \sigma/\sqrt{n}$이 된다.

예제 : 모평균 $\mu = 3$, 모표준편차 $\sigma = 2$인 정규모집단으로 부터 $n = 4$개씩 무작위 추출하고 표본평균을 구하자. 이를 $r = 500$번 반복할 때 표본평균 $(\bar{x})$의 분포를 구해 보자.

```
mu=3; sigma=2;
n=4; r=500; set.seed(11); xbar = c()
for(i in 1:r){
  x <- rnorm(n, mu, sigma) #N(mu,sigma^2)
  xbar <- c(xbar, mean(x))
}
hist(xbar, prob=T)
mu; mean(xbar)              #Mean
#> [1] 3
#> [1] 3.006184
sigma/sqrt(n); sd(xbar)     #SE
#> [1] 1
#> [1] 0.9853343
```



(a) n=4 (b) n=16 (c) n=100

그림 3.5: 표본평균의 분포

3.2.3 중심극한정리

중심극한정리 (central limit theorem) : 모평균이 μ이고, 모분산이 σ^2인 모집단으로부터 임의표본 $X_1, X_2, \cdots, X_n$을 추출하여 구한 표본평균 $\bar{X} = \frac{1}{n}\sum X_i$는 표본의 크기 n이 충분히 크면 근사적으로 $N(\mu, \frac{\sigma^2}{n})$분포를 따르게 된다. 즉, 이러한 사실을 표준화하면

$$\frac{\bar{X} - E(\bar{X})}{s.d.(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - E(X_1))}{s.d.(X_1)}$$

는 근사적으로 $N(0,1)$을 따르게 된다.

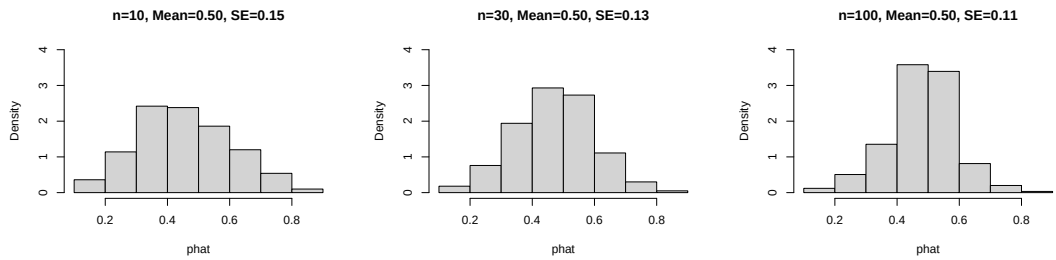
참고 : 중심극한정리에서 표본의 크기 n은 크면 클수록 더 정규분포에 가까이 수렴한다. 흔히, n이 25 정도 이상이면 중심극한정리를 이용해도 무방한 것으로 알려져 있다. 우리는 앞에서 $Y_1, \cdots, Y_n$이 서로 독립인 베르누이 시행이라 하고 $X = \sum_{i=1}^{n} Y_i$라 하면 $X \sim B(n,p)$이 됨을 보인 바 있다. 즉, $Y_1, \cdots, Y_n$의 표본평균 $\bar{Y}$는 X/n과 같다. 이제 $\bar{Y}$에 중심극한정리를 적용하면

$$\frac{\sqrt{n}(\bar{Y} - E(Y_1))}{s.d.(Y_1)} = \frac{\sqrt{n}(\frac{X}{n} - p)}{\sqrt{p(1-p)}} = \frac{(X - np)}{\sqrt{np(1-p)}}$$

가 근사적으로 $N(0,1)$을 따르게 된다. 이를 이항분포의 정규근사화 (normal approximation to the binomial)라 부른다. 단, 근사화가 제대로 되려면 $np > 15,\ \ n(1-p) > 15$ 두 가지 조건이 만족되어야 한다는 사실이 알려져 있다.

예제 : 성공확률 $p = 0.5$인 모집단으로 부터 베르누이 시행을 $n = 30$번 실시했을 때 성공횟수 x와 성공률 $\hat{p} = x/n$을 구하자. 이를 $r = 500$번 반복할 때 표본비율 $(\hat{p})$의 분포를 구해보자.

```
p=0.5; n=30; r=500; set.seed(11); phat = c()
for(i in 1:r){
  x <- rbinom(n, 1, p)       #B(1, p)
  phat <- c(phat, sum(x)/n)
}
hist(phat, prob=T, breaks=6)
p; mean(phat)                #Mean
#> [1] 0.5
#> [1] 0.5026
sqrt(p*(1-p)/n); sd(phat)    #SE
#> [1] 0.09128709
#> [1] 0.09217863
```



(a) n=10 (b) n=30 (c) n=100

그림 3.6: 표본평균의 분포 (이항분포의 정규근사)

3.2.4 t-분포

$X_1, X_2, \cdots, X_n$을 $N(\mu, \sigma^2)$으로 부터의 임의 표본일 경우 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$는 $N(0,1)$을 따른다는 사실을 알고 있다. 만약 σ를 모를 경우에 추정치 s를 대입한

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

의 분포는 무엇일까? 영국의 통계학자 William Sealy Gosset (1876 - 1937)은 Arthur Guinness & Son (Dublin, Ireland)라는 양조회사 근무 중 1908년 Student라는 필명으로 t 분포를 제안하였다. 좀 더 정확히 서술하면 다음과 같다.

Student's t-분포의 정의 : 임의표본 $X_1, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$를 따를 경우

$$T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

는 자유도 (degree of freedom) $n-1$인 t-분포를 따르게 되며 $T \sim t(n-1)$으로 표현한다.

여기서 자유도(degree of freedom (d.f.))는 흔히 관측치 수에서 제약 조건의 개수를 뺀 것으로 해석할 수 있다. 예를 들어, 3개의 관측치 X_1, X_2, X_3 가 있을 때 아무런 제약이 없다면 자유도는 $3-0=3$이 된다. 하지만, $\bar{X}=3$이라는 제약조건이 하나 주어지면 자유도는 $3-1=2$가 된다. t-분포에서 자유도가 n이 아닌 $n-1$이 되는 이유는 σ 대신 추정치인 s를 대입함으로써 1개의 자유도를 잃어버린 것으로 해석할 수 있다.

```
par(bty="l")

x <- seq(-3,3,length=100)
y1 <- dnorm(x, 0, 1) #N(0,1)
y2 <- dt(x, 1)       #t(1)

plot(x, y1, type="l")
lines(x, y2)
```

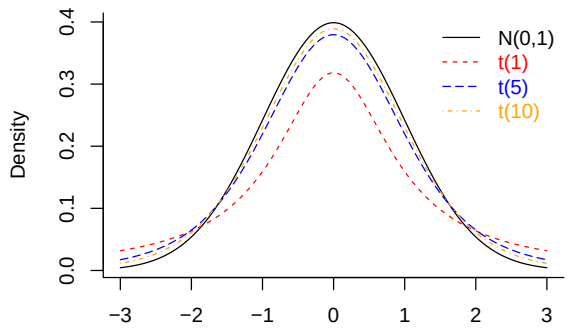


그림 3.7: t-분포

3.3 점추정

주어진 자료가 크기 n인 임의표본 $X_1, \cdots, X_n$일 때 추정하고자 하는 모수를 한 개의 값으로 나타내는 것을 점 추정 (point estimation)이라 하고, 이러한 값을 주는 함수를 점 추정량 (point estimator), 그 함수의 구체적 값을 점 추정치 (point estimate)라 부른다.

정의 : 어떤 모수 θ에 대한 점 추정량 $\hat{\theta} \equiv \hat{\theta}(X_1, \cdots, X_n)$이 $E(\hat{\theta}) = \theta$를 만족하면 $\hat{\theta}$은 θ에 대한 불편 추정량 (unbiased estimator)이라 부른다.

3.3.1 모평균 μ에 대한 점 추정치 $\bar{X}$의 성질

모평균 μ에 대한 점 추정량으로서 표본평균 $\bar{X}$를 고려하면

$$E(\bar{X}) = \mu\,, \;\; Var(\bar{X}) = \sigma^2/n\,, \;\; s.d.(\bar{X}) = \sigma/\sqrt{n}$$

이므로 $\bar{X}$는 μ에 대한 불편추정량이다. 여기서, $s.d.(\bar{X}) = \sigma/\sqrt{n}$이지만 실제로 σ를 모르기 때문에 σ대신 그것의 점 추정량인 $s = \sqrt{s^2}$ (즉, 표본 표준편차)를 대입한 값 $s/\sqrt{n}$을 표준오차 (standard error)라 부르며 $s.e.(\bar{X}) = s/\sqrt{n}$으로 나타낸다. 즉, $\bar{X}$의 표준오차는 $\bar{X}$의 표준편차의 추정치가 된다.

3.3.2 모분산 σ^2에 대한 점 추정치 s^2의 성질

모분산 σ^2에 대한 점 추정량으로 표본분산인 s^2을 고려하면 $E(s^2) = \sigma^2$이 되어 s^2은 σ^2에 대한 불편추정량이 된다. 한편, 표본분산의 분산 즉, $Var(s^2)$은 모집단이 정규분포를 따를 경우 계산 할 수 있으나 그 외의 경우는 계산이 매우 어려워서 여기서는 다루지 않기로 한다.

3.3.3 모비율 p에 대한 점 추정치 $\hat{p}$의 성질

임의표본 $X_1, \cdots, X_n$이 성공의 확률 p인 베르누이 시행을 따를 경우 p에 대한 점 추정량 역시 표본평균을 사용하며 이 때 $X \equiv \sum_{i=1}^{n} X_i \sim B(n, p)$이므로 흔히 점 추정치는 $\hat{p} = X/n$으로 표현한다.

3.4 구간추정

모수 θ에 대한 구간 추정치는 구간 $(L(\hat{\theta}), U(\hat{\theta}))$로 주어지는데, 모수 θ에 대한 $100(1-\alpha)\%$ (단, $0 < \alpha < 1$) 신뢰구간은 $P(L(\hat{\theta}) < \theta < U(\hat{\theta})) = 1 - \alpha$ 을 만족하는 구간 $(L(\hat{\theta}), U(\hat{\theta}))$이다. 여기서, $L(\hat{\theta})$과 $U(\hat{\theta})$은 점 추정치 $\hat{\theta} \equiv \hat{\theta}(X_1, \cdots, X_n)$의 함수이다. α는 신뢰수준 (level of confidence)이라 하며 흔히 0.01 (99% 신뢰구간), 0.05, (95% 신뢰구간) 또는 0.1 (90% 신뢰구간)의 값을 주로 가진다.

3.4.1 모평균 μ에 대한 신뢰구간 (대표본의 경우)

대표본 (표본의 크기 n이 25보다 큰 경우)을 가정한다. 모집단에 대한 분포의 가정이 없을 경우 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간을 정확하게 계산하는 것은 불가능하다. 그러나, 대표본인 경우 모평균 μ에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은 중심극한정리에 의하여

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

이 근사적으로 $N(0, 1)$을 따르므로 다음의 부등식을 만족하는 구간을 계산하면 된다.

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

이 부등식을 μ에 대해 풀면

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

이 된다. 즉, 모평균 μ에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

의 형태를 가진다. 만약 σ를 모를 경우에도 대표본이면 점 추정치 s를 대입한 것을 사용하여도 여전히

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

이 근사적으로 $N(0,1)$을 따르므로 모평균 μ에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

이 된다. 위의 식은

$$\bar{X} \pm z_{\alpha/2}\ s.e.(\bar{X})$$

으로 표현할 수 있다.

일반적으로 모수 θ에 대한 점 추정치를 $\hat{\theta}$라 하면 모수 θ에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$$\hat{\theta} \pm z_{\alpha/2}\ s.e.(\hat{\theta})$$

으로 표현된다.

예제 : 모평균 $\mu = 0$, 모표준편차 $\sigma = 1$인 정규모집단으로 부터 $n = 30$개를 무작위 추출하고 표본평균과 95% 신뢰구간을 계산하자. 신뢰구간안에 모평균 $\mu = 0$이 포함되는지 확인하자. 이를 $r = 100$번 반복할 때 모평균 $\mu = 0$이 신뢰구간안에 포함되는 비율을 구해보자.

```
mu=0; sigma=1
n=30; rep=100; set.seed(123); out=c()
plot(Inf, Inf, xlim=c(1,rep), ylim=c(-1,1)); abline(h=mu)
for(index in 1:rep){
  x <- rnorm(n, mu, sigma)          #N(mu, sigma^2)
  z <- qnorm(0.025, lower.tail=F)   #Z_0.25 = 1.96
  m <- mean(x); se <- sd(x)/sqrt(n)
  lower <- m - z*se; upper <- m + z*se
  rst <- (mu >= lower & mu <=upper) #true or false
  out <- c(out, rst)
  lines(c(index,index), c(lower,upper), col=ifelse(rst, 1, 2))
}
out[1:5]
#> [1] TRUE TRUE TRUE TRUE TRUE
sum(out)/rep
#> [1] 0.95
```

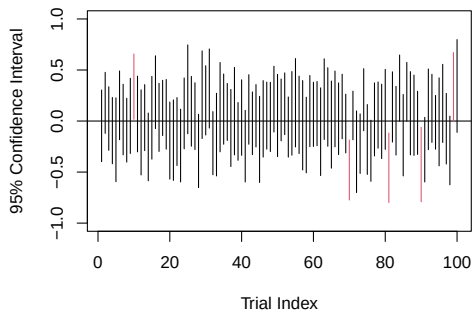


그림 3.8: 모평균에 대한 95% 신뢰구간 (대표본)

3.4.2 모비율 p에 대한 신뢰구간 (대표본의 경우)

표본 $X_1, \cdots, X_n$ 은 성공의 확률 p인 베르누이 시행이고 $np > 15,\ n(1-p) > 15$ 를 만족한다고 가정한다. 이 경우

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{p(1-p)/n}}$$

이 근사적으로 $N(0,1)$을 따르므로 $\hat{p} = X/n$이라 두면

$$-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}$$

이 되고 이를 p에 대해 풀면

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

이 된다. 그런데 위의 구간은 알 수 없는 p를 포함하고 있으므로 이를 점 추정치인 $\hat{p}$로 대체시켜도 여전히 근사적 성질을 만족하게 되며 모비율 p에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

이 된다. 이 또한,

$$\hat{p} \pm z_{\alpha/2}\ s.e.(\hat{p})$$

으로 표현할 수 있다.

예제 : 성공확률 $p = 0.5$인 모집단으로 부터 베르누이 시행을 $n = 50$번 실시하고 표본비율과 95% 신뢰구간을 계산하자. 신뢰구간안에 모비율 $(p = 0.5)$이 포함되는 지를 확인하자. 이를 $r = 100$번 반복할 때 모비율 $(p = 0.5)$이 신뢰구간안에 포함되는 비율을 구해보자.

```
p=0.5; n=50; rep=100; set.seed(123); out=c()
plot(Inf, Inf, xlim=c(1,rep), ylim=c(0,1)); abline(h=p)
for(index in 1:rep){
  x <- rbinom(n, 1, p)              #B(1, 0.5)
  z <- qnorm(0.025, lower.tail=F)   #Z_0.25 = 1.96
  phat <- sum(x)/n; se <- sqrt((phat*(1-phat))/n)
  lower <- phat - z*se; upper <- phat + z*se
  rst <- (p >= lower & p <=upper)
  out <- c(out, rst)
  lines(c(index,index), c(lower,upper), col=ifelse(rst, 1, 2))
}
out[1:5]
#> [1] TRUE TRUE TRUE TRUE TRUE
sum(out)/rep
#> [1] 0.95
```

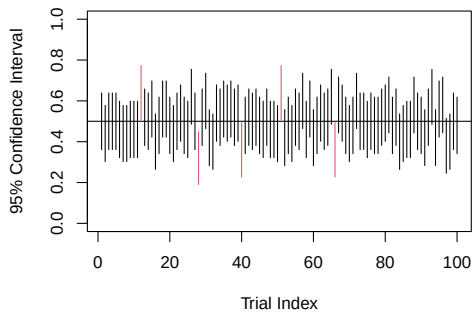


그림 3.9: 모비율에 대한 95% 신뢰구간 (대표본)

3.4.3 모평균 μ에 대한 신뢰구간 (소표본 정규모집단의 경우)

앞서 우리는 모집단의 분포에 관계없이 대표본일 경우에 모평균 및 모비율에 대한 구간추정을 논의하였다. 만약 표본의 크기가 25이하로 작을 경우 (소표본) 구간추정은 어떻게 실시되는가? 소표본인 경우에는 이 필요하다. 만약, σ가 알려져 있다면

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

이므로 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

이 된다. σ를 모르는 경우에는 표본표준편차인 s로 대체시키면

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$$

이므로 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

의 형태를 가진다.

```
n <- 10; set.seed(123)
x <- rnorm(n)                       #N(0, 1)

## Method I
t <- qt(0.025, n-1, lower.tail=F) #t_0.025(df)
m <- mean(x); se <- sd(x)/sqrt(n)
lower <- m - t*se; upper <- m + t*se
c(mean=m, t=z, se=se, low=lower, high=upper)
#>        mean           t          se         low        high
#>  0.07462564  1.95996398  0.30161300 -0.60767036  0.75692165
```

```
## Method II
t.test(x)
#>
#>  One Sample t-test
#>
#> data:  x
#> t = 0.24742, df = 9, p-value = 0.8101
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  -0.6076704  0.7569217
#> sample estimates:
#>  mean of x
#> 0.07462564
```

3.4.4 모분산 σ^2에 대한 신뢰구간 (정규모집단의 경우)

임의표본 $X_1, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$를 따른다고 가정하자. 모분산 σ^2에 대한 점추정치는 표본분산 $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ 을 주로 사용하고 이는 불편추정치가 된다. 모분산 σ^2에 대한 구간추정을 위해 다음과 같은 정의가 필요하다.

정의 :

$$W \equiv \frac{\sum(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

는 자유도 $n-1$인 카이제곱 (chi-square)분포를 따르게 되고 $W \sim \chi^2(n-1)$으로 나타낸다.

따라서, 모분산 σ^2에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$1 - \alpha = P[\, \chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}} \,]$$

이므로 이를 σ^2에 대해서 풀면

$$(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}})$$

이 되고, 같은 방법으로 모표준편차 σ에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$(s\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}}}, s\sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}}})$$

으로 주어진다.

```
n <- 10; set.seed(123)
x <- rnorm(n) #N(0, 1)
sd(x)
#> [1] 0.9537841
sd(x)*sqrt((n-1)/qchisq(0.975, n-1)) #lower
#> [1] 0.6560462
sd(x)*sqrt((n-1)/qchisq(0.025, n-1)) #upper
#> [1] 1.741238
```

제 4 장

통계적 가설검정

4.1 정의와 용어

4.1.1 가설검정의 절차

가설 (hypothesis)이란 아직 증명되지 않은 문제에 대하여 문장으로 설정한 것이라 정의할 수 있다. 그 중에서도 통계적 가설이란 가설의 참과 거짓을 귀납법, 연역법 등으로 증명할 수 없고 관련된 통계 자료를 이용하여 가장 가능성이 높은 결론을 내리고자 하는 것이다. 가설검정의 개념을 설명하기 위해 다음과 같은 가상적 자료를 소개한다.

문제 : 현재 시중에서 판매되는 진통제의 지속효과 시간은 평균 10.0시간으로 알려져 있다. 어느 제약회사에서는 새로 개발한 진통제가 지속효과 시간을 늘여준다고 주장한다. 새로 개발된 약을 여러 사람에게 투여한 후 지속시간에 대한 자료를 바탕으로 이 가설의 참 또는 거짓을 논할 수밖에 없으므로 이는 통계적 가설의 대표적인 예다. 이 회사에서는 통증환자 50명을 임의로 선정하여 진통제의 지속시간을 구해본 결과 표본평균 $\bar{X}$ = 10.5시간, 표본분산 4.0 시간으로 나타났다. 과연 이 회사에서 개발한 진통제는 지속시간을 늘여준다는 주장이 옳은가?

위의 문제를 이용한 통계적 가설 검정의 절차는 다음과 같다.

(1) 귀무가설과 대립가설을 세운다.

- H_0 : 부정하고 싶은 문장 (귀무가설 : null hypothesis)
- H_1 : 주장하고 싶은 문장 (대립가설 : alternative hypothesis)

$\Rightarrow\ H_0 : \mu = 10\ \ vs\ \ H_1 : \mu > 10$

(2) 검정통계량을 선택한다.

- 검정통계량 (test statistic)이란 귀무가설의 기각 여부를 결정하는데 사용되는 통계량으로서 이 문제에서 연비의 참 값 μ에 대한 점 추정치로 표본평균 $\bar{X}$를 사용하는 것이 타당하다.

(3) 기각역을 설정한다.

- 기각역 (rejection region)이란 귀무가설이 기각되는 영역으로 이 문제의 경우 어떤 값 c에 대하여 $\bar{X} > c$ 이면 H_0를 기각한다. 여기서 c는 기각치 (critical value)라 부른다.

통계적 가설이 다른 가설과 근본적으로 차이가 나는 것은 귀무가설과 대립가설 어느 것을 채택하더라도 오류 (error)가 발생된다는 것이다. 왜냐하면, 어느 가설이 참인지는 결코 알 수 없기 때문이다. 이를 표로 정리하면 다음과 같다.

분석가의 결정	H_0 참	H_1 참
H_0을 기각함	제1종 오류	올바른 결정
H_0을 기각하지 않음	올바른 결정	제2종 오류

제1종 오류를 일으킬 확률을 유의수준 (level of significance)이라 부르며 주로 α로 표현하고, 제 2종 오류를 일으킬 확률을 β로 나타낸다.

- α : 귀무가설이 맞음에도 불구하고 귀무가설을 기각하는 오류를 일으킬 확률
- β : 귀무가설이 틀림에도 불구하고 귀무가설을 기각하지 않는 오류를 일으킬 확률

- $1-\beta$: 검정력 (power of test)

최적의 가설검정은 α와 β 모두 최소화시키는 검정법이지만 만약 자료의 크기 n이 정해져 있다면 α와 β 모두 최소화하는 것은 불가능하다. 따라서, 가능한 최적의 검정법은 α를 작은 값 (흔히 0.01, 0.05, 또는 0.1)으로 고정시킨 후 검정력 $(1-\beta)$을 최대화 시키는 검정으로서 이를 일양최강검정 (uniformly most powerful test)이라 한다.

유의수준 α가 주어졌을 때 기각역 계산하는 방법은 귀무가설하에서 검정통계량의 분포를 구한 다음 유의수준에 맞는 기각역을 결정한다. 위의 문제에서 유의수준 α가 주어졌을 때 기각역을 계산해 보자.

$$\alpha = P[\,\bar{X} > c\,] = P[\,\frac{\bar{X}-10}{2/\sqrt{50}} > \frac{c-10}{2/\sqrt{50}}\,] = P[Z > z_\alpha]$$

이므로 $c = z_\alpha \frac{2}{\sqrt{50}} + 10$ 이 된다. 만약 유의수준 $\alpha = 0.05$가 주어졌다면 $c = \frac{1.6452}{\sqrt{50}} + 10 = 10.465$ 가 된다. 따라서, 관측된 검정통계량은 $\bar{X} = 10.5$이므로 유의수준 0.05에서 귀무가설을 기각한다. 즉, 새로 개발한 진통제는 지속시간이 기각치인 10.465시간보다 크기 때문에 기각역에 속한다. 즉, 귀무가설이 기각되므로 지속시간이 늘어났다고 할 수 있다.

4.1.2 유의확률

유의확률에 대한 정의는 주어진 검정통계량을 기각시키기 위한 제1종 오류의 최소값이다.

(예) 위에서 검정통계량은 $\bar{X} = 10.5$이므로 이 값이 주어졌을 때 귀무가설을 기각시키기 위한 기각역은 $\bar{X} > 10.5$, $\bar{X} > 10.4$, $\bar{X} > 10.3$ 등 여러 가지가 가능하다. 이 들에 대한 제1종 오류는 각각

$$P[\bar{X} > 10.5] = P[\frac{\bar{X}-10}{2/\sqrt{50}} > \frac{10.5-10}{2/\sqrt{50}}] = P[Z > 1.768] = 0.038$$

$$P[\bar{X} > 10.4] = P[\frac{\bar{X}-10}{2/\sqrt{50}} > \frac{10.4-10}{2/\sqrt{50}}] = P[Z > 1.414] = 0.080$$

$$P[\bar{X} > 10.3] = P[\frac{\bar{X}-10}{2/\sqrt{50}} > \frac{10.3-10}{2/\sqrt{50}}] = P[Z > 1.061] = 0.145$$

이 된다. 당연히, 기각역은 $\bar{X} > 10.5$에 대한 제1종 오류값인 0.038이 가장 작으며 이 값이 유의확률 (p-value)이 된다.

〈유의확률 (p-value)에 대한 몇 가지 사실〉

(i) 유의확률이 작을수록 귀무가설을 기각할 수 있는 정당성이 커진다.
(ii) 주어진 유의수준 α에서 귀무가설이 기각 $\Rightarrow$ 유의확률 〈 유의수준
(iii) 주어진 유의수준 α에서 귀무가설이 채택 $\Rightarrow$ 유의확률 〉 유의수준

- p-value 〈 0.05 : 유의함 (significant) *
- p-value 〈 0.01 : 매우 유의함 (highly significant) **
- p-value 〈 0.001 : 매우 강력하게 유의함 (highly strongly significant) ***

4.2 여러 가지 검정법 (모수적 방법)

4.2.1 일표본 검정

일표본 검정 (one-sample test)이란 한 개의 집단에서 한 개의 모수에 대한 검정을 의미한다. 예를 들어, 어느 모집단에서 모평균에 대한 검정, 모비율에 대한 검정, 또는 모분산에 대한 검정 등이 그것이다. 주어진 귀무가설에 대해 대립가설의 형태에 따라 기각역의 형태가 결정되며, 그 형태에 따라 단측검정 (one-sided test)과 양측검정 (two-sided test)으로 나눈다.

(1) 모평균 μ에 대한 검정 (대표본) - 일표본 Z-검정

임의표본 $X_1, \cdots, X_n$이 평균 μ, 분산 σ^2인 모집단으로부터 추출되었고 표본의 크기 n이 25보다 큰 대표본을 가정하자. 모평균에 대한 검정으로 귀무가설 $H_0 : \mu = \mu_0$ 에 대하여 3가지 형태의 대립가설이 가능하며 그에 따른 기각역의 형태는 다음과 같다. 단, 여기서 검정통계량은 $Z = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ 이고 이를 일표본 Z-검정 (one-sample Z-test)이라 한다.

귀무가설	대립가설	기각역	비고
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$Z > z_\alpha$	단측검정
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$Z < -z_\alpha$	단측검정

귀무가설	대립가설	기각역	비고
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$\|Z\| > z_{\alpha/2}$	양측검정

참고 : 대표본 가정하에서 모평균 μ에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

이고 이는 유의수준 α에서 양측검정

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

에 대한 채택역 (acceptance region, 기각역의 여사건)과 같게 된다.

(2) 모평균 μ에 대한 검정 (소표본, 정규모집단) - 일표본 t-검정

표본의 크기가 25보다 작은 소표본인 경우 임의표본 $X_1, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$를 따른다고 가정하자. 만약 소표본이면서 정규분포를 따르지 않을 경우는 4.3절에서 소개하는 비모수적 방법을 사용해야 한다. 모평균에 대한 검정으로 귀무가설 $H_0 : \mu = \mu_0$ 에 대하여 3가지 형태의 대립가설이 가능하다. 단, 여기서 검정통계량은 $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ 이며 이는 귀무가설하에서 자유도 $n-1$인 t 분포를 따르는데 이를 일표본 t-검정 (one-sample t-test)이라 한다.

귀무가설	대립가설	기각역	비고
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$t > t_\alpha$	단측검정
$H_0 : \mu = \mu_0$	$H_1 : \mu < \mu_0$	$t < -t_\alpha$	단측검정
$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$\|t\| > t_{\alpha/2}$	양측검정

예제 : 피마 인디언 자료(Pima.te)에서 혈당(glu)이 한국인 참조표준과 유의한 차이가 있는지를 양측검정하자. 한국인 평균혈당이 98로 알려져 있다고 가정하자. 일표본 t-검정을 실시하자.

```
t.test(Pima.te$glu, mu=98)
#>
#>  One Sample t-test
#>
#> data:  Pima.te$glu
#> t = 12.7, df = 331, p-value < 2.2e-16
#> alternative hypothesis: true mean is not equal to 98
#> 95 percent confidence interval:
#>  115.9661 122.5520
#> sample estimates:
#> mean of x
#>   119.259
```

(3) 모비율 p에 대한 검정 (대표본)

임의표본 $X_1, \cdots, X_n$은 성공의 확률 p인 베르누이 시행이고 $np > 15,\ n(1-p) > 15$를 만족한다고 가정한다. 모비율에 대한 검정으로 귀무가설 $H_0 : p = p_0$에 대하여 3가지 형태의 대립가설에 대한 기각역은 다음과 같다. 검정통계량은 $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$ 인데 $\hat{p} = \sum X_i/n$ 이며, 이 또한 일표본 Z-검정 (one-sample Z-test)이라 한다.

귀무가설	대립가설	기각역	비고
$H_0 : p = p_0$	$H_1 : p > p_0$	$Z > z_\alpha$	단측검정
$H_0 : p = p_0$	$H_1 : p < p_0$	$Z < -z_\alpha$	단측검정
$H_0 : p = p_0$	$H_1 : p \neq p_0$	$\|Z\| > z_{\alpha/2}$	양측검정

예제 : 폐암환자 자료(lung)에서 사망자(status=2)는 228명 중 165명(72.4%)이다. 50%보다 유의하게 큰지를 단측검정하자. 일표본 비율차이검정을 실시하자.

```
death = as.numeric(lung$status ==2)


x = sum(death); x
#> [1] 165
n = length(death); n
#> [1] 228
p = x/n; p
#> [1] 0.7236842

prop.test(x, n, p=0.5, alternative="greater", correct=F)
#>
#>  1-sample proportions test without continuity correction
#>
#> data:  x out of n, null probability 0.5
#> X-squared = 45.632, df = 1, p-value = 7.136e-12
#> alternative hypothesis: true p is greater than 0.5
#> 95 percent confidence interval:
#>  0.6725643 1.0000000
#> sample estimates:
#>         p
#> 0.7236842
```

(4) 모분산 σ^2에 대한 검정 - 카이제곱 검정

임의표본 $X_1, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$를 따른다고 가정하자. 모분산에 대한 검정으로 귀무가설 H_0 : $\sigma^2 = \sigma_0^2$ 에 대하여 3가지 형태의 대립가설에 대한 기각역은 다음과 같다. 검정통계량 $\chi^2 = (n-1)s^2/\sigma_0^2$ 은 귀무가설하에서 자유도 $n-1$인 카이제곱분포를 따른다. 여기서 $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$이다.

귀무가설	대립가설	기각역	비고
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{\alpha}$	단측검정
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{1-\alpha}$	단측검정
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2}$ 또는 $\chi^2 < \chi^2_{1-\alpha/2}$	양측검정

4.2.2 이표본 검정

이표본 검정이란 서로 독립인 두 모집단에서 관심있는 모수들 간의 비교에 대한 검정이다. 예를 들어, 두 모집단의 모평균이 같은가에 대한 검정이다.

(1) 이표본 모평균의 차이에 대한 검정 (대표본) - 이표본 Z-검정

표본에 대한 가정

- $X_1, \cdots, X_{n_1}$: 평균 μ_1, 분산 σ_1^2 을 갖는 모집단 1로 부터의 표본

$$\bar{X} = \sum_{i=1}^{n_1} X_i/n_1 \, , \qquad s_1^2 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2/(n_1 - 1)$$

- $Y_1, \cdots, Y_{n_2}$: 평균 μ_2, 분산 σ_2^2 을 갖는 모집단 2로 부터의 표본

$$\bar{Y} = \sum_{i=1}^{n_2} Y_i/n_2 \, , \qquad s_2^2 = \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2/(n_2 - 1)$$

- 두 모집단은 서로 독립이다.
- n_1, n_2는 충분히 크다. (각각 25이상)

목표 : $\mu_1 - \mu_2$ 에 대한 추론

점 추정치 : $\bar{X} - \bar{Y}$

구간 추정

$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$, $Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$\Rightarrow$ $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ 는 근사적으로 $N(0,1)$

$\Rightarrow$ $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ 역시 근사적으로 $N(0,1)$

$\Rightarrow$ $\mu_1 - \mu_2$ 에 대한 $100 \times (1-\alpha)\%$ 근사적 신뢰구간은

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

귀무가설 $H_0 : \mu_1 - \mu_2 = \mu_0$ 에 대하여 3가지 형태의 대립가설이 가능하며 그에 따른 기각역의 형태는 다음과 같이 요약할 수 있다. 검정통계량은

$$Z = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

이고 귀무가설 하에서 근사적으로 표준정규분포를 따른다.

귀무가설	대립가설	기각역	비고
$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_1 : \mu_1 - \mu_2 > \mu_0$	$Z > z_\alpha$	단측검정
$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_1 : \mu_1 - \mu_2 < \mu_0$	$Z < -z_\alpha$	단측검정
$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_1 : \mu_1 - \mu_2 \neq \mu_0$	$\lvert Z \rvert > z_{\alpha/2}$	양측검정

(2) 이표본 모평균의 차이에 대한 검정 (소표본, 정규모집단) - 이표본 t-검정

표본에 대한 가정 (정규성, 독립성, 등분산성)

- $X_1, \cdots, X_{n_1} : N(\mu_1, \sigma^2)$ 부터의 표본,
  $\bar{X} = \sum_{i=1}^{n_1} X_i / n_1$, $s_1^2 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2 / (n_1 - 1)$
- $Y_1, \cdots, Y_{n_2} : N(\mu_2, \sigma^2)$ 부터의 표본,
  $\bar{Y} = \sum_{i=1}^{n_2} Y_i / n_2$, $s_2^2 = \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 / (n_2 - 1)$
- 두 표본은 서로 독립이다.

목표 : $\mu_1 - \mu_2$ 에 대한 추론

점 추정치 : $\bar{X} - \bar{Y}$

구간 추정

$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2, Var(\bar{X} - \bar{Y}) = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$

$\Rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$

$\Rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$

단, $s_p = \frac{(n-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$: 합동분산(合同分散) : pooled variance

$\Rightarrow$ $\mu_1 - \mu_2$ 에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

귀무가설 $H_0 : \mu_1 - \mu_2 = \mu_0$ 에 대하여 3가지 형태의 대립가설이 가능하며 그에 따른 기각역의 형태는 다음과 같이 요약할 수 있다. 검정통계량은

$$t = \frac{(\bar{X} - \bar{Y}) - \mu_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

으로 주어지고 이는 귀무가설 하에서 자유도 $n_1 + n_2 - 2$인 t-분포를 따른다.

귀무가설	대립가설	기각역	비고
$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_1 : \mu_1 - \mu_2 > \mu_0$	$t > t_\alpha$	단측검정
$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_1 : \mu_1 - \mu_2 < \mu_0$	$t < -t_\alpha$	단측검정
$H_0 : \mu_1 - \mu_2 = \mu_0$	$H_1 : \mu_1 - \mu_2 \neq \mu_0$	$\lvert t \rvert > t_{\alpha/2}$	양측검정

참고 : 표본이 정규성, 독립성은 만족하나 등분산성 (homoscedasticity)을 만족하지 않는 경우에 대한 추론을 소개한다. 만약 $0.5 < s_1/s_2 < 2$을 만족하면 등분산성이 성립된다고 가정해도 좋으나 이 조건을 만족하지 않으면 이분산성 (heteroscedasticity) 가정하에서 새터쓰웨이트 근사 (Satterthwaite approximation)를 사용한다. 즉, 다음과 같은 사실에 근거하여 이표본 t-검정을 실시한다.

$$t^* = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu)$$

여기서,

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{s_2^2}{n_2})^2}$$

인데 정수값이 아니므로 반올림하여 사용한다.

예제 : 피마 인디언 자료(Pima.te)에서 당뇨병 유무(type)에 따라 나이(age)에 유의한 차이가 있는지를 양측검정하자. 독립표본 t-검정을 실시하자(등분산 가정).

```
t.test(age ~ type, Pima.te, var.equal=T)
#>
#>  Two Sample t-test
#>
#> data:  age by type
#> t = -5.3594, df = 330, p-value = 1.571e-07
#> alternative hypothesis: true difference in means between group No and group Yes is
#> 95 percent confidence interval:
#>  -8.748356 -4.050508
#> sample estimates:
#>  mean in group No mean in group Yes
#>          29.21525          35.61468
```

예제 -계속) 등분산 가정이 만족하는 지를 검정하자.

```
var.test(age ~ type, Pima.te)
#>
#>  F test to compare two variances
#>
#> data:  age by type
#> F = 0.95078, num df = 222, denom df = 108, p-value = 0.7461
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#>  0.6795178 1.3057539
#> sample estimates:
#> ratio of variances
#>          0.9507776
```

등분산 가정이 만족되지 않는 경우 Welch 검정을 하거나 비모수 검정을 실시할 수 있다.

```
t.test(age ~ type, Pima.te, var.equal=F)
#>
#>  Welch Two Sample t-test
#>
#> data:  age by type
#> t = -5.313, df = 209.68, p-value = 2.746e-07
#> alternative hypothesis: true difference in means between group No and group Yes is not e
#> 95 percent confidence interval:
#>  -8.773874 -4.024990
#> sample estimates:
#>  mean in group No mean in group Yes
#>          29.21525          35.61468
```

(3) 쌍체비교 (matched pair comparisons)

자료가 쌍(pair)으로 관측되는 경우가 있다. 예를 들어, 개인별로 왼쪽과 오른쪽의 시력 차이가 있는가? 아스피린의 복용전과 복용후 혈압에 차이는 있는가? 등이다. 이렇게 쌍으로 관측되는 자료에 다음과 같은 가정을 한다.

- $D_1, \cdots, D_n$: 임의 표본, 단 $D_i = X_i - Y_i, i = 1, \cdots, n$
- 표본평균 $\bar{D} = \sum_{i=1}^{n_1} D_i/n$, 표본분산 $s_D^2 = \sum_{i=1}^{n_1}(D_i - \bar{D})^2/(n-1)$
- $E(D_i) = \delta, Var(D_i) = \sigma_D^2$
- $Z = \frac{\bar{D}-\delta}{s_D/\sqrt{n}}$ 는 대표본인 경우 근사적으로 $N(0,1)$

예제 : 12마리의 마우스를 대상으로 약물처리 전과 후의 종양크기를 기록하였다. 약물처리 전과 후에 유의한 차이가 있는지를 양측검정하자. 대응표본 t-검정을 실시하자.

```
pre = c(16.4, 10.3, 15.8, 16.5, 12.5, 8.3, 12.1, 10.1, 12.9, 12.6, 17.3, 9.4)
post = c(14.3, 9.8, 16.9, 17.2, 10.5, 7.9, 12.4, 8.6, 13.1, 11.6, 15.5, 8.6)
t.test(pre, post, paired=T)
#>
#>  Paired t-test
```

```
#>
#> data:  pre and post
#> t = 2.0947, df = 11, p-value = 0.06015
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -0.03297204  1.33297204
#> sample estimates:
#> mean of the differences
#>                    0.65
```

(4) 이표본 모비율의 차이에 대한 검정 (대표본) - 이표본 Z-검정

자료의 구조

	시행횟수	성공횟수	실패횟수
모집단 1	n_1	X	n_1-X
모집단 2	n_2	Y	n_1-Y

가정 : $X \sim B(n_1, p_1),\ Y \sim B(n_2, p_2)$, $X,\ Y$: 독립, $n_1 > 25, n_2 > 25$

목표 : $p_1 - p_2$에 대한 추론

점 추정치 : $\hat{p}_1 - \hat{p}_2, \hat{p}_1 = X/n_1, \hat{p}_2 = Y/n_2$

구간 추정

$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2\ ,\ \ Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$,

$\Rightarrow\ \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}$ 는 근사적으로 $N(0,1)$

$\Rightarrow\ \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$ 역시 근사적으로 $N(0,1)$

$\Rightarrow\ p_1 - p_2$ 에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

귀무가설 $H_0 : p_1 - p_2 = p_0$ 에 대하여 3가지 형태의 대립가설이 가능하며 그에 따른 기각역의 형태는 다음과 같이 요약할 수 있다. 검정통계량은 $Z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$ 으로 주어지고 귀무가설 하에서 근사적으로 표준정규분포를 따른다.

귀무가설	대립가설	기각역	비고
$H_0 : p_1 - p_2 = p_0$	$H_1 : p_1 - p_2 > p_0$	$Z > z_\alpha$	단측검정
$H_0 : p_1 - p_2 = p_0$	$H_1 : p_1 - p_2 < p_0$	$Z < -z_\alpha$	단측검정
$H_0 : p_1 - p_2 = p_0$	$H_1 : p_1 - p_2 \neq p_0$	$\|Z\| > z_{\alpha/2}$	양측검정

예제 : 흑색종 환자자료(Melanoma)에서 성별(sex)에 따라 궤양여부(ulcer)에 차이가 있는지를 양측검정하자. 독립표본 모비율검정을 실시하자.

```
sex <- factor(Melanoma$sex, 0:1, c("Male","Female"))
ulcer <- factor(Melanoma$ulcer, 1:0, c("Yes","No"))


tab <- table(ulcer,sex); tab
#>       sex
#> ulcer Male Female
#>   Yes   47     43
#>   No    79     36
addmargins(tab)
#>       sex
#> ulcer Male Female Sum
#>   Yes   47     43  90
#>   No    79     36 115
#>   Sum  126     79 205
prop.table(tab, 2)*100
#>       sex
#> ulcer     Male   Female
#>   Yes 37.30159 54.43038
#>   No  62.69841 45.56962
```

```
prop.test(c(47,43), c(126, 79), correct=F)
#>
#>  2-sample test for equality of proportions without continuity
#>  correction
#>
#> data:  c(47, 43) out of c(126, 79)
#> X-squared = 5.7845, df = 1, p-value = 0.01617
#> alternative hypothesis: two.sided
#> 95 percent confidence interval:
#>  -0.3098209 -0.0327549
#> sample estimates:
#>    prop 1    prop 2
#> 0.3730159 0.5443038
chisq.test(tab, correct=F)
#>
#>  Pearson's Chi-squared test
#>
#> data:  tab
#> X-squared = 5.7845, df = 1, p-value = 0.01617
```

4.2.3 다중 검정

다중검정 (multiple hypothesis testing)이란 3개 이상의 모집단에 대한 가설 검정이다. 예를 들어, k개 모집단의 모평균 (또는 처리효과)이 모두 같은지에 대한 검정, 즉

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

이 다중검정으로 귀무가설이 기각 될 수도 있고 그렇지 않을 수도 있다. 만약 귀무가설이 기각된다면 k개의 처리효과가 모두 같은 것은 아니라는 것이다. 그렇다면 어떤 모집단들의 평균이 다른가? 서로 다른 두 처리효과간의 차이를 보려면 $\binom{k}{2} = d$ 경우의 수 만큼의 차이를 계산해야 한다. 이를 위해 $\binom{k}{2} = d$ 경우의 수 만큼의 동시신뢰구간 (simultaneous C.I.) $\mu_i - \mu_j,\ i < j$ 을 계산해야 한다. 신뢰구간이 0을 포함하면 두 처리효과는 차이가 없음을 나타내고, 신뢰구간이 0을 포함하지 않으면 두 처리효과는 차이가 있음을 나타낸다.

예를 들어 k=3일 경우 $\mu_1 - \mu_2,\ \mu_2 - \mu_3,\ \mu_1 - \mu_3$ 에 대한 3개의 동시신뢰구간이 필요 각각에 대한 95% 신뢰구간을 $A_1,\ A_2,\ A_3$ 라 하면

$$P(A_1) = P(A_2) = P(A_3) = 0.95$$

그러나,

$$P(A_1 \cap A_2 \cap A_3) = (0.95)^3 = 0.857$$

즉, 동시에 3개의 신뢰구간을 구하면 그 신뢰도는 0.95에서 0.857로 낮아진다. 이를 보완하기 위해 주로 본페로니 보정법 (Bonferroni's adjustment)을 사용하는데 구체적으로 다음과 같은 논리에 근거한다. 일반적으로 d개의 동시신뢰구간 $A_1, \cdots, A_d$ 이 $1-\alpha$의 신뢰도를 가지기 위해 다음의 본페로니 부등식 (Bonferroni inequality)을 이용한다.

$$1 - \alpha = P(\cap_{i=1}^{d} A_i) = 1 - P(\cup_{i=1}^{d} A_i^c) \geq 1 - \sum_{i=1}^{d} P(A_i^c) = 1 - d\alpha^*$$

따라서, α/d를 α 대신 사용하면 원하는 $1-\alpha$의 신뢰도를 얻게 된다. 예를 들어, 3개의 동시신뢰구간 ($d = 3$)이 95% (즉, $\alpha = 0.05$)의 신뢰도를 얻으려면 각 신뢰구간은 $\alpha^* = \alpha/3 = 0.05/3 = 0.0167$ 로 계산하면 된다. 따라서, $\mu_i - \mu_j$에 대한 신뢰구간을

구하려면 이 모수가 포함된 통계량과 그 통계량의 분포를 구해야 되는데 이는 다음과 같은 과정을 거친다.

점 추정치 : $\bar{Y}_i - \bar{Y}_j$

$E(\bar{Y}_i - \bar{Y}_j) = \mu_i - \mu_j$

$Var(\bar{Y}_i - \bar{Y}_j) = \sigma^2(\frac{1}{n_i} + \frac{1}{n_j})$

$\Rightarrow \frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t(n-k)$ 단, $s^2 = SSE/(n-k)$는 σ^2의 불편추정치

결론적으로 $\mu_i - \mu_j,\ i < j$에 대한 $100(1-\alpha)\%$ d개의 동시신뢰구간은

$$(\bar{Y}_i - \bar{Y}_j) \pm t_{\alpha/2d}\, s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

으로 주어진다.

예제 : 폐암 환자자료(lung)에서 ECOG(ph.ecog)에 따라 지난 6개월간 체중감소량(wt.loss)에 차이가 있는지를 양측검정하자.

```
d = na.omit(lung[,c('wt.loss','ph.ecog')])
y = d$wt.loss
group = d$ph.ecog
group = factor(ifelse(group >=2, 2, group))

tapply(y, group, function(x) sprintf("%.2f (%.2f)", mean(x), sd(x)))
#>               0               1               2
#>   "6.00 (8.81)" "10.59 (14.18)" "12.67 (14.28)"

oneway.test(y ~ group, var.equal=T)
#>
#>  One-way analysis of means
#>
#> data:  y and group
#> F = 3.9847, num df = 2, denom df = 210, p-value = 0.02002
```

```
pairwise.t.test(y, group, p.adjust = "none", pool.sd=F, var.equal=T)
#>
#>  Pairwise comparisons using t tests with non-pooled SD
#>
#> data:  y and group
#>
#>   0      1
#> 1 0.0235 -
#> 2 0.0036 0.4086
#>
#> P value adjustment method: none
pairwise.t.test(y, group, p.adjust = "bonferroni", pool.sd=F, var.equal=T)
#>
#>  Pairwise comparisons using t tests with non-pooled SD
#>
#> data:  y and group
#>
#>   0     1
#> 1 0.070 -
#> 2 0.011 1.000
#>
#> P value adjustment method: bonferroni
pairwise.t.test(y, group, p.adjust = "fdr", pool.sd=F, var.equal=T)
#>
#>  Pairwise comparisons using t tests with non-pooled SD
#>
#> data:  y and group
#>
#>   0     1
#> 1 0.035 -
#> 2 0.011 0.409
```

```
#>
#> P value adjustment method: fdr
```

4.3 여러 가지 검정법 (비모수적 방법)

우리는 앞에서 주어진 검정 통계량이 귀무가설하에서 특정 분포를 따르는 모수적 방법을 소개하였다. 대표본 $(n > 25)$인 경우 중심극한정리에 의해 근사적으로 정규분포를 이용하였고, 소표본이면서 정규모집단인 경우 t-분포를 이용하였다. 소표본이면서 정규모집단이 아닌 경우의 검정은 흔히 비모수적 방법 (nonparametric methods)을 사용하는데 다음과 같은 귀무가설과 대립가설에 대한 검정이다.

귀무가설 : 두 모집단 A와 B의 분포는 동일하다.

대립가설 : 모집단 A의 분포는 모집단 B의 분포를 오른쪽으로 δ만큼 이동시킨 것이다.

$$H_0 : \delta = 0 \ \ vs \ \ H_1 : \delta > 0$$

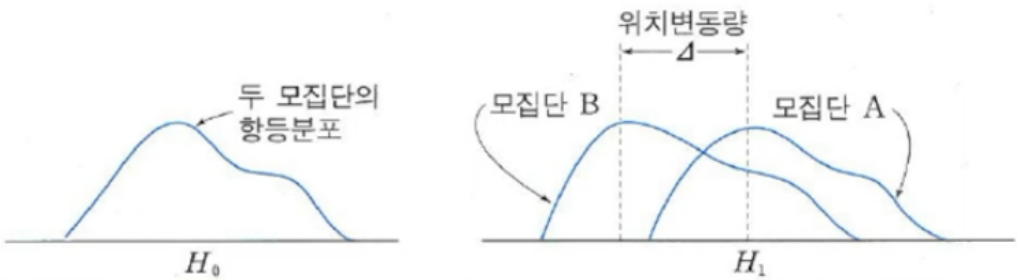


그림 4.1: 비모수적 방법

4.3.1 윌콕슨 순위합 검정

$X_{11}, ..., X_{1n_A}$ 와 $X_{21}, ..., X_{2n_B}$ 를 각각 연속인 모집단 A와 B에서의 독립확률표본이라 하자. 귀무가설 "H_0: 두 모집단의 분포는 같다"을 검정하는 윌콕슨 순위합 검정 (Wilcoxon rank sum test) 절차는 다음과 같다.

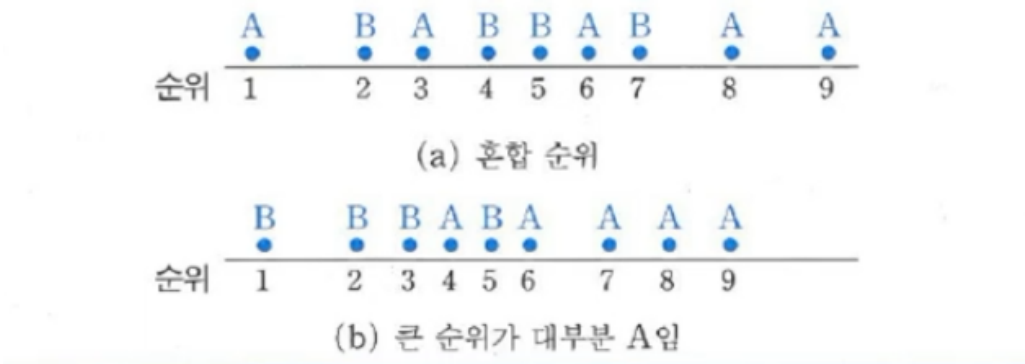


그림 4.2: 비모수적 방법

(1) $n = n_A + n_B$개의 측정치를 크기순으로 나열하고 순위를 매긴다.

(2) 순위합 W_A를 계산한다.

(3) 대립가설에 따른 기각역은 아래와 같다.

대립가설	기각역
H_1: 모집단 A가 모집단 B의 오른쪽(큰쪽)에 위치	W_A의 큰쪽 꼬리부분
H_1: 모집단 A가 모집단 B의 왼쪽(작은쪽)에 위치	W_A의 작은쪽 꼬리부분
H_1: 모집단이 서로 다르다.	W_A의 양쪽 꼬리부분에서 같은 확률을 갖도록 설정

위의 기각역에 의해 정확하게 p-value를 계산할 수 있으나 표본의 크기가 충분히 클 경우 윌콕슨 순위합 검정통계량은 다음과 같은 정규근사화에 의해

$$Z = \frac{W_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B (n_A + n_B + 1)/12}}$$

는 귀무가설 H_0하에서 근사적으로 $N(0,1)$분포를 따른다.

예제 : 종양을 이식한 마우스를 대상으로 3마리는 처리A를, 2마리는 처리B를 실시한 후 14일 뒤 종양크기를 재어 보았다. 두 처리방법간에 유의한 차이가 있는지를 양측검정하자.

```
g1 = data.frame(group="A", value=c(31.8, 39.1))
g2 = data.frame(group="B", value=c(35.5, 27.6, 21.3))
data = rbind(g1, g2); data
#>   group value
#> 1     A  31.8
#> 2     A  39.1
#> 3     B  35.5
#> 4     B  27.6
#> 5     B  21.3

wilcox.test(value ~ group, data)
#>
#>  Wilcoxon rank sum exact test
#>
#> data:  value by group
#> W = 5, p-value = 0.4
#> alternative hypothesis: true location shift is not equal to 0
```

4.3.2 부호 검정과 부호순위 검정

쌍체비교 (Matched Pair Comparisons)에 대한 예로서

- 개인별로 왼쪽과 오른쪽의 시력 차이가 있는가?
- 아스피린의 복용전과 복용후 혈압에 차이는 있는가?

등이 있으며 모수적 검정법은 이미 소개한 바 있다. 우선 자료에 대한 가정으로

- $D_1, \cdots, D_n$: 임의 표본, 단 $D_i = X_i - Y_i, i = 1, \cdots, n$
- 표본평균 $\bar{D} = \sum_{i=1}^{n} D_i/n$, 표본분산 $s_D^2 = \sum_{i=1}^{n}(D_i - \bar{D})^2/(n-1)$

- $E(D_i) = \delta, Var(D_i) = \sigma_D^2$
- $Z = \frac{\bar{D}-\delta}{s_D/\sqrt{n}}$ 는 대표본인 경우 근사적으로 $N(0,1)$

을 생각할 수 있는데 만약 대표본이 아닌 경우는 비모수적 방법을 주로 사용하며 여기서는 부호검정과 부호순위 검정을 소개한다.

부호검정 (sign test)

가설은 $H_0 : \delta = 0 \ \ vs \ \ H_1 : \delta > 0$로 표현할 수 있고 (대립가설은 원하는 주장에 따라 $\delta < 0$이 될 수도 있음), 이에 대한 검정통계량 $S = \sum_{i=1}^{n} I(D_i > 0)$인데 이는 $D_1, \cdots, D_n$ 중에서 양의 부호를 가진 것의 개수가 된다. 귀무가설 하에서 $I(D_i > 0)$는 성공의 확률 1/2인 베르누이 시행이므로 $S \sim B(n, 0.5)$이 된다. 따라서, 가설을 성공의 확률 p를 이용하여 표현하면 $H_0 : p = 1/2 \ \ vs \ \ H_1 : p > 1/2$이 되고 이에 대한 기각역은 적절한 c (주어진 유의수준에 의해 정해짐) 대해 $S > c$ 이면 귀무가설을 기각하게 된다. 만약 표본의 크기가 충분히 크면 표준화시킬 경우 근사적으로 정규분포를 따르게 된다. 즉,

$$Z = \frac{S - n/2}{\sqrt{n/4}}$$

는 귀무가설 H_0하에서 근사적으로 $N(0,1)$분포를 따른다.

부호순위검정 (Wilcoxon signed-rank test)

쌍체비교에 대한 또 다른 비모수 검정으로 부호순위 검정을 소개하는데 이는 다음의 절차에 의해 시행된다.

(1) 차이 $D_i = X_{i1} - X_{i2}, i = 1, \cdots, n$를 계산한다.
(2) D_i의 절대값을 크기순으로 배열하여 순위를 매기고 해당하는 부호를 붙인다.
(3) 부호순위통계량 $T^+ = \sum_{i=1}^{n} I(D_i > 0) r(|D_i|)$: 양의 값을 갖는 D_i들 중에서 D_i의 절대값 순위의 합을 계산한다.
(4) 대립가설이 D_i가 0보다 큰지, 작은지, 아니면 다른지에 따라서 기각역을 T^+의 큰쪽 꼬리부분, 작은쪽 꼬리부분, 양쪽 꼬리부분에서 설정한다. 만약 표본의 크기가 크면

$$Z = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

는 귀무가설 H_0하에서 근사적으로 $N(0,1)$분포를 따른다.

예제 : 12마리의 마우스를 대상으로 약물처리 전과 후의 종양크기를 기록하였다. 약물처리 전과 후에 유의한 차이가 있는지를 양측검정하자. 윌콕슨 부호순위 검정을 실시하자.

```
pre = c(16.4, 10.3, 15.8, 16.5, 12.5, 8.3, 12.1, 10.1, 12.9, 12.6, 17.3, 9.4)
post = c(14.3, 9.8, 16.9, 17.2, 10.5, 7.9, 12.4, 8.6, 13.1, 11.6, 15.5, 8.6)

wilcox.test(pre, post, paired=T)
#>
#>  Wilcoxon signed rank exact test
#>
#> data:  pre and post
#> V = 62, p-value = 0.07715
#> alternative hypothesis: true location shift is not equal to 0
```

4.3.3 순위상관계수

이변량 자료 $(x_1, y_1), \cdots, (x_n, y_n)$: 서로독립이고 연속형 분포를 가짐

피어슨 표본 상관계수 (Pearson's sample correlation coefficient)

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

- R_i : $X_1, \cdots, X_n$ 중에서 X_i의 순위
- S_i : $Y_1, \cdots, Y_n$ 중에서 Y_i의 순위

스피어만의 순위상관계수 (Spearman's rank correlation coefficient)

$$\begin{aligned} r_{SP} &= \frac{\sum(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum(R_i - \bar{R})^2}\sqrt{\sum(S_i - \bar{S})^2}} \\ &= \frac{\sum(R_i - \frac{(n+1)}{2})(S_i - \frac{(n+1)}{2})}{n(n^2-1)/12} \end{aligned}$$

r_{SP}의 대표본 근사

$\sqrt{n-1}r_{SP}$는 귀무가설 H_0하에서 근사적으로 $N(0,1)$분포를 따른다.

예제 : 평가 순서와 평가 점수와의 상관관계를 알고 싶다.

```
x1 <- c(5, 2, 3, 1, 6, 4)
x2 <- c(47, 32, 29, 28, 56, 38)

r1 <- rank(x1); r1
#> [1] 5 2 3 1 6 4
r2 <- rank(x2); r2
#> [1] 5 3 2 1 6 4

r_sp <- cov(r1, r2)/sqrt(var(r1)*var(r2)); r_sp
#> [1] 0.9428571

cor.test(x1, x2, method="spearman")
#>
#>  Spearman's rank correlation rho
#>
#> data:  x1 and x2
#> S = 2, p-value = 0.01667
#> alternative hypothesis: true rho is not equal to 0
#> sample estimates:
#>       rho
#> 0.9428571
```

제2부 의학자료분석을 위한 통계적 방법론

제 5 장

회귀모형

5.1 통계적 모형

5.1.1 수학적 모형과 통계적 모형

어떤 변수들이 또 다른 변수와 어떤 관계에 있는지 규명하고자 하는 연구분야는 매우 많다. 흡연여부, 나이, 음주량 등이 폐암 생존율에 어떤 영향을 미치는지 알고 싶으며, 비만지수, 나이 등의 변수가 혈압과 어떤 관계가 있는지 궁금하다. 즉, p개의 변수 $X_1, \cdots, X_p$가 변수 Y에 어떤 영향을 미치는지 알고 싶다. 이를 도식화하면

$$input : X_1, \cdots, X_p \quad \rightarrow \quad function\ f \quad \rightarrow \quad output : Y$$

로 나타낼 수 있으며 이러한 연구에서 입력과 출력의 관계를 나타내주는 함수를 규명하는 것이 연구의 가장 주된 목적이다. 이러한 함수를 흔히 인공지능 (artificial intelligence)이라 부르기도 한다.

이러한 함수를 흔히 모형 (model)이라 부르며, 모형에는 크게 두 가지 형태의 모형이 있다.

(i) 결정적 (수학적) 모형 : input 과 output의 관계가 오차없이 명확함

$$Y = f(X_1, \cdots, X_p)$$

(예)

- 힘 = 질량 x 가속도
- 화씨 = 32 + 1.8 x 섭씨

(ii) 통계적 모형 : output이 input에 의해 영향을 받는 경향을 보이며 언제나 오차를 수반함

$$Y = f(X_1, \cdots, X_p) + \epsilon$$

(예)

- 매출액 = 100 + 0.1 x 광고비 + ϵ
- 수축기 혈압 = 110 + 0.1 x 연령 + 0.15 x 몸무게 + ϵ

5.1.2 통계적 모형의 종류

(1) 회귀모형 회귀모형은 다음과 같이 정의된다.

$$Y = f(X_1, \cdots, X_p) + \epsilon$$

- Y : 반응변수 (response variable), 종속변수 (dependent variable)
- $X_1, \cdots, X_p$: 공변량, 설명변수 (covariates)
- ϵ : 오차항 (error term)
- $f(\cdot)$: 회귀함수 (regression function)

회귀분석이란 반응변수와 설명변수를 관측하여 회귀함수를 추정하고, 추정된 회귀식을 바탕으로 새로운 설명변수가 주어졌을 때 반응변수의 값을 예측하는 일련의 과정을 일컫는다.

회귀모형은 여러 종류가 있으며 다음과 같이 분류한다.

(i) 모수회귀모형 (parametric regression model)

- 중선형회귀모형 (multiple linear regression model)

$$f(X_1, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- 비선형회귀모형 (non-linear regression model)

$$f(X) = \frac{\beta_0 X}{\beta_1 + X}$$

- k-차 다항회귀모형 (k-th degree polynomial regression model)

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

- 로지스틱회귀모형 (logistic regression model)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \quad p = P(Y = 1), \ Y \sim B(1, p)$$

- 로그선형모형 (log-linear regression model)

$$log(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \quad Y \sim Poisson(\mu)$$

(ii) 비모수회귀모형 (nonparametric regression model) :
회귀함수 f의 형태를 구체적으로 명시하지 않고 자료에 근거하여 함수의 추정치를 계산한다. 비모수회귀모형의 종류로 평활 스플라인 (smoothing spline), 국소다항회귀 (local polynomial regression) 등이 있다.

5.2 훈련자료와 시험자료

회귀모형을 구축하기 위해 주어진 자료를 모두 사용하여 모형을 적합하는 것은 바람직하지 않다. 따라서, 많은 경우 자료의 일부만 적합에 사용하고 나머지 자료는 적합된 회귀모형이 제대로 되었는지 시험하는데 사용될 수 있다. 이처럼 모형의 구축에 사용된 자료를 훈련자료 (training data), 구축된 모형을 여러 측면에서 시험하는데 사용되는 자료를 시험자료 (test data)라 부른다. 흔히 전체 자료의 70% 내외를 훈련자료, 나머지 30% 내외를 시험자료로 사용한다.

시험자료를 이용하여 구축된 모형의 신뢰상 정도는 흔히 예측오차 (prediction error)로 판별한다. 예측오차를 정의하기 위해 훈련자료의 설명변수와 반응변수를 $(X_i, Y_i),\ i = 1, \cdots, m$, 시험자료의 설명변수와 반응변수를 $(Z_j, W_j),\ j = 1, \cdots, n$라 하면 j번째 예측오차는

$$W_j - \hat{f}_X(Z_j)$$

으로 주어지는데 여기서 회귀함수의 추정치 $\hat{f}_X$는 훈련자료에 근거한 것이다. 즉, 예측오차란 시험자료의 반응변수값 W와 훈련자료로 추정한 회귀함수에 시험자료의 설명변수를 대입했을 때의 추정치 $\hat{f}_X(Z)$ 와의 차이를 나타낸다.

여러 모형을 비교하는 경우 흔히 이러한 예측오차를 제곱하여 합한 예측오차제곱합 (sum of squares of the prediction errors)을 시험자료의 개수로 나눈 값, 즉

$$SSPE = \frac{1}{n}\sum_{j=1}^{n}(W_j - \hat{f}_X(Z_j))^2$$

을 주로 사용한다. 이 값은 평균오차제곱합 (MSE; mean squared error)

$$MSE = E[(W - f(Z))^2]$$

의 추정치라 할 수 있다. 논문에서 흔히 RMSE (root mean squared error)로 표현하는 것은 실제로 $\sqrt{PE}$를 의미한다.

한편, 예측오차의 제곱 대신 절대값을 취하여 사용하기도 하는데 이를 예측오차절대값 합 (sum of absolute of the prediction error)라고 하며 이는

$$SAPE = \frac{1}{n}\sum_{j=1}^{n}|W_j - \hat{f}_X(Z_j)|$$

로 표현할 수 있다.

5.3 선형회귀모형

선형회귀모형 (linear regression model)이란 설명변수 (covariates)들의 선형결합 (linear combination)이 반응변수 (response variable)를 설명할 수 있다는 가정하에 설정된 모형이다. 즉,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \ + \ \epsilon$$

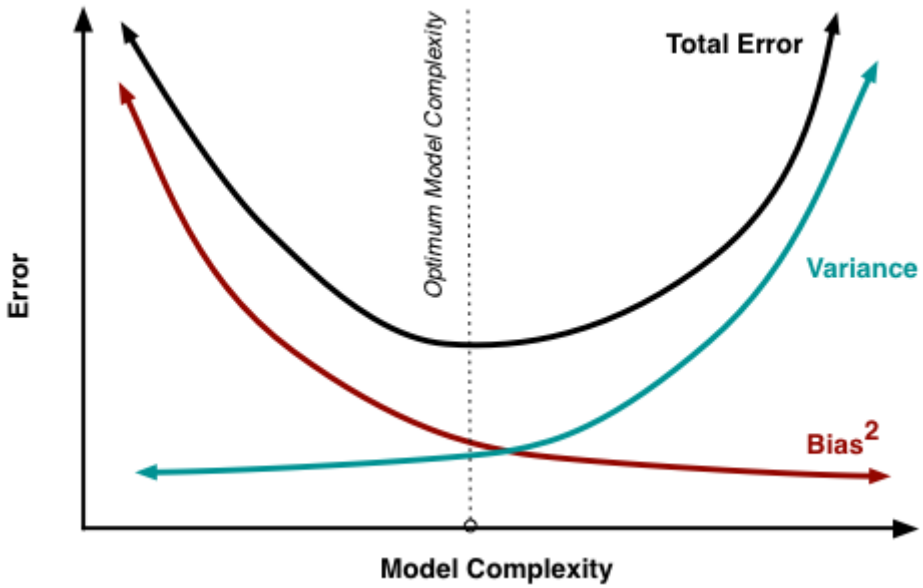


그림 5.1: 분산과 편의, 모형의 복잡도

을 가정한 것으로 각 기호는 다음과 같다. - Y: 반응변수 - $X_1, \cdots, X_p$: 설명변수 또는 공변량 - ϵ : 오차항 (error term)으로 흔히 $N(0, \sigma^2)$을 따른다고 가정함 - $\beta_0, \beta_1, \cdots, \beta_p$: 회귀계수 (regression coefficients)로 추정되어야할 모수 모수 $\beta_0, \beta_1, \cdots, \beta_p, \sigma^2$ 을 추정하기 위해 n 개의 자료를 관측한다. 즉,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \,, \quad i = 1, \cdots, n$$

적합된 회귀계수를 $\hat{\beta}_j, \ j = 1, \cdots, p$ 라 하면 $\hat{\beta}_j$는 X_j를 제외한 다른 공변량들이 고정되어 있다는 가정 하에 X_j가 한 단위 증가할 때 반응변수의 증가값으로 해석할 수 있다.

5.4 선형회귀모형의 변환

5.4.1 능형회귀

능형회귀 (ridge regression)는 Hoerl and Kennard (1970)에 의해 소개되었는데 $\hat{\beta}$에 대한 추정치로 최소제곱추정치 대신

$$\hat{\beta}(\theta) = (\mathbf{X}'\mathbf{X} + \theta\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{5.1}$$

을 사용하는 추정치다. 여기서 θ는 양수로서 추정되어야 할 모수이며 흔히 편의모수 (biasing parameter) 또는 축소모수 (shrinkage parameter)라 한다. 능형추정량의 아이디어는, 0에 가까운 수의 역수는 매우 불안정하고 여기에 적당한 양수를 더해주면 그 역수는 안정해지는 것이다.

능형추정치 $\hat{\beta}(\theta)$가 모수 θ를 포함하고 있으므로, 완전한 추정치가 되기 위해선 θ의 값을 결정해야 한다. θ의 추정법에는 여러 가지가 제안되었으며 다음의 몇 가지 방법을 소개한다.

(1) 능형 트레이스

$\hat{\beta}_j(\theta),\ j = 0,\ \cdots,\ p-1$를 여러 가지 θ 값에 대해 그림을 그린 다음 $\hat{\beta}_j(\theta)$ 값들이 수렴하기 시작하는 값을 찾는 것이다. 이러한 $\hat{\beta}_j(\theta)$와 θ의 그림을 능형 트레이스 (ridge trace)라 부른다. 능형 트레이스에 의한 θ의 추정은 매우 주관적임을 알 수 있다.

(2) GCV_θ

GCV_θ는 Wahba, Golub and Heath (1979)가 제안한 일반화 교차확인법 (GCV; generalized cross-validation)으로

$$GCV_\theta = \sum e_{i,\theta}^2 \ / \ \left[1 - \frac{1}{n}\text{tr}(\mathbf{H}_\theta)\right]^2$$

로 정의된다. 여기서, $e_{i,\theta}$는 θ가 주어졌을 때의 잔차이고 $\mathbf{H}_\theta = \mathbf{X}(\mathbf{X}'\mathbf{X} + \theta\mathbf{I})^{-1}\mathbf{X}'$이다. GCV_θ는 $PRESS_\theta$와 밀접한 관계가 있으며, GCV_θ를 최소화 하는 θ를 추정치로 사용한다.

5.4.2 LASSO

LASSO (least absolute shrinkage and selection operator)는 Tibshirani (1996)에 의해 제안된 편의 추정의 한 방법이며 다음과 같은 동기에서 제안되었다. 첫째, 변수선택은 해석하기 좋은 모형을 제공하지만 이산적 과정 (특정 변수가 최종 선택된 모형에 포함되거나 제거되는 과정)으로 인해 변동성이 매우 심하다. 둘째, 능형회귀는 연속적 과정 (특정 변수를 포함시키거나 제거시키는 대신 추정치의 값을 축소시키는 과정)으로 보다 안정적이지만 변수선택의 기능이 없어 해석하기 좋은 모형을 제공하지 못한다. 따라서, 변수 선택과 능형회귀의 장점인 모형해석의 용이함과 연속적 과정을 동시에 가진 방법이 필요한데 LASSO가 바로 그것이라고 주장한다. 즉, LASSO는 변수선택의 기능을 연곳적 과정으로 수행한다는 것이다. LASSO 추정법은 능형회귀와 매우 흡사한데, 능형회귀의 β에 대한 제약조건이 제곱형태 (L_2 restriction)라면, LASSO는 제약조건이 절대값 형태 (L_1 restriction)이다. 즉, LASSO는

$$\begin{aligned}\hat{\beta}_L &= \underset{\beta}{arg\,min}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\ s.t.\ \sum_{j=1}^{p}|\beta_j| \le s \\ &= \underset{\beta}{arg\,min}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+\lambda\sum_{j=1}^{p}|\beta_j|\end{aligned}$$

으로 주어진다. 여기서, s와 λ는 일대일로 대응하는 미지의 모수인데 능형회귀의 편의모수처럼 추정되어야 할 값이다.

능형회귀는 편의모수가 주어지면 추정치가 식으로 주어지지만 LASSO 추정치는 s가 주어지더라도 식으로 표현할 수 없는 단점이 있다. LASSO를 추정하기 위한 여러 알고리즘이 개발되어 있으나 그 중에서도 LARS (least angle regression)라는 알고리즘이 가장 뛰어난 것으로 알려져 있으며 이는 R 패키지에서 사용 가능하다.

5.4.3 Elastic Net

Elastic net은 Zou and Hastie (2005)에 의해 제안되었으며, 이는 능형회귀와 LASSO의 가중평균을 구한 형태를 가진다. 즉, elastic net 추정치는

$$arg\min_{\beta}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2$$

으로 주어지는데 흔히 $\lambda_1 = 1-\lambda_2$을 가정하여 한개의 평활모수로 축소하거나, 더 나아가서 $\lambda_1 = \lambda_2 = 1/2$을 가정하기도 한다.

예제 : 전립선암 환자자료 (Prostate)에서 lpsa (전립선 항원 수준의 로그값)를 반응변수, age (나이)를 설명변수라고 하자. 단순회귀모형에 적합해 보자.

Prostate data

```
library(lasso2)
data(Prostate)
str(Prostate)
#> 'data.frame':    97 obs. of  9 variables:
#>  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
#>  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
#>  $ age    : num  50 58 74 58 62 50 64 58 47 63 ...
#>  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
#>  $ svi    : num  0 0 0 0 0 0 0 0 0 0 ...
#>  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
#>  $ gleason: num  6 6 7 6 6 6 6 6 6 6 ...
#>  $ pgg45  : num  0 0 20 0 0 0 0 0 0 0 ...
#>  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
```

correlation

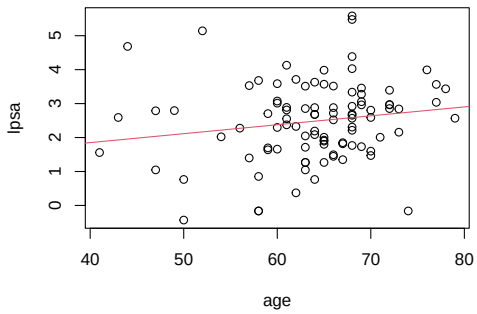
```
cor(Prostate[, c('lpsa','age')])
#>            lpsa       age
#> lpsa 1.0000000 0.1695928
#> age  0.1695928 1.0000000
```

simple linear regression

```
fit <- lm(lpsa ~ age, data=Prostate)
summary(fit)
#>
#> Call:
#> lm(formula = lpsa ~ age, data = Prostate)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.90738 -0.71234  0.06967  0.66187  2.99584
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.79906    1.00793   0.793   0.4299
#> age          0.02629    0.01568   1.677   0.0968 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.144 on 95 degrees of freedom
#> Multiple R-squared:  0.02876,    Adjusted R-squared:  0.01854
#> F-statistic: 2.813 on 1 and 95 DF,  p-value: 0.09677
```

scatter plot and fitting line

```
plot(lpsa ~ age, data=Prostate)
abline(fit, col=2)
```



coefficients and 95% CI

```
#summary(fit)$coef
cbind(coef(fit), confint.default(fit))
#>                                2.5 %     97.5 %
#> (Intercept) 0.79906015 -1.176443790 2.77456409
#> age         0.02629454 -0.004431546 0.05702063
```

anova table

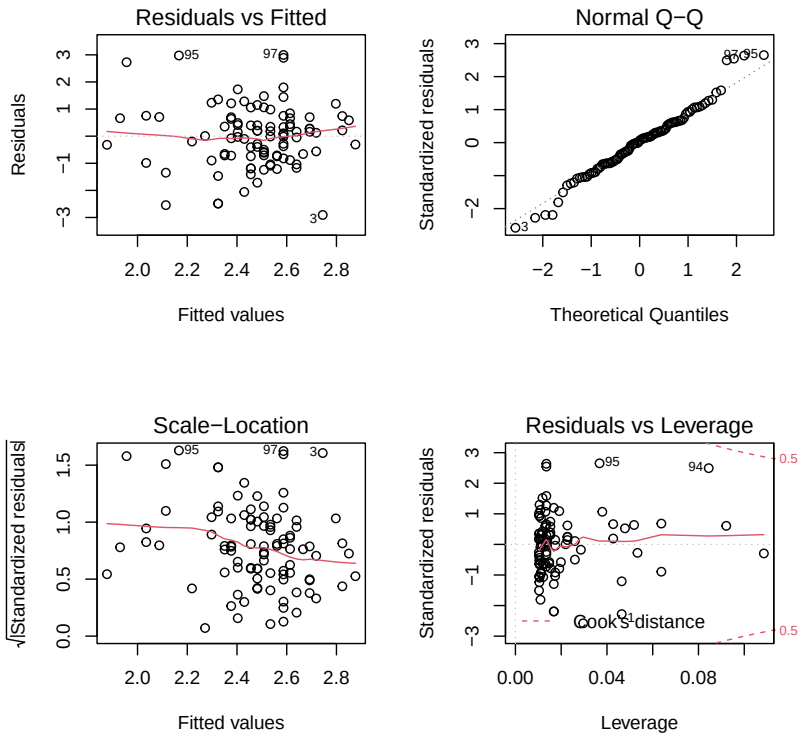
```
anova(fit)
#> Analysis of Variance Table
#>
#> Response: lpsa
#>           Df  Sum Sq Mean Sq F value  Pr(>F)
#> age        1   3.679  3.6791  2.8133 0.09677 .
#> Residuals 95 124.239  1.3078
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

goodness-of-fit

```
#names(summary(fit))
cbind(RMSE = summary(fit)$sigma,
      R2 = summary(fit)$r.squared,
      adj_R2 = summary(fit)$adj.r.squared,
      AIC = AIC(fit),
      BIC = BIC(fit))
#>          RMSE         R2      adj_R2      AIC      BIC
#> [1,] 1.143579 0.02876173 0.01853817 305.2808 313.005
```

model diagnostics

```
par(mfrow=c(2,2))
plot(fit); #dev.off()
```

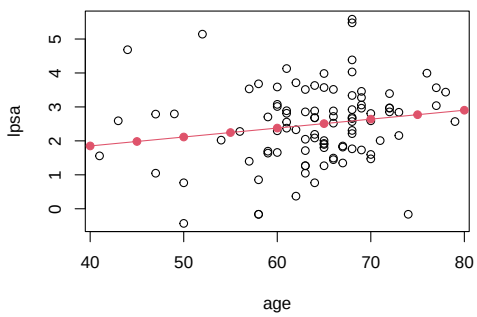


predicted value

```
pred <- predict(fit)
pred <- predict(fit, interval="confidence")
head(pred,4)
#>        fit      lwr      upr
#> 1 2.113787 1.624536 2.603038
#> 2 2.324144 2.030093 2.618195
#> 3 2.744856 2.354202 3.135511
#> 4 2.324144 2.030093 2.618195
```

predicted value of new data

```
x <- seq(40,80,by=5)
y <- predict(fit, data.frame(age = x))
plot(lpsa ~ age, data=Prostate)
lines(x, y, type="o", col=2, pch=20, cex=1.5)
```



예제 : 전립선암 환자자료 (Prostate)에서 lpsa (전립선 항원 수준의 로그값)를 반응변수, 나머지를 설명변수라고 하자. 중선형회귀모형을 적합하고 변수선택, Lasso 회귀모형 등과 비교해 보자.

pairs plot and correlation

```
pairs(Prostate); cor(Prostate)
```

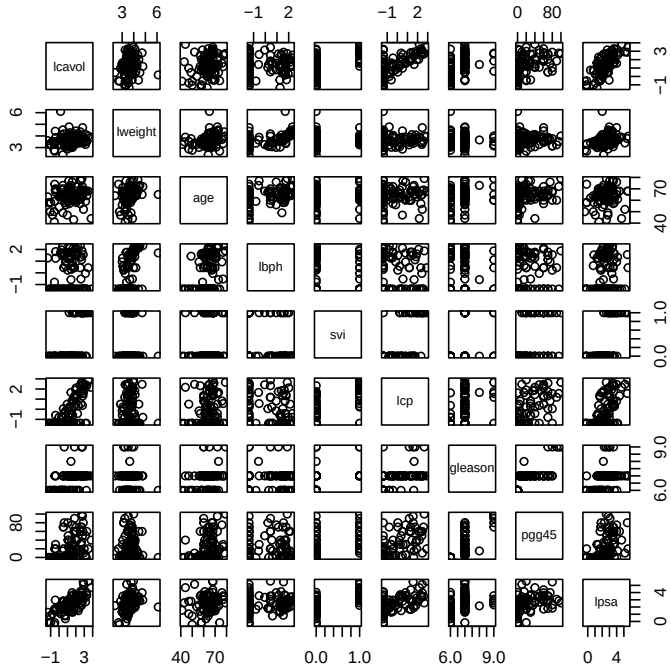


그림 5.2: 전립선암 환자자료에 대한 산점도행렬

표 5.1: 전립선암 환자자료에 대한 상관계수

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
lcavol	1.00	0.19	0.22	0.03	0.54	0.68	0.43	0.43	0.73
lweight	0.19	1.00	0.31	0.43	0.11	0.10	0.00	0.05	0.35
age	0.22	0.31	1.00	0.35	0.12	0.13	0.27	0.28	0.17
lbph	0.03	0.43	0.35	1.00	-0.09	-0.01	0.08	0.08	0.18
svi	0.54	0.11	0.12	-0.09	1.00	0.67	0.32	0.46	0.57
lcp	0.68	0.10	0.13	-0.01	0.67	1.00	0.51	0.63	0.55
gleason	0.43	0.00	0.27	0.08	0.32	0.51	1.00	0.75	0.37
pgg45	0.43	0.05	0.28	0.08	0.46	0.63	0.75	1.00	0.42
lpsa	0.73	0.35	0.17	0.18	0.57	0.55	0.37	0.42	1.00

multiple linear regression

```
fit <- lm(lpsa ~ ., data=Prostate)
summary(fit)
#>
#> Call:
#> lm(formula = lpsa ~ ., data = Prostate)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -1.73316 -0.37133 -0.01702  0.41414  1.63811
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.669399   1.296381   0.516  0.60690
#> lcavol       0.587023   0.087920   6.677 2.11e-09 ***
#> lweight      0.454461   0.170012   2.673  0.00896 **
#> age         -0.019637   0.011173  -1.758  0.08229 .
#> lbph         0.107054   0.058449   1.832  0.07040 .
#> svi          0.766156   0.244309   3.136  0.00233 **
#> lcp         -0.105474   0.091013  -1.159  0.24964
#> gleason      0.045136   0.157464   0.287  0.77506
#> pgg45        0.004525   0.004421   1.024  0.30885
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.7084 on 88 degrees of freedom
#> Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
#> F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

coefficients and 95% CI

```
cbind(coef(fit), confint.default(fit))
#>                                 2.5 %      97.5 %
#> (Intercept)  0.669399027 -1.871461587 3.210259641
#> lcavol       0.587022881  0.414702115 0.759343647
#> lweight      0.454460641  0.121243105 0.787678177
#> age         -0.019637208 -0.041535382 0.002260966
#> lbph         0.107054351 -0.007504234 0.221612936
#> svi          0.766155885  0.287318079 1.244993690
#> lcp         -0.105473570 -0.283856721 0.072909582
#> gleason      0.045135964 -0.263488720 0.353760648
#> pgg45        0.004525324 -0.004140039 0.013190686
```

anova table

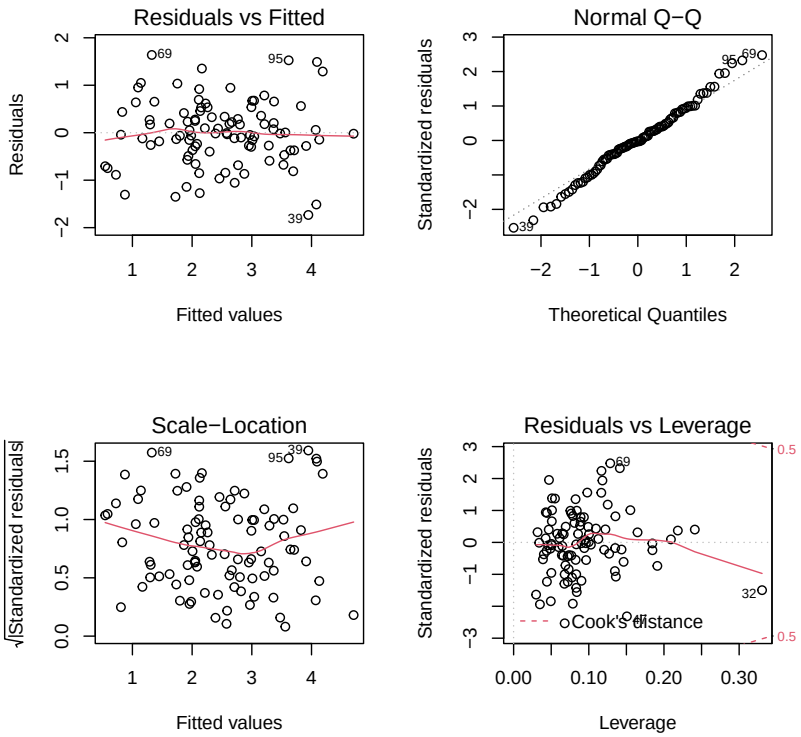
```
anova(fit)
#> Analysis of Variance Table
#>
#> Response: lpsa
#>           Df Sum Sq Mean Sq  F value    Pr(>F)
#> lcavol     1 69.003  69.003 137.4960 < 2.2e-16 ***
#> lweight    1  5.948   5.948  11.8529 0.0008833 ***
#> age        1  0.420   0.420   0.8369 0.3627907
#> lbph       1  1.069   1.069   2.1302 0.1479844
#> svi        1  5.952   5.952  11.8594 0.0008806 ***
#> lcp        1  0.129   0.129   0.2576 0.6130570
#> gleason    1  0.707   0.707   1.4097 0.2382955
#> pgg45      1  0.526   0.526   1.0477 0.3088513
#> Residuals 88 44.163   0.502
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

goodness-of-fit

```
#names(summary(fit))
cbind(RMSE = summary(fit)$sigma,
      R2 = summary(fit)$r.squared,
      adj_R2 = summary(fit)$adj.r.squared,
      AIC = AIC(fit),
      BIC = BIC(fit))
#>           RMSE        R2     adj_R2      AIC      BIC
#> [1,] 0.7084164 0.6547535 0.6233674 218.9525 244.6996
```

model diagnostics

```
par(mfrow=c(2,2))
plot(fit); #dev.off()
```



predicted value

```
pred <- predict(fit)
pred <- predict(fit, interval="confidence")
head(pred)
#>         fit        lwr       upr
#> 1 0.8744063 0.47937891 1.2694337
#> 2 0.7240557 0.35807267 1.0900388
#> 3 0.5437102 0.02460792 1.0628124
#> 4 0.5842042 0.19187151 0.9765369
#> 5 1.7214934 1.45811454 1.9848722
#> 6 0.8072678 0.40089705 1.2136385
```

predicted value of new data

```
## new data
#summary(Prostate)
x <- data.frame(
  lcavol = 1.35,
  lweight = 3.6,
  age = 63,
  lbph = 0.1,
  svi = 0.2,
  lcp = -0.17,
  gleason = 6.7,
  pgg45 = 24.38
)
predict(fit, x)
#>      1
#> 2.4554
```

변수선택 (subset selection method)

전진선택법, 후진제거법, 단계선택법에 대해 소개한다.

```
## variable selection method
m_null <- lm(lpsa ~ 1, data=Prostate)
m_full <- lm(lpsa ~ ., data=Prostate)

### forward variable selection
fit <- step(m_null,
            scope=list(lower=formula(m_null), upper=formula(m_full)),
            direction="forward", trace=F)

### backward variable elimination
fit <- step(m_full, direction="backward", trace=F)

### stepwise variable selection
fit <- step(m_null,
            scope=list(lower=formula(m_null), upper=formula(m_full)),
            direction="both", trace=F)
```

Stepwise variable selection

```
#> Start:  AIC=28.84
#> lpsa ~ 1
#>
#>           Df Sum of Sq     RSS     AIC
#> + lcavol   1    69.003  58.915 -44.366
#> + svi      1    41.011  86.907  -6.658
#> + lcp      1    38.528  89.389  -3.926
#> + pgg45    1    22.814 105.103  11.783
#> + gleason  1    17.416 110.502  16.641
#> + lweight  1    16.041 111.877  17.841
```

```
#> + lbph     1     4.136 123.782  27.650
#> + age      1     3.679 124.239  28.007
#> <none>                 127.918  28.838
#>
#> Step:  AIC=-44.37
#> lpsa ~ lcavol
#>
#>           Df Sum of Sq     RSS     AIC
#> + lweight  1     5.948  52.966 -52.690
#> + svi      1     5.237  53.677 -51.397
#> + lbph     1     3.266  55.649 -47.898
#> + pgg45    1     1.698  57.217 -45.203
#> <none>                  58.915 -44.366
#> + lcp      1     0.656  58.259 -43.452
#> + gleason  1     0.416  58.499 -43.053
#> + age      1     0.003  58.912 -42.370
#> - lcavol   1    69.003 127.918  28.838
#>
#> Step:  AIC=-52.69
#> lpsa ~ lcavol + lweight
#>
#>           Df Sum of Sq     RSS     AIC
#> + svi      1     5.181  47.785 -60.676
#> + pgg45    1     1.949  51.017 -54.327
#> <none>                  52.966 -52.690
#> + lcp      1     0.837  52.129 -52.235
#> + gleason  1     0.781  52.185 -52.131
#> + lbph     1     0.675  52.291 -51.935
#> + age      1     0.420  52.546 -51.462
#> - lweight  1     5.948  58.915 -44.366
#> - lcavol   1    58.910 111.877  17.841
#>
```

```
#> Step:  AIC=-60.68
#> lpsa ~ lcavol + lweight + svi
#>
#>           Df Sum of Sq    RSS     AIC
#> + lbph     1    1.3001 46.485 -61.352
#> <none>                 47.785 -60.676
#> + pgg45    1    0.5735 47.211 -59.847
#> + age      1    0.4025 47.382 -59.497
#> + gleason  1    0.3890 47.396 -59.469
#> + lcp      1    0.0641 47.721 -58.806
#> - svi      1    5.1814 52.966 -52.690
#> - lweight  1    5.8923 53.677 -51.397
#> - lcavol   1   28.0446 75.830 -17.884
#>
#> Step:  AIC=-61.35
#> lpsa ~ lcavol + lweight + svi + lbph
#>
#>           Df Sum of Sq    RSS     AIC
#> + age      1    0.9593 45.526 -61.374
#> <none>                 46.485 -61.352
#> - lbph     1    1.3001 47.785 -60.676
#> + pgg45    1    0.3533 46.132 -60.092
#> + gleason  1    0.2125 46.272 -59.796
#> + lcp      1    0.1023 46.383 -59.565
#> - lweight  1    2.8013 49.286 -57.675
#> - svi      1    5.8063 52.291 -51.935
#> - lcavol   1   27.8299 74.315 -17.841
#>
#> Step:  AIC=-61.37
#> lpsa ~ lcavol + lweight + svi + lbph + age
#>
#>           Df Sum of Sq    RSS     AIC
```

```
#> <none>                  45.526 -61.374
#> - age       1    0.9593 46.485 -61.352
#> + pgg45     1    0.6590 44.867 -60.788
#> + gleason   1    0.4560 45.070 -60.351
#> + lcp       1    0.1293 45.396 -59.650
#> - lbph      1    1.8568 47.382 -59.497
#> - lweight   1    3.2250 48.751 -56.735
#> - svi       1    5.9517 51.477 -51.456
#> - lcavol    1   28.7666 74.292 -15.870
```

Best subset selection

```
library(leaps)
fit <- regsubsets(lpsa ~ ., data=Prostate)
#names(summary(fit))
R2 = summary(fit)$rsq
adj_R2 = summary(fit)$adjr2
BIC = summary(fit)$bic
Cp = summary(fit)$cp
```

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	R2	adj_R2	BIC	Cp
1 (1)	*								0.54	0.53	-66.05	24.39
2 (1)	*	*							0.59	0.58	-71.80	14.54
3 (1)	*	*			*				0.63	0.61	-77.21	6.22
4 (1)	*	*		*	*				0.64	0.62	-75.32	5.63
5 (1)	*	*	*	*	*				0.64	0.62	-72.76	5.71
6 (1)	*	*	*	*	*			*	0.65	0.63	-69.60	6.40
7 (1)	*	*	*	*	*	*		*	0.65	0.63	-66.47	7.08
8 (1)	*	*	*	*	*	*	*	*	0.65	0.62	-61.99	9.00

축소추정법 (shrinkage method)

전립선암 자료를 능형회귀, LASSO, Elastic net에 적합한다.

```
library(glmnet)
x = model.matrix(lpsa ~., data=Prostate)[,-1]
y = Prostate$lpsa

## lambda: tuning parameter
grid=10^seq(-3, 3, length=100)

set.seed(1234)
## alpha: Ridge=0, Lasso=1, Elasticnet=0.5
cv <- cv.glmnet(x, y, lambda=grid, alpha=0, standardize=F,
                type.measure="mse")
bestlam = cv$lambda.min

## coefficients
fit <- glmnet(x, y, lambda=grid, alpha=0, standardize=F)
beta <- coef(fit, s=bestlam)

## predicted value
pred <- predict(fit, x, type="response", s=bestlam)[,1]

par(mfrow=c(1,2))
plot(fit, xvar="lambda", col=1:8)
abline(v=log(bestlam), col=2, lty=2)
legend("topright", colnames(x),
       col=1:8, text.col=1:8, lty=1, bty="n")

plot(cv)
title(sub=paste("Best Log Lambda = ", round(log(bestlam), 2)))
```

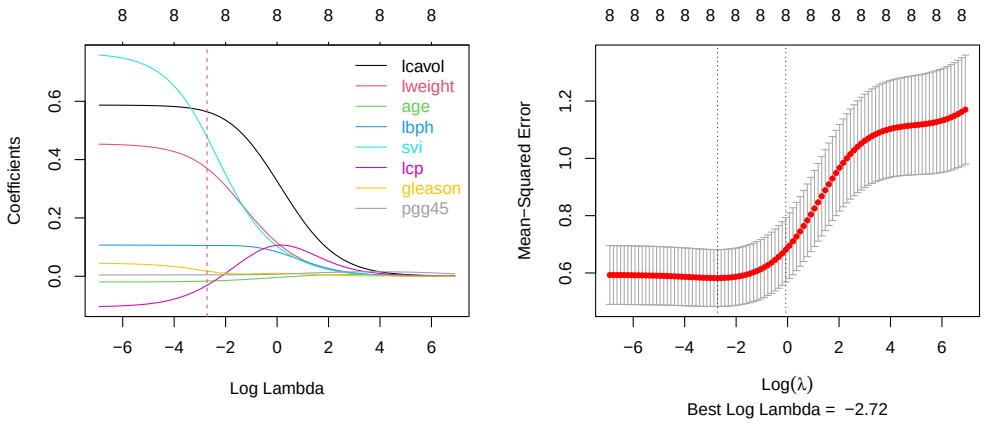


그림 5.3: Ridge regression

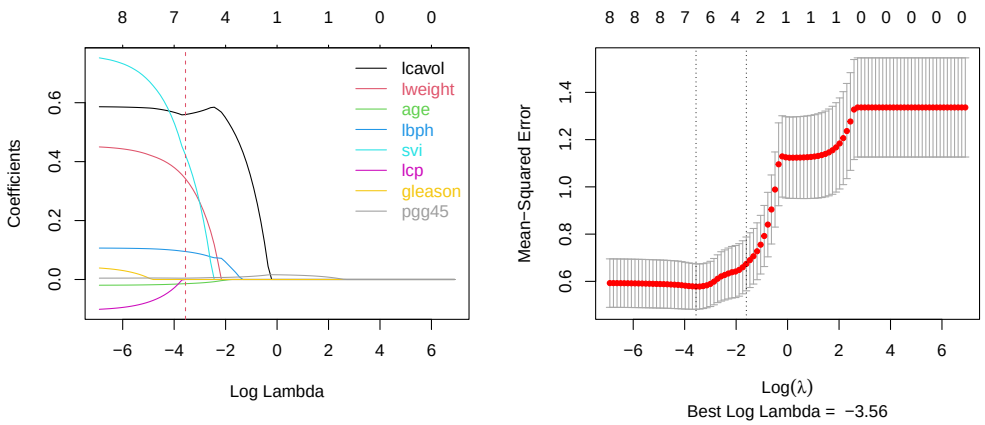


그림 5.4: LASSO regression

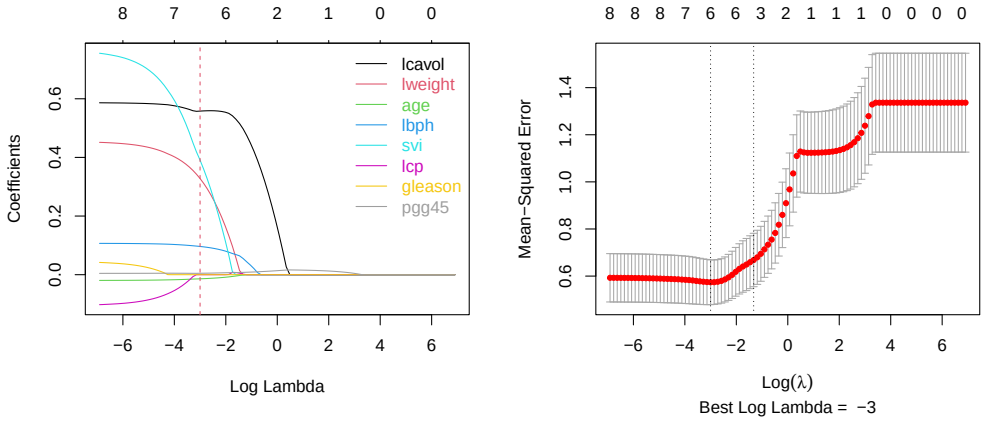


그림 5.5: Elasticnet regression

표 5.2: Coefficients of linear regression

Variables	LSE	Stepwise	Ridge	LASSO	Elasticnet
(Intercept)	0.669	0.951	1.065	1.167	1.212
lcavol	0.587	0.566	0.564	0.560	0.558
lweight	0.454	0.424	0.370	0.342	0.328
age	-0.020	-0.015	-0.017	-0.014	-0.014
lbph	0.107	0.112	0.106	0.095	0.096
svi	0.766	0.721	0.477	0.422	0.391
lcp	-0.105		-0.031	0.000	0.000
gleason	0.045		0.018	0.000	0.000
pgg45	0.005		0.005	0.005	0.005

표 5.3: MSE of linear regression

	LSE	Stepwise	Ridge	LASSO	Elasticnet
	44.163	45.526	45.191	45.768	46.042
df	8	5	8	6	6
MSE	0.496	0.495	0.508	0.503	0.506

5.5 로지스틱 회귀모형

5.5.1 모형의 적합

반응변수가 연속형 변수일 때 중선형회귀분석을 이용하지만 이항분포인 경우에는 로지스틱 회귀모형이 사용된다. 예를 들어, 피마 인디언 자료에서 k개의 변수로 구성된 공변량 $\mathbb{X} = (X_1, \cdots, X_k)$가 주어졌을 때 반응변수 Y가 당뇨병 여부 (예를 들어, 당뇨병에 걸리면 $Y = 1$, 걸리지 않았으면 $Y = 0$)를 나타낸다고 하자. 이 때 공변량 $\mathbb{X}$를 가진 사람이 당뇨병에 걸릴 학률을 $p_{\mathbf{x}} = P(Y = 1|\mathbb{X})$라 하면 로지스틱 회귀는

$$log(\frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

으로 주어진다. 한편,

$$\frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}}$$

는 성공과 실패의 비로서 이를 오즈비 (odds ratio)라 하며, 로그 오즈비를 로짓 함수 (logit function)라 부른다. 즉, $\text{logit}(p) = log(p/(1-p))$이다. 이제 적합된 회귀계수를 $\hat{\beta}_j,\ j = 1, \cdots, k$라 하면 $\hat{\beta}_j$는 j번째 공변량 X_j가 아닌 다른 공변량들이 고정되어 있다는 가정하에 X_j가 한 단위 증가할 때 오즈비는 $exp(\hat{\beta}_j)$만큼 증가한다고 해석할 수 있다. 한편, 로지스틱 회귀를 다시 표현하면

$$p_{\mathbf{x}} = \frac{exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\}}{1 + exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\}}$$

로 쓸 수 있다.

5.5.2 ROC 곡선

의학연구의 가장 중요한 것 중의 하나는 올바른 진단 (diagnosis)이다. 논의의 단순화를 위해 진단은 두 가지 (병의 유무)로 이루어진다고 가정한다. 흔히 병이 있는 경우를 positive, 없는 경우를 negative로 나타낸다. 최적의 진단이란 실제로 병이 있는 경우에 있다고 판단 (TP - true positive, 진양)하고, 없는 경우에 없다고 판단 (TN - true negative, 진음)할 확률을 최대화하는 것이다. 즉, 실제로 병이 없는 경우에 있다고 판단 (FP - false positive, 위양)하고, 있는 경우에 없다고 판단 (FN - false negative, 위음)할 확률을 최소화하는 것이다. 이러한 상황은 가설검정과 동일하다. FP는 제1종 오류 (귀무가설이 맞음에도 불구하고 귀무가설을 기각시키는 오류)와 동일하고, FN은 제2종 오류 (대립가설이 맞음에도 불구하고 대립가설을 기각시키지 않는 오류)에 해당된다.

	실제 상황 (true state)	
분석가의 결정	positive	negative
positive	TP (true positive)	FP (false positive)
negative	FN (false negative)	TN (true negative)

이를 바탕으로 다음과 같은 여러 가지 개념에 대한 정의를 소개한다.

- 민감도 (sensitivity, recall) = TPR (true positive rate, 진양율) = TP / (TP + FN)
- 특이도 (specificity, selectivity) = TNR (true negative rate, 진음율) = TN / (TN + FP)
- 유병율 (prevalence) = (TP + FN) / total
- 정확도 (accuracy) = (TP + TN) / total
- 양성예측도 = PPV (positive predicted value) = TP / (TP + FP)
- 음성예측도 = NPV (negative predicted value) = TN / (TN + FN)
- 위발견율 = FDR (false discovery rate) = FP / (TP + FP) = 1- PPV

이러한 개념을 바탕으로 최적진단을 선택하는 방법으로 ROC 곡선 (Receiver Operating Characteristics curve)을 소개한다. 이 곡선의 명칭이 ROC인 유래는 1941년 제2차 세계대전 당시 레이더 수신기 (radar receiver)를 작용하여 (operator) 아군과 적군을 구분하는 (binary classification) 용도로 사용된 것이라 한다. ROC 곡선의 Y축은 민감도 (TPR), X축은

FPR (1-특이도)를 나타낸다. 어떤 진단을 하느냐에 따라 2차원 평면 위에 하나의 점이 주어진다. 로지스틱 회귀를 한 다음 추정된 확률 $\hat{p}_x$가 $k\ (0 < k < 1)$보다 크면 병이 있다 (즉, 반응변수 = 1)라고 진단할 경우 k값이 주어지면 TPR과 FPR이 정해지고 2차원 평면 위에 하나의 점이 정해진다. 이제 k를 0부터 1까지 변화시키면 여러 개의 점이 정해지고 이러한 점들을 연결하면 한 개의 곡선이 생기는데 이것이 바로 ROC 곡선이다. 이제 어떤 k값을 정해야 최적의 진단인가? Y축 값 (TPR)은 클수록 X축 값 (FPR)은 작을수록 좋다. 마치 가설검정에서 제1종 오류는 작을수록, 검정력은 클수록 좋은 것과 같은 논리다. 주어진 진단법에서 k를 선택하기 위해 여러 가지 방법이 있으나 가장 널리 사용되는 것은 기울기 1인 직선으로 ROC 곡선과 접하는 점에 해당되는 k를 선택하는 방법이 가장 널리 사용된다.

AUC (area under curve)는 ROC 곡선 아래 부분의 면적이다. 따라서, AUC가 1에 가까울수록 좋은 진단법이다. 한편, 임의로 진단하면 (wild guess 또는 coin flip) AUC는 (0,0)과 (1,1)을 연결하는 직선의 형태로서 AUC는 0.5에 가깝다.

예제 : 피마 인디언 자료 (Pima.te)에서 type (당뇨병 유무)를 반응변수로, 나머지를 설명변수로 하여 로지스틱 회귀모형에 적합해 보자.

Pima.te data

```
library(MASS)
str(Pima.te)
#> 'data.frame':    332 obs. of  8 variables:
#>  $ npreg: int  6 1 1 3 2 5 0 1 3 9 ...
#>  $ glu  : int  148 85 89 78 197 166 118 103 126 119 ...
#>  $ bp   : int  72 66 66 50 70 72 84 30 88 80 ...
#>  $ skin : int  35 29 23 32 45 19 47 38 41 35 ...
#>  $ bmi  : num  33.6 26.6 28.1 31 30.5 25.8 45.8 43.3 39.3 29 ...
#>  $ ped  : num  0.627 0.351 0.167 0.248 0.158 0.587 0.551 0.183 0.704 0.263 ...
#>  $ age  : int  50 31 21 26 53 51 31 33 27 29 ...
#>  $ type : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 1 1 2 ...

## pairs plot
pairs(Pima.te)
```

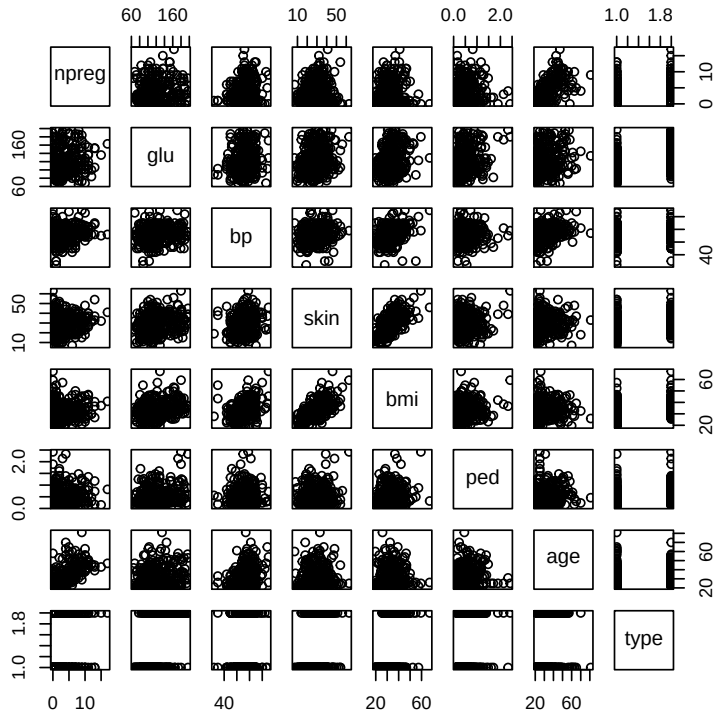


그림 5.6: 피마 인디언 자료에 대한 산점도 행렬

multiple logistic regression

```
fit <- glm(type=="Yes" ~ ., data=Pima.te, family=binomial)
summary(fit)
#>
#> Call:
#> glm(formula = type == "Yes" ~ ., family = binomial, data = Pima.te)
#>
#> Deviance Residuals:
#>     Min       1Q   Median       3Q      Max
#> -2.9647  -0.6582  -0.3608   0.6158   2.4646
#>
#> Coefficients:
#>              Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -9.514019   1.229278  -7.740 9.98e-15 ***
#> npreg        0.140944   0.059652   2.363  0.01814 *
#> glu          0.037481   0.005558   6.743 1.55e-11 ***
#> bp          -0.008675   0.012589  -0.689  0.49076
#> skin         0.013167   0.020025   0.658  0.51084
#> bmi          0.078951   0.028432   2.777  0.00549 **
#> ped          1.110131   0.446921   2.484  0.01299 *
#> age          0.018055   0.018359   0.983  0.32537
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 420.30  on 331  degrees of freedom
#> Residual deviance: 285.79  on 324  degrees of freedom
#> AIC: 301.79
#>
#> Number of Fisher Scoring iterations: 5
```

OR and 95% CI

```
cbind(exp(coef(fit)), exp(confint.default(fit)))
#>                                  2.5 %       97.5 %
#> (Intercept) 7.380982e-05 6.633623e-06 0.0008212541
#> npreg       1.151360e+00 1.024320e+00 1.2941558991
#> glu         1.038192e+00 1.026943e+00 1.0495640454
#> bp          9.913625e-01 9.672010e-01 1.0161276318
#> skin        1.013254e+00 9.742552e-01 1.0538144046
#> bmi         1.082151e+00 1.023497e+00 1.1441673147
#> ped         3.034757e+00 1.263874e+00 7.2869197427
#> age         1.018219e+00 9.822328e-01 1.0555242956
```

AUC and 95% CI

```
library(pROC)
y = as.numeric(Pima.te$type=="Yes")
pred <- predict(fit, type="response")
(roc.fit <- roc(y, pred, ci=T))
#>
#> Call:
#> roc.default(response = y, predictor = pred, ci = T)
#>
#> Data: pred in 223 controls (y 0) < 109 cases (y 1).
#> Area under the curve: 0.8686
#> 95% CI: 0.8294-0.9078 (DeLong)
```

cut-off value

```
rets = c("threshold", "tp", "fn", "fp", "tn",
         "sensitivity", "specificity", "accuracy", "ppv", "npv")
out <- coords(roc.fit, x=seq(0,1,by=0.2), ret=rets, transpose=F)
```

표 5.5: Sensitivity and Specificity

threshold	tp	fn	fp	tn	sensitivity	specificity	accuracy	ppv	npv
0.0	109	0	223	0	1.000	0.000	0.328	0.328	
0.2	95	14	81	142	0.872	0.637	0.714	0.540	0.910
0.4	77	32	33	190	0.706	0.852	0.804	0.700	0.856
0.6	57	52	14	209	0.523	0.937	0.801	0.803	0.801
0.8	32	77	4	219	0.294	0.982	0.756	0.889	0.740
1.0	0	109	0	223	0.000	1.000	0.672		0.672

optimal cut-off value

```
coords(roc.fit, x="best", best.method="youden", transpose=F)
#>   threshold specificity sensitivity
#> 1 0.2709196   0.7668161   0.8440367
```

ROC curv

```
par(mfrow=c(1,2))
plot(roc.fit, print.auc=TRUE, grid=c(0.1, 0.1))
plot(roc.fit, print.thres="best", grid=c(0.1, 0.1))
```

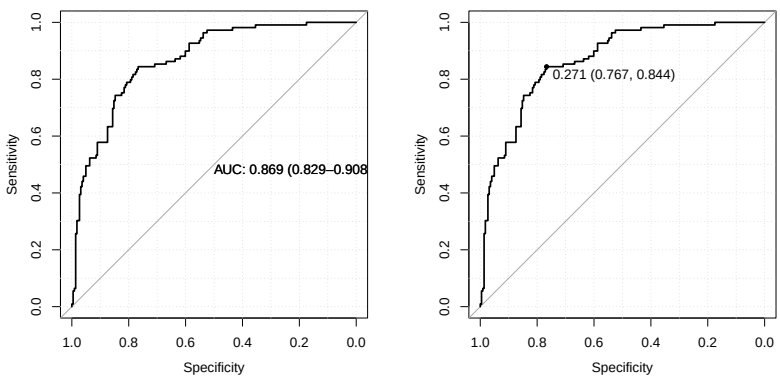


그림 5.7: ROC 곡선

variable selection method

```
m_null <- glm(type=="Yes" ~ 1, data=Pima.te, family=binomial)
m_full <- glm(type=="Yes" ~ ., data=Pima.te, family=binomial)

### forward variable selection
fit <- step(m_null,
            scope=list(lower=formula(m_null), upper=formula(m_full)),
            direction="forward", trace=F)

### backward variable elimination
fit <- step(m_full, direction="backward", trace=F)

### stepwise variable selection
fit <- step(m_null,
            scope=list(lower=formula(m_null), upper=formula(m_full)),
            direction="both", trace=F)
```

Stepwise variable selection

```
#> Start:  AIC=422.3
#> type == "Yes" ~ 1
#>
#>         Df Deviance    AIC
#> + glu    1   325.99 329.99
#> + bmi    1   386.84 390.84
#> + age    1   394.44 398.44
#> + skin   1   395.98 399.98
#> + ped    1   399.79 403.79
#> + npreg  1   401.55 405.55
#> + bp     1   410.44 414.44
#> <none>       420.30 422.30
#>
```

```
#> Step:  AIC=329.99
#> type == "Yes" ~ glu
#>
#>         Df Deviance    AIC
#> + npreg  1   311.02 317.02
#> + bmi    1   312.49 318.49
#> + age    1   314.87 320.87
#> + skin   1   315.16 321.16
#> + ped    1   317.52 323.52
#> + bp     1   323.88 329.88
#> <none>       325.99 329.99
#> - glu    1   420.30 422.30
#>
#> Step:  AIC=317.02
#> type == "Yes" ~ glu + npreg
#>
#>         Df Deviance    AIC
#> + bmi    1   294.54 302.54
#> + skin   1   301.07 309.07
#> + ped    1   303.72 311.72
#> <none>       311.02 317.02
#> + age    1   310.08 318.08
#> + bp     1   310.16 318.16
#> - npreg  1   325.99 329.99
#> - glu    1   401.55 405.55
#>
#> Step:  AIC=302.54
#> type == "Yes" ~ glu + npreg + bmi
#>
#>         Df Deviance    AIC
#> + ped    1   287.44 297.44
#> <none>       294.54 302.54
```

```
#> + age    1   293.35 303.35
#> + skin   1   293.93 303.93
#> + bp     1   294.28 304.28
#> - bmi    1   311.02 317.02
#> - npreg  1   312.49 318.49
#> - glu    1   364.90 370.90
#>
#> Step:  AIC=297.44
#> type == "Yes" ~ glu + npreg + bmi + ped
#>
#>         Df Deviance    AIC
#> <none>       287.44 297.44
#> + age    1   286.73 298.73
#> + skin   1   286.96 298.96
#> + bp     1   287.23 299.23
#> - ped    1   294.54 302.54
#> - bmi    1   303.72 311.72
#> - npreg  1   304.01 312.01
#> - glu    1   349.80 357.80
```

Shrinkage method

```
library(glmnet)
x = model.matrix(type=="Yes" ~., data=Pima.te)[,-1]
y = as.numeric(Pima.te$type=="Yes")

## lambda: tuning parameter
grid=10^seq(-4, 3, length=100)

set.seed(1234)
## alpha: Ridge=0, Lasso=1, Elasticnet=0.5
cv <- cv.glmnet(x, y, lambda=grid, alpha=0, standardize=F,
```

```
                family="binomial", type.measure = "auc")
fit <- glmnet(x, y, lambda=grid, alpha=0, standardize=F, family="binomial")

bestlam = cv$lambda.min

## coefficients
beta <- coef(fit, s=bestlam)

## predicted value
pred <- predict(fit, x, type="response", s=bestlam)[,1]

par(mfrow=c(1,2))
plot(fit, xvar="lambda", col=1:8)
abline(v=log(bestlam), col=2, lty=2)
legend("topright", colnames(x),
       col=1:8, text.col=1:8, lty=1, bty="n")

plot(cv)
title(sub=paste("Best Log Lambda = ", round(log(bestlam), 2)))
```

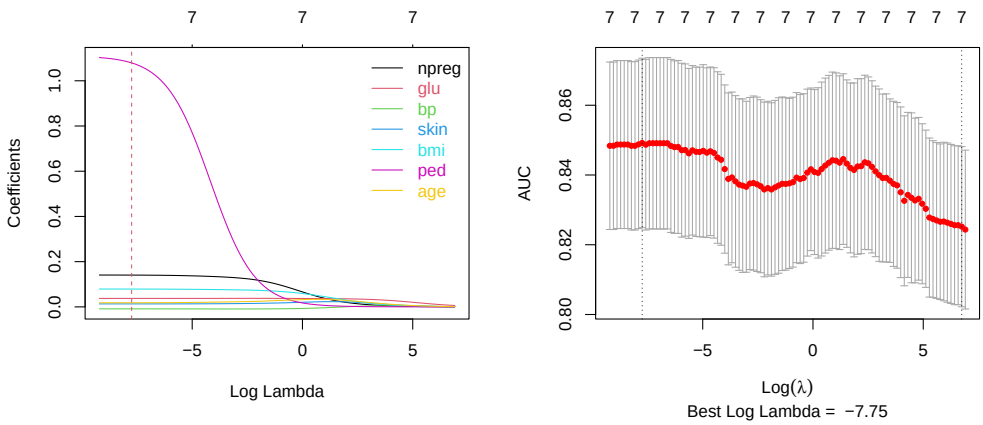


그림 5.8: Ridge logistic regression

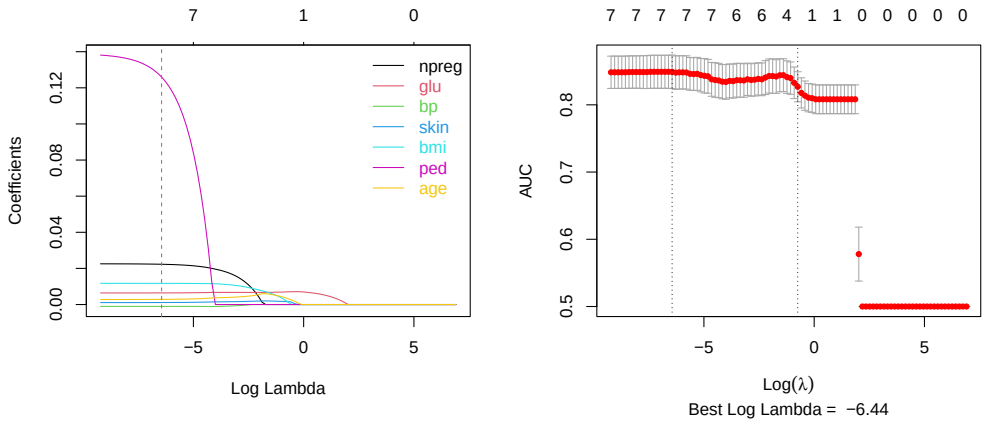


그림 5.9: LASSO logistic regression

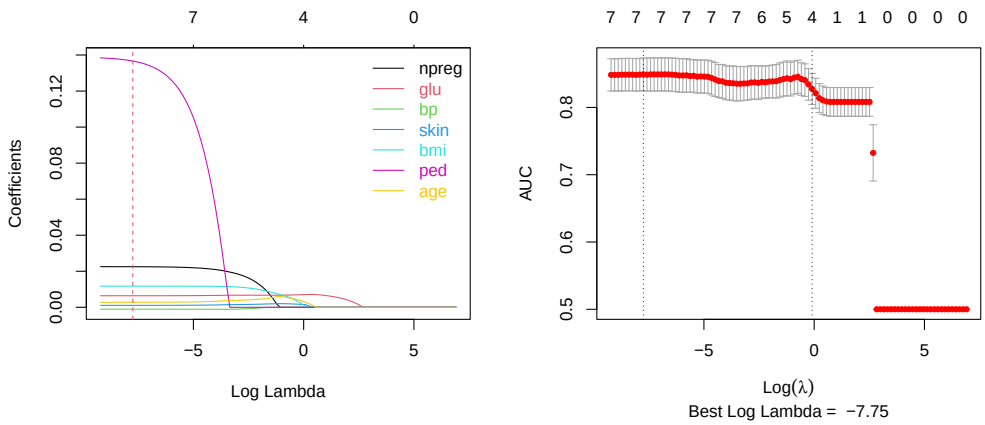


그림 5.10: Elasticnet regression

표 5.6: Coefficients of logistic regression

Variables	Logistic	Stepwise	Ridge	LASSO	Elasticnet
(Intercept)	-9.514	-9.552	-9.493	-1.019	-1.021
npreg	0.141	0.178	0.141	0.022	0.023
glu	0.037	0.038	0.037	0.006	0.006
bp	-0.009		-0.009	-0.001	-0.001
skin	0.013		0.013	0.001	0.001
bmi	0.079	0.084	0.079	0.012	0.012
ped	1.110	1.166	1.079	0.126	0.137
age	0.018		0.018	0.003	0.003

표 5.7: AUC of logistic regression

	Logistic	Stepwise	Ridge	LASSO	Elasticnet
AUC	0.869	0.867	0.868	0.868	0.868

5.6 분산분석모형

5.6.1 완전 확률화 디자인

분산분석모형중에 가장 간단한 모형으로 완전 확률화 디자인 (CRD : completely randomized design)을 소개한다. 먼저, 자료의 구조는 다음과 같다.

처리	관측치	평균	제곱합
1	$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$	$\bar{Y}_1$	$\sum_{j=1}^{n_1}(Y_{1j}-\bar{Y}_1)^2$
2	$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$	$\bar{Y}_2$	$\sum_{j=1}^{n_2}(Y_{2j}-\bar{Y}_2)^2$
$\vdots$			
k	$Y_{k1}, Y_{k2}, \cdots, Y_{kn_k}$	$\bar{Y}_k$	$\sum_{j=1}^{n_k}(Y_{kj}-\bar{Y}_k)^2$

여기서 $n \equiv \sum_{j=1}^{k} n_j$은 전체 관측치 수를 나타내고, $\bar{Y} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}/n$는 전체 평균을 나타낸다.

완전확률화 디자인의 목적은 처리효과 간에 차이가 있는지를 검정하고자 하는 것이다.

예제 : 4종류의 치료방법이 치료효과에 미치는 영향

처리	관측치	평균	제곱합
A	10, 15, 8, 12, 15	12	38
B	14, 18, 21, 15	17	30
C	17, 16, 14, 15, 17, 15, 18	16	12
D	12, 15, 17, 15, 16, 15	12	38

- $k = 4$, $n = 5 + 4 + 7 + 6 = 22$, $\bar{Y} = 15$
- 목표 : 치료방법에 따라 치료효과에 차이가 있는가?

완전 확률화 디자인의 추론은 일원분산분석모형 (One-way ANOVA model 또는 One-way classification model)을 통해 이루어진다. 보다 구체적으로,

$$Y_{ij} = \mu_j + \epsilon_{ij} \, , \;\; i = 1, \cdots, k \, ; \;\; j = 1, \cdots, n_i$$

- Y_{ij} : i 번째 처리 했을 때 j번째 관측값
- μ_j : j 번째 처리효과의 참값 (모수)
- $\epsilon_{ij} \sim N(0, \sigma^2)$: 오차항 (설명 안되는 부분)

목적 : $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ 를 검정

```
a <- data.frame(group="a", value=c(10,  15,  8,  12,  15))
b <- data.frame(group="b", value=c(14,  18,  21, 15))
c <- data.frame(group="c", value=c(17,  16,  14, 15,  17, 15, 18))
d <- data.frame(group="d", value=c(12,  15,  17, 15,  16, 15))
data <- rbind(a,b,c,d)

## Mean (SD)
tapply(data$value, data$group, function(x) sprintf("%.2f (%.2f)", mean(x), sd(x)))
```

```
#>              a              b              c              d
#> "12.00 (3.08)" "17.00 (3.16)" "16.00 (1.41)" "15.00 (1.67)"
#boxplot(value ~ group, data)

## Test for Homogeneity of Variances
bartlett.test(value ~ group, data)
#>
#>  Bartlett test of homogeneity of variances
#>
#> data:  value by group
#> Bartlett's K-squared = 4.123, df = 3, p-value = 0.2485

## Test for Equal Means
oneway.test(value ~ group, data, var.equal = T)
#>
#>  One-way analysis of means
#>
#> data:  value and group
#> F = 4.3404, num df = 3, denom df = 18, p-value = 0.01814
summary(aov(value ~ group, data))
#>             Df Sum Sq Mean Sq F value Pr(>F)
#> group        3     68  22.667    4.34 0.0181 *
#> Residuals   18     94   5.222
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.6.2 분할표 분석

두 변수가 모두 질적 자료 (첫 번째 변수는 r개의 범주, 두 번째 자료는 c개의 범주로 구성)인 경우 다음과 같은 $r \times c$ 분할표 (contingency table)로 나타낸다.

	1	2	$\cdots$	c
1	Y_{11}	Y_{12}	$\cdots$	Y_{1c}
2	Y_{21}	Y_{22}	$\cdots$	Y_{2c}
$\vdots$				
r	Y_{r1}	Y_{r2}	$\cdots$	Y_{rc}

분할표에서 $r = 1$인 경우는 c개의 범주에서 각 범주가 발생될 확률이 귀무가설에서 주어진 값과 같은지에 대한 검정이다. 이러한 검정을 피어슨의 적합도 검정 (Pearson's goodness-of-fit test)이라 한다. 즉, 귀무가설은

$$H_0 : p_1 = p_{10}, \cdots, p_c = p_{c0}$$

으로 주어지고, 이에 대한 검정통계량은

$$\chi^2 = \sum_{j=1}^{c} (Y_{1j} - np_{0j})^2 / np_{j0}$$

로 주어지며 이 검정통계량은 귀무가설하에서 자유도 $c-1$인 χ^2 분포를 따른다. 단, $n = \sum Y_{1j}$이다.

두 변수에 대한 분할표의 분석은 두 변수간의 관련성 여부에 대한 검정이다. 첫번째 변수 A가 i번째 범주 A_i를 취하고, 두번째 변수 B가 j번째 범주 B_j일 확률을 $p_{ij} = P(A_i \cap B_j)$라 하자. 한편, $p_{i.} = P(A_i), p_{.j} = P(B_j)$라고 하자. 이제, 두 변수간의 관련성 (association)이 없다는 귀무가설은

$$H_0 : p_{ij} = p_{i.}p_{.j}, \ \forall i, j$$

로 표현된다. 이에 대한 검정통계량은

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}(Y_{ij} - n(Y_{i.}/n)(Y_{.j}/n))^2/n(Y_{i.}/n)(Y_{.j}/n)$$

로 주어지며 이 검정통계량은 귀무가설하에서 자유도 $(r-1)(c-1)$인 χ^2 분포를 따른다. 단, $n = \sum_{i=1}^{r}\sum_{j=1}^{c} Y_{ij}$이다.

예제 : 흑색종 환자자료(Melanoma)에서 성별(sex)에 따라 궤양여부(ulcer)에 차이가 있는지를 양측검정하자. 독립표본 모비율검정 또는 카이제곱검정을 실시하자.

```
sex <- factor(Melanoma$sex, 0:1, c("Male","Female"))
ulcer <- factor(Melanoma$ulcer, 1:0, c("Yes","No"))

tab <- table(ulcer,sex); tab #
#>       sex
#> ulcer Male Female
#>   Yes   47     43
#>   No    79     36
addmargins(tab) #
#>       sex
#> ulcer Male Female Sum
#>   Yes   47     43  90
#>   No    79     36 115
#>   Sum  126     79 205
prop.table(tab, 2)*100 # %
#>       sex
#> ulcer     Male   Female
#>   Yes 37.30159 54.43038
#>   No  62.69841 45.56962

prop.test(c(47,43), c(126, 79), correct=F) #
#>
#>  2-sample test for equality of proportions without continuity
```

```
#>  correction
#>
#> data:  c(47, 43) out of c(126, 79)
#> X-squared = 5.7845, df = 1, p-value = 0.01617
#> alternative hypothesis: two.sided
#> 95 percent confidence interval:
#>  -0.3098209 -0.0327549
#> sample estimates:
#>    prop 1    prop 2
#> 0.3730159 0.5443038
chisq.test(tab, correct=F)
#>
#>  Pearson's Chi-squared test
#>
#> data:  tab
#> X-squared = 5.7845, df = 1, p-value = 0.01617
```

제 6 장

생존분석

6.1 생존시간과 중도절단

6.1.1 생존시간

의학 자료의 많은 부분은 생존시간 (survival time)에 관한 것이다. 생존시간은 흔히 실패시간 (failure time)이라고도 불리며 다음과 같은 여러 경우를 포함한다. 특정 질병에 걸린 사람이 사망할 때까지의 시간, 금연을 시도했다가 다시 흡연을 할 때까지의 시간, 특정 질병에 대한 치료 또는 수술 후 재발 (recurrence)까지 걸리는 시간, 범죄자가 다시 범죄를 저지를 때까지의 시간 등 생존시간은 다양한 의미로 사용된다.

생존기간을 나타내는 확률변수를 t라 하고, t는 확률밀도함수 $f(t)$, 분포함수 $F(t)$를 가질 때 생존함수 (survival function) $S(t)$는

$$S(t) = 1 - F(t) = P(T > t)$$

로 정의되며 위험율 또는 치사율 (hazard rate) $\lambda(t)$는

$$\lambda(t) = f(t)/S(t)$$

로 정의된다. $\lambda(t)$의 의미는

$$\begin{aligned}\lambda(t)\cdot\delta t &\simeq P(t<T<t+\delta \mid T>t)\\ &= P((t,\ t+\delta t)\text{사이에서 사망} \mid \text{t 까지 생존})\end{aligned}$$

이므로 t시점까지 생존한 사람이 t직후에 사망할 비율로 볼 수 있다.

$\lambda(t)$와 $S(t)$와의 관계는

$$\int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{1-F(u)}du = -log(1-F(t)) = -logS(t)$$

로부터

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right)$$

가 됨을 알 수 있는데, 흔히

$$\Lambda(t) = \int_0^t \lambda(u)du$$

를 **누적위험함수** (cumulative hazard function)라 부른다.

6.1.2 중도절단

자료는 흔히 완전자료 (complete data)와 불완전자료 (incomplete data)로 나눌 수 있다. 우리가 다루는 대부분의 자료는 완전자료로서 자료 그 자체가 원래의 의미를 제대로 반영하고 있는 경우이다. 예를 들어 어떤 사람의 몸무게가 $64\ kg$이라면 64라는 정보를 제대로 간직하고 있다. 물론 완전자료에는 측정오차 (measurement error)가 발생되었거나 반올림 오류 (round-off error) 등이 포함될 수 있어도 여전히 완전자료라고 부른다. 반면, 어떤 사람이 폐암에 걸린 나이는 65세와 69세 사이라고 할 때 이런 자료를 불완전 자료라고 부른다. 즉, 이 사람이 폐암에 걸린 나이를 정확히 알 수 없고 단지 구간만 알려져 있는 경우이다. 이런 형태의 불완전자료를 중도절단자료 (censored data)라 부른다. 불완전자료의 또 다른 예로서, 어떤 사람이 신체검사에서 (키, 몸무게, 혈압) 등 세 가지를 측정해야 하는데 그 중에서 키와

몸무게만 측정하고 혈압 측정치는 없는 경우, 즉 자료의 일부가 없는 경우도 역시 불완전자료가 된다. 이런 형태의 불완전자료는 빠진자료 (missing data)라 부른다.

의학 자료는 특성상 불완전자료일 경우가 많으며, 그 중에서도 특히 중도절단 자료가 많다. 여기서는 중도절단 자료의 종류를 소개하고 중도절단 자료를 분석하는 방법에 대해 자세히 소개한다. 중도절단은 크게 우측중도절단 (right censored data), 좌측중도절단 (left censored data), 구간중도절단 (interval censored data) 등으로 나눈다. 우측중도절단이란 생존시간을 끝까지 관측하지 못하는 경우이다. 즉, 사건 (사망, 재발, 재범 등 연구의 대상이 되는 것)이 발생할 때까지의 정확한 시간을 관측하지 못하고 연구를 종료하는 것이다. 우측중도절단에는 제1종 중도절단 (type I censoring), 제2종 중도절단 (type II censoring), 임의 중도절단 (random censoring) 등이 있다. 의학 자료는 주로 임의 중도절단인 경우가 많다. 좌측중도절단은 연구를 시작하는 시점에 이미 사건이 발생되어 정확한 생존시간을 알 수 없는 경우다. 예를 들어, 한국 남자 어린이의 어금니 영구치가 생기는 시점에 대한 연구를 위해 특정 초등학교 3학년을 대상으로 3년간 지켜보기로 하였는데 연구시작 시점에 어떤 학생은 이미 어금니 영구치가 발생하였다면 좌측 중도절단된 경우다. 한편, 구간 중도절단이란 사건의 발생시점을 정확하게 알 수 없고 구간만 아는 경우다. 예를 들어, 매년 정기검진을 하는데 전년도 검진에서는 나타나지 않았다가 금년도 검진에서 질병이 나타났다면 질병이 발생한 시점은 알 수 없고 단지 발생 구간만 아는 경우이다.
의학 자료는 주로 우측 중도절단 중에서도 임의 중도절단인 경우가 가장 많아 이에 대해 자세히 살펴보자.

실제 생존시간 $T_i,\ i = 1, \cdots, n$에 대응하는 중도절단 변수 $C_i,\ i = 1, \cdots, n$가 있을 때 우리는 $T_i,\ i = 1, \cdots, n$대신

$$Y_i = min(T_i, C_i),\ \delta_i = I(T_i \leq C_i),\ \ i = 1, \cdots, n$$

를 관측하게 되는데 여기서 $\delta_i = I(T_i \leq C_i),\ i = 1, \cdots, n$는 중도절단지수 (censoring indicator)라 불리운다. 즉, $T_i \leq C_i$이면 생존기간 T_i를 관측할 수 있고 $T_i > C_i$이면 T_i 대신 C_i만을 관측하게 된다.

그림 6.1에서 환자 1은 중도절단 안 된 것으로 T_1의 생존기간이 관측되었고, 환자 2는 연구종점까지 생존해 있으므로 T_2는 관측할 수 없고 중도절단시간 C_2만 관측되었다. 흔히, 이런 경우에 C_2를 T_2^+로 표현한다. 환자 3은 연구종점 이전에 중도절단된 것인데 다른 병원으로 옮겨서 더 이상 추적 (follow up)이 불가능하거나 연구대상이 아닌 다른 원인으로 사망한

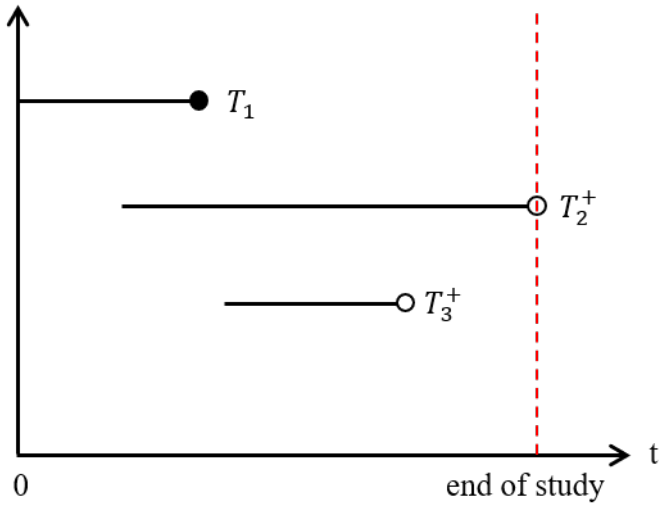


그림 6.1: 임의중도절단

경우이다.

6.2 생존함수의 추정

생존분석에서 가장 중요한 문제중의 하나는 생존함수 $S(t) = P(T > t)$를 추정하는 것이다. 생존함수의 추정은 함수의 형태를 가정하는 모수적 추정 (parametric estimation)과 함수의 형태를 가정하지 않는 비모수적 추정 (nonparametric estimation)이 있다.

6.2.1 모수적 추정

생존함수에 대한 모수적 추정은 위험율 함수나 생존함수를 특정함수로 가정하는 것이다. 따라서, 모수적 추정시에 가장 중요한 점은 가정한 함수가 자료의 성격을 제대로 반영하는 것인지 확인하는 것이다. 기본적으로 생존시간을 나타내는 확률변수는 언제나 양의 값을 가지므로 **지수** 분포, **와이불** 분포, **감마** 분포, **래일리** 분포, **로그-정규** 분포 등이 많이 이용된다.

지수분포 (exponential distribution)는 위험율 함수가 시간과 관계없이 언제나 일정한 상수로 가정한 것이다. 즉, $\lambda(t) = \lambda$ 로 가정한 것인데 이로 부터

$$S(t) = e^{-\lambda t},\ f(t) = \lambda e^{-\lambda t}$$

가 된다. 지수분포의 특징은 비기억성 (memoryless property)을 갖는 것인데 이는

$$P(T > t + s|T > s) \ = \ P(T > t),\ 0 < s < t < \infty$$

을 의미한다. 지수분포는 모수적 추정에서 가장 단순한 분포이므로 현실성이 떨어지지만 기초이론 전개와 모의실험 등에 많이 사용된다.

지수분포의 한계를 극복하기 위해 모수를 한 개 더 추가시켜 확장한 분포가 **감마**분포 (gamma distribution)와 **와이불** 분포 (Weibull distribution)이다. **감마**분포의 확률밀도함수는

$$f(t) \ = \ \frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}\ ,\ \alpha > 0,\ \lambda > 0$$

으로 주어지는데 $\alpha = 1$인 경우가 바로 지수분포이다. **감마**분포의 경우 위험율 함수 및 생존함수는 식으로 표현 할 수 없다 (does not have closed form expressions).

한편, **와이불** 분포의 생존함수는 $S(t) = exp(-(\lambda t)^{\alpha})$인데 이로 부터 $\lambda(t) = \alpha\lambda(\lambda t)^{\alpha-1}$ 임을 알 수 있다. 이 사실을 이용하면

$$f(t) \ = \ \alpha\lambda(\lambda t)^{\alpha-1} exp(-(\lambda t)^{\alpha})$$

으로 표현된다. **와이불** 분포의 생존함수에 **로그-로그** 변환을 하면

$$log[-logS(t)] = \alpha(logt + log\lambda)$$

이므로 $log[-log\hat{S}(t)]$와 $logt$의 관계가 선형으로 주어지면 **와이불** 분포의 가정이 옳음을 알 수 있다.

래일리 분포 (Rayleigh distribution)는 상수 위험율 (지수분포)을 선형 위험율로 확장한 것

이다. 즉, **래일리** 분포는 위험율 함수를

$$\lambda(t) \;=\; \lambda_0 + \lambda_1 t$$

으로 가정한 것인데 이로 부터 생존함수가

$$S(t) \;=\; exp(-\lambda_0 t - \frac{1}{2}\lambda_1 t^2)$$

임을 쉽게 알 수 있고 이로 부터 확률밀도함수는

$$f(t) \;=\; (\lambda_0 + \lambda_1 t)\, exp(-\lambda_0 t - \frac{1}{2}\lambda_1 t^2)$$

으로 표현된다.

한편, **로그-정규** 분포 (log-normal distribution)는 생존시간의 **로그**값인 $logT$ 가 정규분포 $N(\mu, \sigma^2)$를 따른다고 가정한 것이다. 이로 부터 생존함수는

$$S(t) = 1 - \Phi(\frac{logt - \mu}{\sigma})$$

이 된다. **로그-정규** 분포는 생존시간을 **로그** 변환한 후 정규분포로 가정하므로 매우 편리하게 사용될 수 있는 분포이다.

6.2.2 비모수적 추정

생존함수의 분포를 가정하지 않는 비모수적 방법의 대표적인 추정치로 카플란-마이어 추정치 (Kaplan-Meier estimator)를 소개한다.

기호의 편의상 관측치 $(Y_i,\, \delta_i),\, i = 1, \cdots, n$이 크기 순으로 나열되어 있다고 하자. 즉,

$$y_1 \le y_2 \le \cdots \le y_n$$

으로 가정하고 중도절단지수도 이와 대응되게 주어졌다고 가정하자. 이제 중도절단되지 않은 관측치 (즉, 사망이 일어난 경우)들 중 서로 겹치지 않는 관측치들의 갯수가 k개라 하고 이를

$$\tau_1 < \tau_2 < \cdots < \tau_k$$

이라 하자. 또한,

$$d_j = \sum_{i=1}^{n} I(y_i = \tau_j,\, \delta_i = 1)$$

를 τ_j 시점에서 발생한 사망자 수라 하고

$$n_j = \sum_{i=1}^{n} I(y_i \geq \tau_j)$$

는 τ_j 직전까지 사망하지 않은 관측치들의 개수라 하자. τ_j 직전까지 사망하지 않은 관측치들의 집합을 흔히 τ_j에서의 위험집합 (risk set)이라 부른다. 생존함수에 대한 **카플란-마이어** 추정치는

$$\hat{S}(t) = \prod_{j:\tau_j \leq t} (1 - \frac{d_j}{n_j})$$

으로 나타난다. 같은 논리로 누적위험함수에 대한 추정치를

$$\hat{H}(t) = \sum_{j:\tau_j \leq t} \frac{d_j}{n_j}$$

으로 표현할 수 있는데 이를 흔히 **넬슨-알렌** (Nelson-Aalen, 1969) 추정치라 부른다.

특별한 경우로 같은 값을 갖는 관측치들이 하나도 없는 경우 (no ties)를 고려해 보자. 이제는 $y_1 < y_2 < \cdots < y_n$이고 $\tau_j = y_j$라 두면 $d_j = \delta_j$, $n_j = n - j + 1$이 되므로 **카플란-마이어** 추정치는

$$\hat{S}(t) = \prod_{j:y_j \leq t}^{n} (1 - \frac{\delta_j}{n-j+1}) = \prod_{j:y_j \leq t}^{n} (1 - \frac{1}{n-j+1})^{\delta_j}$$

로 표현된다.

카플란-마이어 추정치 $\widehat{S}(t)$의 분산에 대한 추정치로 흔히 그린우드 공식 (Greenwood's formula)

$$\hat{\sigma}_S^2(t) \equiv \widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:\tau_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

을 주로 사용한다. 같은 값을 갖는 관측치들이 하나도 없는 경우에는 **그린우드** 공식이

$$\hat{\sigma}_S^2(t) = \hat{S}(t)^2 \sum_{j:y_j \le t} \frac{\delta_j}{(n-j)(n-j+1)}$$

으로 표현된다. 이를 이용하여 분포함수에 대한 신뢰구간을 구할 수 있는데 $100 \times (1-\alpha)\%$ 신뢰구간은

$$\widehat{S}(t) \pm z_{\alpha/2}\hat{\sigma}_S(t)$$

으로 주어진다. 여기서, 한 가지 주의해야 할 점은 t가 0에 가까운 값이거나 매우 큰 값이면 신뢰구간이 $(0,1)$을 벗어나는 값을 포함할 수 있다. 이를 피하기 위해 로짓변환 (logistic transformation)이나 로그-로그 변환등이 있다.

예제 : 흑색종 환자자료(Melanoma)를 이용하여 생존함수를 추정하자.

```
library(survival)
fit <- survfit(Surv(time/365.25, status==1) ~ 1, data=Melanoma)


## survival table
out <- summary(fit, time=seq(0,5,by=1))
```

표 6.1: Survival table

time	n.risk	n.event	surv	lower	upper
0	205	0	1.000	1.000	1.000
1	193	6	0.970	0.947	0.994
2	183	9	0.925	0.889	0.962
3	167	15	0.849	0.800	0.900
4	160	6	0.818	0.766	0.874
5	122	9	0.769	0.712	0.831

```
## Kaplan-Meier curve
#plot(fit, conf.int = T)


#library(survminer)
#ggsurvplot(fit, data=Melanoma)
```

6.2.3 로그-순위 검정

특정 암을 앓고 있는 환자에게 새로 개발된 항암제가 효과가 있는지 알아보기 위해 환자 전체를 임의로 두 그룹으로 나눈 다음 한 그룹에는 항암제를 투여하고 나머지 한 그룹에는 위약(placebo)을 투여한 후 시간이 경과함에 따라 각 환자들의 생존시간을 측정하였다. 항암제를 투여한 그룹의 생존함수를 $S_1(t)$, 위약을 투여한 그룹의 생존함수를 $S_2(t)$라 하면 항암제의 효과 여부는

$$H_0 : S_1(t) = S_2(t)$$

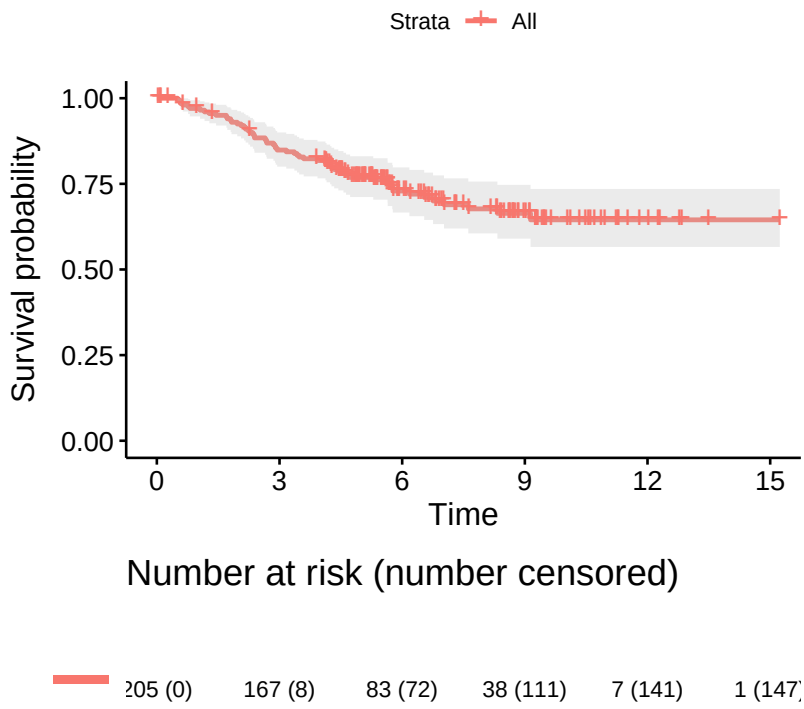


그림 6.2: Kaplan-Meier curve

라는 귀무가설을 검정하는 것과 같다. 중도절단 자료의 경우 가장 많이 사용되는 검정법으로 로그-순위 검정 (log-rank test)이 있는데 이는 흔히 Mantel test, Mantel-Cox test, 또는 Cochran-Mantel-Haenszel test 등으로 불리운다.

예제 : 흑색종 환자자료 (Melanoma)에서 ulcer (궤양여부)에 따른 생존함수의 차이에 대한 로그-순위 검정의 결과는 다음과 같다.

```
fit <- survfit(Surv(time/365.25, status==1) ~ ulcer, data=Melanoma)

## log-rank test
survdiff(Surv(time, status==1) ~ ulcer, data=Melanoma)
#> Call:
#> survdiff(formula = Surv(time, status == 1) ~ ulcer, data = Melanoma)
#>
#>            N Observed Expected (O-E)^2/E (O-E)^2/V
#> ulcer=0 115       16     35.8      10.9      29.6
#> ulcer=1  90       41     21.2      18.5      29.6
#>
#>  Chisq= 29.6  on 1 degrees of freedom, p= 5e-08

## Kaplan-Meier curve
#plot(fit, col=1:2, lty=1:2)

#library(survminer)
#ggsurvplot(fit, data=Melanoma)
```

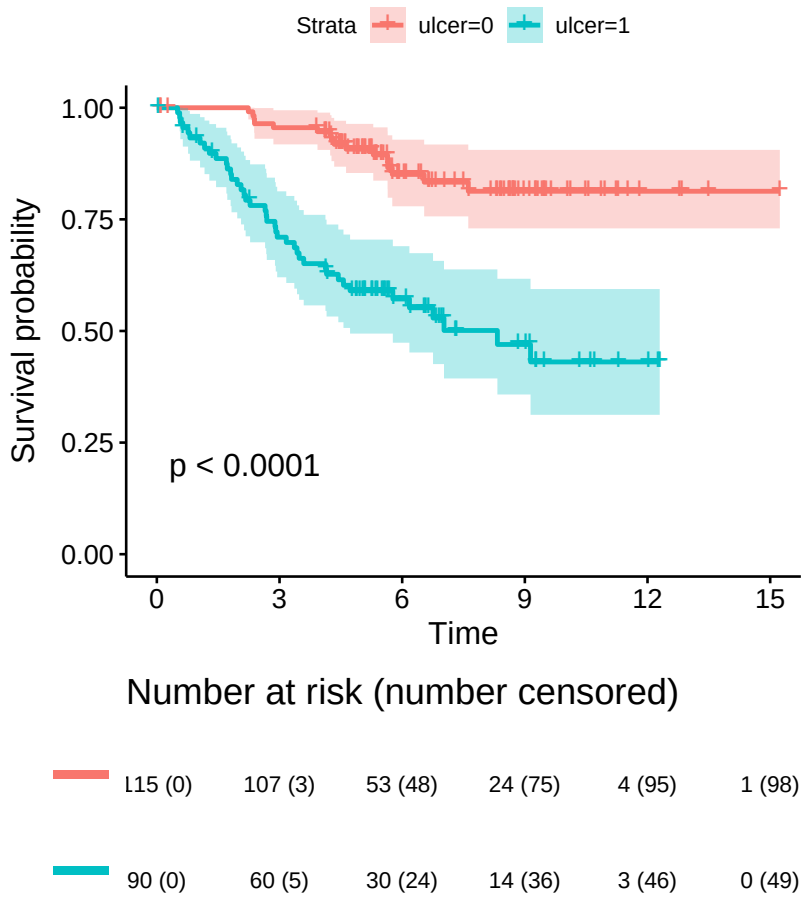


그림 6.3: log-rank test

6.2.4 중간값과 위험율의 비

항암제를 투여한 그룹 (treatment group)의 생존함수를 $S_T(t)$, 위약을 투여한 그룹 (control group, saline group)의 생존함수를 $S_C(t)$라 하자. 항암제의 효과 여부를 판단하는 검정으로 로그-순위 검정을 소개하였는데 보다 구체적으로 각 그룹의 중간값 median survival time)과 위험율(hazard rate)를 비교하는 통계량이 매우 유용하며 자주 언급된다. 중간값 m은

$$S(m) = 0.5$$

를 만족하는 값으로 생존시간을 크기순으로 나열했을 때 중간에 해당되는 값으로 해석할 수 있다. 이제 두 그룹의 중간값을 각각 m_T와 m_C라 하면 중간값 비 (MR, median survival times ratio)는

$$MR = m_T/m_C$$

로 정의하고 MR에 대한 신뢰구간의 하한값이 1보다 큰 값이면 항암제의 중간값이 위약의 중간값보다 유의하게 큰 것으로 판단한다. 같은 논리로 만약 신뢰구간이 1을 포함하면 중간값의 차이는 없는 것으로 판단한다.

한편, 위험율 $\lambda(t)$는 t시점까지 생존한 사람이 t직후에 사망할 비율로 정의하였으며 이는 t시점에서 순간 사망율 (instantaneous death rate)로 해석할 수 있다. 이제 t 시점에서 두 그룹의 위험율을 각각 $\lambda_T(t)$와 $\lambda_C(t)$라 하면 위험율 비 (HR, hazard ratio)는

$$HR(t) = \lambda_T(t)/\lambda_C(t)$$

로 정의하고 $HR(t)$에 대한 신뢰구간의 하한값이 1보다 작으면 항암제의 위험율이 위약의 위험율보다 유의하게 작은 것으로 판단한다. 같은 논리로 만약 신뢰구간이 1을 포함하면 위험율의 차이는 없는 것으로 판단한다. 위험율 비 $HR(t)$는 시간 t에 의존하므로 여러 시점에서 계산해야 한다. 이러한 불편을 없애기 위해 대부분의 경우 위험율 비는 시간에 관계없이 일정하다고 가정 (proportional hazard assumption)하여 하나의 값

$$\gamma := HR(t) = \lambda_T(t)/\lambda_C(t),\ \forall t$$

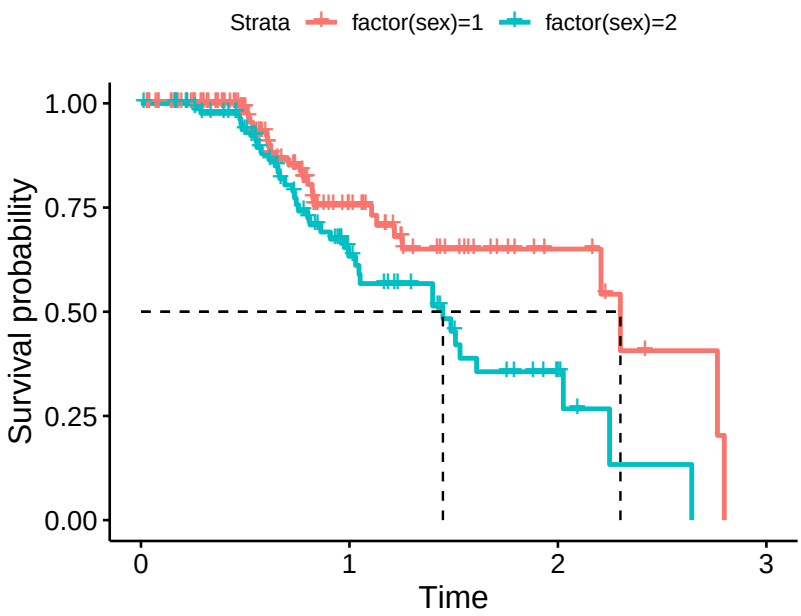
으로 주어진다. 즉, 시점에 관계없이 위험율의 비는 일정하다고 가정하며, 이 경우 두 그룹의 생존함수는

$$S_T(t) = S_C^{\gamma}(t)$$

이 됨을 쉽게 증명할 수 있다. 예를 들어 시점 $t = 10$에서 위약의 생존함수 값이 $S_C(10) = 0.36$이고 같은 시점 $t = 10$에서 항암제의 생존함수 값은 $S_T(10) = 0.6$이 라고 하면 위험율 비는 $\gamma = 0.5$로 추정할 수 있다.

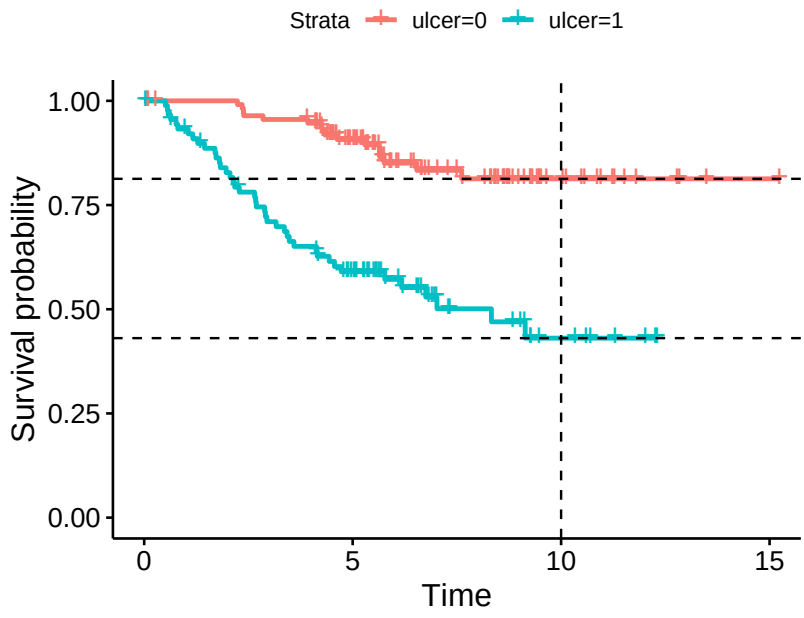
예제 : 폐암 환자자료(lung)를 이용하여 중간값 비 MR을 계산해 보자.

```
fit <- survfit(Surv(time/365.25, status==1) ~ factor(sex), data=lung,
               conf.type="none")
m <- quantile(fit, p=0.5); m
#>                      50
#> factor(sex)=1 2.299795
#> factor(sex)=2 1.448323
m[1] / m[2]
#> [1] 1.587902
```



예제 : 흑색종 환자자료(Melanoma)를 이용하여 $t = 10$에서 위험율 비 γ를 계산해 보자.

```
fit <- survfit(Surv(time/365.25, status==1) ~ ulcer, data=Melanoma)
st <- summary(fit, time=10)$surv; st
#> [1] 0.8129165 0.4306245
st[2] / st[1]
#> [1] 0.5297278
```



6.3 칵스회귀모형

6.3.1 칵스회귀모형

중도절단자료에서 공변량 (설명변수)이 생존시간 (반응변수)에 미치는 영향을 알기 위한 회귀모형을 구축하는 방법에는 크게 두 가지가 있다. 첫째, 공변량과 반응시간과의 함수 관계를 바로 설정하는 방법이다. 이는 앞에서 다루었던 완전자료에 대한 모수적 회귀와 비모수적 회귀 모두 여기에 해단된다. 또한, 다음 절에서 소개하는 선형회귀모형도 이 방법에 의한 것이다. 둘째, 공변량이 위험율 함수에 미치는 영향을 가정하는 방법이다. **칵스** (Cox, 1972)는 두 번째 방법으로 중도절단자료에서의 회귀모형을 다음과 같이 제안하였다.

편의상 $Y_1 < Y_2 < \cdots < Y_n$ 이라 하고 이에 해당되는 중도절단지수를 $\delta_1,\ \delta_2,\ \cdots,\ \delta_n$ 이라 하자. 이때 R_i 를 Y_i 직전까지 생존해 있는 환자의 집합을 나타낸다고 할 때 R_i 를 위험집합 (risk set)이라 부른다.

생존기간 $T_1,\ T_2,\ \cdots,\ T_n$ 이 임의중도절단에 의해 $(Y_1,\ \delta_1),\ \cdots, (Y_n,\ \delta_n)$ 만 관측가능하고, p 개의 설명변수로 구성된 $\mathbf{x}_1,\ \cdots,\ \mathbf{x}_n$ (즉, $\mathbf{x}_i = (X_{i1},\ \cdots,\ X_{ip})'$)이 있다고 하자. 또한, $\mathbb{x}$ 가 주어져 있을 때 시간 t 에서의 위험율을 $\lambda(t : \mathbb{x})$ 라 하고 $\mathbb{x} = 0$ 에서의 위험율, 즉, 기저 위험율 (baseline hazard)을 $\lambda_0(t)$ 라 하자. 이때

$$\lambda(t : \mathbf{x}) = \exp(\mathbf{x}'\beta)\lambda_0(t)$$

로 가정한 모형을 비례위험모형 (proportional hazard model)이라 한다.

비례위험모형에 대한 가정이 맞는지를 확인하는 방법 중의 하나로 각 집단의 생존함수 추정치들이 시간에 관계없이 일정한 차이를 유지하는 형태를 보여야 한다. 회귀계수는 최우추정치를 사용하는데 뉴튼-랩슨의 반복법 (Newton-Raphson iteration)에 의해 구해야 한다. 이렇게 구한 $\hat{\beta}$ 는 최우추정치의 근사적 정규성

$$\sqrt{n}(\hat{\beta} - \beta)\ \underrightarrow{D}\ N_p(0,\ I^{-1}(\beta))$$

이 만족됨을 증명할 수 있다. 여기서, $I^{-1}(\beta)$ 는 피셔정보행렬의 역행렬이다. 회귀계수의 추정치를 이용하여 주어진 공변량에 대한 생존함수를 추정할 수 있다. 먼저

$$S(t : \mathbb{x}) = [S_0(t)]^{exp(\mathbb{x}'\beta)}$$

, $S_0(t) = exp\{-\int_0^t \lambda_0(u)du\}$임을 보일 수 있는데 생존함수 $S(t:\mathbb{x})$를 추정하기 위해 β 와 λ_0 또는 Λ_0를 추정해야 하는데 β는 앞에서 구한 $\hat{\beta}$를 대입하면 되지만 λ_0 또는 Λ_0를 추정해야 하는 문제가 남아 있다. 이를 위해 세 가지 추정법을 소개한다. 우선 **브레즐로우** (Breslow, 1972)는

$$\hat{S}_0(t) = \prod_{i:y_i \le t}^{n} (1 - \frac{\delta_i}{\sum_{j \in R_i} exp(\mathbb{x}_j'\hat{\beta})})$$

을 제안하였는데 이 추정치의 약점은 음수를 가질 수 있다는 것이다.

한편, **치아티스** (Tsiatis, 1978)는

$$\hat{\Lambda}_0(t) = \sum_{i:y_i \le t}^{n} \frac{\delta_i}{\sum_{j \in R_i} exp(\mathbb{x}_j'\hat{\beta})}$$

을 제안하였는데 이는 계단함수이므로 링크 (Link, 1979)는 이 계산함수를 선형으로 연결한 함수를 제안하였다.

6.3.2 c-index

지금까지 중도절단된 생존시간에 대한 칵스회귀모형을 소개하였다. 적합된 칵스모형이 얼마나 좋은 것인가는 특정 환자의 공변량을 대입하였을 때 계산된 생존시간에 대한 예측값 (prediction)이 실제 그 환자의 생존시간과 얼마나 가까운가 하는 것이다. 이제 적합된 모형의 예측력을 평가하기 위해 다음과 같은 방법을 사용한다. 먼저 훈련자료를 이용하여 칵스모형에 적합시킨 후 시험자료의 관측치 $Y(\mathbf{x})$와 시험자료의 공변량을 훈련자료에 의하여 적합된 모형에 대입하여 구한 적합치 $\hat{Y}(\mathbf{x})$의 차이를 이용하여 모형의 예측력을 측정할 수 있다. 생존분석에서 모형간의 예측력을 비교하기 위한 통계량으로 c-index가 주로 이용되는데 c-index는 concordant index의 줄임말로 c-statistic 또는 concordance 라고도 한다.

c-index의 개념을 이해하기 위해 우선 완전자료의 경우부터 설명한다. c-index란 n개의 관측치 $Y_1(\mathbf{x}), \cdots, Y_n(\mathbf{x})$와 개의 적합지 $\hat{Y}_1(\mathbf{x}), \cdots, \hat{Y}_n(\mathbf{x})$이 있을 때 모든 가능한 쌍의 개수인 $\binom{n}{2} = n(n-1)/2$개 중에서 $Y_i(\mathbf{x}), Y_j(\mathbf{x})$의 대소 관계와 $\hat{Y}_i(\mathbf{x}), \hat{Y}_j(\mathbf{x})$의 대소 관계가 일치하는 쌍 (pair)의 비율을 나타내는 것이다. 좀 더 구체적으로 알아보자.

$n = 5$개의 관측치를 1, 2, 3, 4, 5 (즉, $Y_i(\mathbf{x}) = i,\ i = 1, \cdots, 5$)라 하고, 적합치를 1, 2, 3, 4, 5 (즉, $\hat{Y}_i(\mathbf{x}) = i, i = 1, \cdots, 5$)라 하면 모든 가능한 쌍의 개수는 10개이고 $Y_i(\mathbf{x}), Y_j(\mathbf{x})$의 대소 관계와 $\hat{Y}_i(\mathbf{x}), \hat{Y}_j(\mathbf{x})$의 대소 관계가 일치하는 쌍의 개수 역시 10개다. 따라서 c-index는 10/10=1.0이 된다. 이 경우 적합치가 10, 20, 30, 40, 50 (즉, $\hat{Y}_i(\mathbf{x}) = 10i,\ i = 1, \cdots, 5$)라 하더라도 c-index는 여전히 1.0이다. 즉, c-index는 대소 관계만 따지지 적합치 값 그 자체는 전혀 문제되지 않는다. 같은 이유로 적합치가 0.1, 0.2, 0.3, 0.4, 0.5 (즉, $\hat{Y}_i(\mathbf{x}) = 0.1i,\ i = 1, \cdots, 5$)라 하더라도 c-index는 여전히 1.0이다. 관측치는 여전히 1, 2, 3, 4, 5 라 하고, 적합치를 5, 4, 3, 2, 1 이라 하면 관측치의 대소 관계와 적합치의 대소 관계가 일치하는 쌍이 하나도 없으므로 c-index는 0/10 = 0.0이 된다. 만약 적합치가 1, 2, 3, 5, 4 라 하면 관측치의 쌍 (4,5)와 적합치의 쌍 (5,4)를 제외한 모든 경우의 대소 관계가 일치하므로 c-index는 9/10 = 0.9가 된다.

적합치가 같은 값 (tie)인 경우는 1/2의 가중치를 부여한다. 즉, 관측치는 여전히 1, 2, 3, 4, 5 라 하고, 적합치를 1, 1, 2, 3, 4 라 하면 관측치의 쌍 (1,2)와 적합치의 쌍 (1,1)을 제외한 모든 경우의 대소 관계가 일치하므로 c-index는 9.5/10 = 0.95가 된다.

실제로 c-index는 중도절단자료의 적합에 대한 측도이다. 여기서 한 가지 명심해야할 중요한 사실은 중도절단자료는 관측치에만 나타나고 적합치는 중도절단으로 표현되지 않는다는 사실이다. 자, 이제 3번째 관측치가 중도절단되었다고 하자. 즉, 관측치는 $1, 2, 3^+, 4, 5$라 하고, 적합치를 1, 2, 3, 5, 4 라 하자. 3^+는 3이상이므로 4, 5와는 대소 관계가 불가능하다. 즉, 완전자료의 경우에는 모든 가능한 쌍의 개수가 10가지였으나 3이 중도절단됨으로써 $(3^+, 4), (3^+, 5)$의 대소관계는 알 수 없다. 즉, 모든 가능한 쌍의 개수는 10개가 아닌 10-2=8개인 것이다. 이 8개의 쌍에서 적합치의 대소 관계가 일치하는 것은 (1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5)로서 7쌍이다. 즉, c-index는 7/8 = 0.875가 된다. 일반적으로 c-index는 다음과 같이 정의된다.

$$c = \frac{\sum_{i \neq j} \sum I(Y_i > Y_j) I(\hat{Y}_i > \hat{Y}_j) \delta_j}{\sum_{i \neq j} \sum I(Y_i > Y_j) \delta_j}$$

적합치를 임의로 정하는 경우 (random guess) c-index는 0.5에 가까운 값이 나오는데 이는 ROC곡선의 AUC와 같은 개념으로 생각할 수 있다. 적합된 모형이 어느 정도의 예측력을 가지기 위해서는 적어도 0.5 보다는 훨씬 큰 값 (0.7 내지 0.8 이상)이어야 한다.

예제 : 흑색종 자료를 이용하여 칵스회귀모형에 적합시켜 보자. 먼저, 모든 공변량을 사용하여 칵스모형에 적합시킨 결과는 다음과 같다. 각 공변량의 회귀계수 추정치 $\hat{\beta}$는 coef, 회귀계수 추정치의 지수승은 exp(coef), 회귀계수 추정치의 표준오차는 se(coef), z는 표준화 계수 coef / se(coef), $Pr(>|z|)$는 $p-$value를 나타낸다. 이러한 값들이 의미하는 바를 이해하기 위해 칵스모형의 정의를 다시 살펴보자.

$$\lambda(t:\mathbf{x}) = \exp(\mathbf{x}'\beta)\lambda_0(t)$$

에서 $\lambda(t:\mathbf{x})/\lambda_0(t) = \exp(\mathbf{x}'\beta)$ 이므로 어떤 공변량 x가 1단위 증가할 때 회귀계수 추정치의 지수승이 바로 위험율 (hazard rate)의 증가치가 되는 것이다. 예를 들어, 암 두께 (thickness)의 경우 $\hat{\beta} = 0.109,\ exp\hat{\beta} = 1.115,\ se(\hat{\beta}) = 0.038,\ z = 0.109/0.038 = 2.887$로서 암 두께가 $1mm$ 증가하면 흑색종 위험율이 2.887배 증가한다는 뜻으로 해석할 수 있다. 그 다음 출력물은 회귀계수 추정치의 지수승에 대한 95% 신뢰구간이 제시되어 있다. 회귀계수의 지수승 $exp(\beta)$에 대한 95% 신뢰구간은 $exp(\hat{\beta} \pm 1.96se(\hat{\beta}))$으로 주어진다. 예를 들어, 암 두께 (thickness)의 경우 $exp(0.109 \pm 1.96 \times 0.0378) = (1.0356, 1.201)$ 이 된다. 일반적으로 회귀계수의 지수승 $exp(\beta)$에 대한 95% 신뢰구간이 1을 포함하지 않으면 $p-$value는 0.05보다 작게 되어 위험율에 유의하게 영향을 미친다고 할 수 있다. 이런 관점에서 흑색종 암의 경우 sex와 age는 위험율에 유의한 공변량이 아니고, thickness와 ulcer는 유의한 공변량이다.

한편, 흑색종 자료에 대한 칵스 회귀모형의 c-index는 0.753으로 예측력이 비교적 뛰어나다고 할 수 있다.

cox regression

```
fit <- coxph(Surv(time, status==1) ~ ., data=Melanoma)
summary(fit)
#> Call:
#> coxph(formula = Surv(time, status == 1) ~ ., data = Melanoma)
#>
#>   n= 205, number of events= 57
#>
#>                coef exp(coef)  se(coef)      z Pr(>|z|)
#> sex        0.448121  1.565368  0.266861  1.679 0.093107 .
```

```
#> age        0.016805  1.016947  0.008578  1.959 0.050094 .
#> year      -0.102566  0.902518  0.061007 -1.681 0.092719 .
#> thickness  0.100312  1.105516  0.038212  2.625 0.008660 **
#> ulcer      1.194555  3.302087  0.309254  3.863 0.000112 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#>           exp(coef) exp(-coef) lower .95 upper .95
#> sex          1.5654     0.6388    0.9278     2.641
#> age          1.0169     0.9833    1.0000     1.034
#> year         0.9025     1.1080    0.8008     1.017
#> thickness    1.1055     0.9046    1.0257     1.191
#> ulcer        3.3021     0.3028    1.8012     6.054
#>
#> Concordance= 0.757  (se = 0.031 )
#> Likelihood ratio test= 44.4  on 5 df,   p=2e-08
#> Wald test            = 40.89  on 5 df,   p=1e-07
#> Score (logrank) test = 48.14  on 5 df,   p=3e-09

#fit$concordance
concordance(fit)
#> Call:
#> concordance.coxph(object = fit)
#>
#> n= 205
#> Concordance= 0.7572 se= 0.03115
#> concordant discordant     tied.x     tied.y    tied.xy
#>       6559       2103          0          0          0
```

회귀모형에서 사용된 공변량중 반응변수를 가장 잘 설명할 수 있는 변수들만 선택하여 모형을 간결하게 만들어야 하며 이를 변수선택 (variable selection) 과정이라 한다. 칵스회귀모형 또한 예외가 아니며, 앞으로 부터 선택 (forward selection), 뒤로 부터 제거 (backward eleimination), 단계적 선택 (stepwise selection) 등이 있다. 흑색종 자료에서 단계적 선택에 대한 R code 및 출력은 다음과 같다.

variable selection method

```
m_null <- coxph(Surv(time, status==1) ~ 1, data=Melanoma)
m_full <- coxph(Surv(time, status==1) ~ ., data=Melanoma)

### forward variable selection
fit <- step(m_null, scope=list(lower=formula(m_null),
            upper=formula(m_full)), direction="forward", trace=F)

### backward variable elimination
fit <- step(m_full, direction="backward", trace=F)

### stepwise variable selection
fit <- step(m_null, scope=list(lower=formula(m_null),
            upper=formula(m_full)), direction="both", trace=F)
```

Result of stepwise variable selection

```
#> Start:  AIC=566.4
#> Surv(time, status == 1) ~ 1
#>
#>             Df    AIC
#> + ulcer      1 539.96
#> + thickness  1 549.21
#> + sex        1 562.25
#> + age        1 563.40
#> <none>         566.40
```

```
#> + year       1 566.75
#>
#> Step:  AIC=539.96
#> Surv(time, status == 1) ~ ulcer
#>
#>             Df    AIC
#> + thickness  1 533.96
#> + sex        1 538.24
#> + age        1 538.60
#> <none>         539.96
#> + year       1 540.44
#> - ulcer      1 566.40
#>
#> Step:  AIC=533.96
#> Surv(time, status == 1) ~ ulcer + thickness
#>
#>             Df    AIC
#> + sex        1 533.01
#> + age        1 533.39
#> <none>         533.96
#> + year       1 534.99
#> - thickness  1 539.96
#> - ulcer      1 549.21
#>
#> Step:  AIC=533.01
#> Surv(time, status == 1) ~ ulcer + thickness + sex
#>
#>             Df    AIC
#> + age        1 532.78
#> <none>         533.01
#> + year       1 533.91
#> - sex        1 533.96
```

```
#> - thickness  1 538.24
#> - ulcer      1 546.58
#>
#> Step:  AIC=532.78
#> Surv(time, status == 1) ~ ulcer + thickness + sex + age
#>
#>             Df    AIC
#> + year       1 532.00
#> <none>         532.78
#> - age        1 533.01
#> - sex        1 533.39
#> - thickness  1 537.55
#> - ulcer      1 546.50
#>
#> Step:  AIC=532
#> Surv(time, status == 1) ~ ulcer + thickness + sex + age + year
#>
#>             Df    AIC
#> <none>         532.00
#> - year       1 532.78
#> - sex        1 532.80
#> - age        1 533.91
#> - thickness  1 535.72
#> - ulcer      1 546.62
#> Call:
#> concordance.coxph(object = fit.step)
#>
#> n= 205
#> Concordance= 0.7572 se= 0.03115
#> concordant discordant     tied.x     tied.y    tied.xy
#>       6559       2103          0          0          0
```

Penalized Cox regression

```
library(glmnet)
x = model.matrix(Surv(time, status==1) ~., data=Melanoma)[,-1]
y = Surv(Melanoma$time, Melanoma$status==1)

## lambda: tuning parameter
grid=10^seq(-3, 2, length=100)

## alpha: Ridge=0, Lasso=1, Elasticnet=0.5
alpha=1; set.seed(1234)
cv <- cv.glmnet(x, y, lambda=grid, alpha=alpha, standardize=F,
                family="cox", type.measure = "C")
(bestlam = cv$lambda.min); (bestlam2 = cv$lambda.1se)
#> [1] 0.00178865
#> [1] 0.05857021

## coefficients
fit <- glmnet(x, y, lambda=grid, alpha=alpha, standardize=F, family="cox")
coef(fit, s=bestlam)
#> 5 x 1 sparse Matrix of class "dgCMatrix"
#>                      1
#> sex        0.42388058
#> age        0.01675322
#> year      -0.10041329
#> thickness  0.10113888
#> ulcer      1.16178035

## c-index
pred <- predict(fit, x, s=bestlam)
Cindex(pred, y)
#> [1] 0.7573309
```

```
## Plot of Lasso regression
par(mfrow=c(1,2))
plot(fit, xvar="lambda", col=1:8); abline(v=log(bestlam), col=2, lty=2)
legend("topright", colnames(x), col=1:8, text.col=1:8, lty=1, bty="n")

plot(cv); title(sub=paste("Best Log Lambda = ", round(log(bestlam), 2)))
```

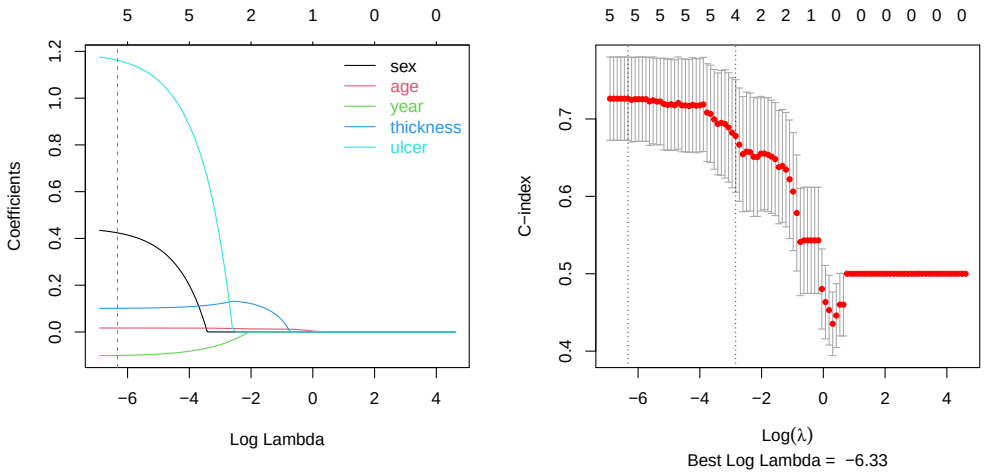


그림 6.4: Plot of Lasso regression

6.4 머신러닝

6.4.1 기계학습

훈련자료에서 공변량 $\mathbf{x}_i$와 그에 해당되는 반응변수 y_i를 이용하여 모형을 적합시키는 과정을 기계학습 (machine learning)이라 한다. 기계학습의 1차적 목표는 새로운 공변량을 학습된 모형에 대입하였을 때 반응변수 값을 예측하기 위한 것이다. 이러한 기계학습과정은 인공지능 (artificial intelligence)의 가장 핵심적 기술이다. 즉, 뛰어난 인공지능이란 제대로 학습된 기계라는 것이다.

기계학습은 반응변수의 형태에 따라 크게 두 가지로 나뉜다. 반응변수가 연속형이면 회귀모형 (regression model)이 되고, 이산형이면 분류모형 (classification model)이 된다. 한편, 반응변수가 아예없는 경우는 흔히 군집화 (clustering)이 목적인 경우가 대부분이다. 기계학습의 종류는 매우 많은데 대부분의 기계학습 방법들은 회귀모형과 분류모형 둘 다 적용이 가능하다.

6.4.2 기계학습의 종류

기계학습의 방법에는 매우 많은 종류가 있으나 대표적인 몇 가지만 소개한다. 기계학습의 방법에 대한 자세한 내용들은 본 도서에서 다루기에는 너무 방대하여 해당 문헌들을 참고하기 바란다. 본 교재의 1저자인 김충락 교수의 KMOOC 강좌 “AI 연구자를 위한 통계적 학습론”도 그 중의 하나로 소개한다.

(1) 로지스틱 회귀모형 (logistic regression model)

(2) 서포트 벡터기계 (svm - support vector machine)

(3) 배깅 (bagging)

(4) 부스팅 (boosting)

(5) 랜덤 포리스트 (random forest)

(6) 심층신경망 (dnn - deep neural network)

예제 : 흑색종 환자(Melanoma)의 생존여부 예측을 위한 머신러닝 모형으로 로지스틱회귀, SVR, random forest, boosting 등을 소개한다.

```
library(MASS)
data(Melanoma)
str(Melanoma)
#> 'data.frame':    205 obs. of  7 variables:
#>  $ time     : int  10 30 35 99 185 204 210 232 232 279 ...
#>  $ status   : int  3 3 2 3 1 1 1 3 1 1 ...
#>  $ sex      : int  1 1 1 0 1 1 1 0 1 0 ...
#>  $ age      : int  76 56 41 71 52 28 77 60 49 68 ...
#>  $ year     : int  1972 1968 1977 1968 1965 1971 1972 1974 1968 1971 ...
#>  $ thickness: num  6.76 0.65 1.34 2.9 12.08 ...
#>  $ ulcer    : int  1 0 0 0 1 1 1 1 1 1 ...
```

training data and test data

```
library(caret)
library(tidyverse)


set.seed(1234)
wd <- Melanoma %>% mutate(status = factor(as.numeric(status==1)))
id <- createDataPartition(wd$status, p=0.75, list=F)
train <- wd[id,]; test <- wd[-id,]


DescTools::Freq(train$status)
#>    level  freq   perc  cumfreq  cumperc
#> 1      0   111  72.1%      111    72.1%
#> 2      1    43  27.9%      154   100.0%
DescTools::Freq(test$status)
#>    level  freq   perc  cumfreq  cumperc
#> 1      0    37  72.5%       37    72.5%
#> 2      1    14  27.5%       51   100.0%
```

logistic regression

```
library(pROC)

ff <- formula("status ~ time + sex + age + year + thickness + ulcer")

fit <- glm(status=='1' ~., data=train, family=binomial)
pred <- predict(fit, test, type="response")
roc(test$status, pred)$auc
#> Area under the curve: 0.9228
```

classification trees

```
require(rpart)
fit <- rpart(status ~., data=train)
pred <- predict(fit, test)[,'1']
roc(test$status, pred)$auc
#> Area under the curve: 0.7809
```

bagging

```
library(adabag)
fit <- bagging(status ~ ., data=train, mfinal=10)
pred <- predict(fit, test)$prob[,2]
roc(test$status, pred)$auc
#> Area under the curve: 0.916
```

boosting

```
library(adabag)
fit <- boosting(status ~ ., data=train, mfinal=10)
pred <- predict(fit, test)$prob[,2]
roc(test$status, pred)$auc
#> Area under the curve: 0.9093
```

random forest

```
library(randomForest)
fit <- randomForest(status ~ ., data=train)
pred <- predict(fit, test, type="prob")[,2]
roc(test$status, pred)$auc
#> Area under the curve: 0.9237
```

support vector machine (SVM)

```
library(e1071)
fit <- svm(status ~ ., data=train, probability=T)
pred <- predict(fit, test, probability=T)
pred <- attr(pred, "probabilities")[,2]
roc(test$status, pred)$auc
#> Area under the curve: 0.9537
```

deep neural network (DNN)

```
library(neuralnet)
fit <- neuralnet(ff, data=train, hidden=c(10,10), linear.output=F)
pred <- compute(fit, test)$net.result[,2]
roc(test$status, pred)$auc
#> Area under the curve: 0.8359
```

표 6.2: Comparison of AUC

Logit	CART	Bagging	Boosting	RandomForest	SVM	DNN
0.923	0.781	0.916	0.909	0.924	0.954	0.836

예제 : 흑색종 환자(Melanoma)의 생존시간 예측을 위한 머신러닝 모형(칵스회귀, SVR, random forest, boosting 등)을 소개한다.

training data and test data

```
library(caret)
library(tidyverse)

set.seed(1234)
wd <- Melanoma %>% mutate(status = as.numeric(status==1))
id <- createDataPartition(wd$status, p=0.75, list=F)
train <- wd[id,]; test <- wd[-id,]
```

Cox regression

```
library(survival)
library(Hmisc)

ff <- formula("Surv(time, status) ~ sex + age + year + thickness + ulcer")

fit <- coxph(ff, data=train)
concordance(fit, newdata = test)$concordance
#> [1] 0.7524272

pred <- predict(fit, test)
rcorr.cens(-pred, Surv(test$time, test$status==1))["C Index"]
#>   C Index
#> 0.7524272
```

survival trees

```
require(rpart)
fit <- rpart(ff, data=train)
pred <- predict(fit, test)
rcorr.cens(1-pred, Surv(test$time, test$status==1))["C Index"]
#>   C Index
#> 0.6294498
```

random forest

```
library(ranger)
fit <- ranger(ff, data=train)
1-fit$prediction.error
#> [1] 0.7060345
pred <- predict(fit, test)$survival[,length(fit$unique.death.times)]
rcorr.cens(pred, Surv(test$time, test$status))["C Index"]
#>   C Index
#> 0.7119741
```

support vector machine (SVM)

```
library(survivalsvm)
fit <- survivalsvm(ff, data=train, gamma.mu=0.1)
pred <- predict(fit, test)
conindex(pred, Surv(test$time, test$status))
#>   C Index
#> 0.4708738
rcorr.cens(pred$predicted, Surv(test$time, test$status))["C Index"]
#>   C Index
#> 0.4708738
```

boosting

```
library(gbm)
fit <- gbm(ff, data=train, distribution="coxph")
pred <- predict(fit, test)
rcorr.cens(1-pred, Surv(test$time, test$status==1))["C Index"]
#>   C Index
#> 0.7249191
```

표 6.3: Comparison of C-index

	Cox	SurvialTree	RandomForest	SVM	Boosting
C Index	0.752	0.629	0.712	0.471	0.725

제 7 장

임상시험

7.1 임상연구의 종류

임상연구 (clinical studies) 또는 임상시험 (clinical trials)은 항암치료를 위한 신약개발등 인간의 질병에 대한 원인 파악 및 질병치료를 위한 약의 개발 등에 관한 연구를 일컫는다. 임상연구는 자료의 형태에 따라 크게 두 가지로 나누는데 관측연구 (observational study)와 실험연구 (experimental study)등이 있다.

7.1.1 관측연구

관측연구는 다시 3가지로 분류되는데 첫째, 이미 발생한 사건에 대해 사례군 (case group)과 대조군 (control group)으로 나누어서 두 군간의 차이를 연구하는 것으로 이를 후향적 연구 (retrospective study)라 부른다. 예를 들어, 아스피린을 자주 복용하는 것이 위궤양에 어떤 영향을 주는가를 알아보기 위해 위궤양이 있는 환자 (case) 50명에 대해 아스피린의 복용여부를 조사하고, 이들과 나이, 성별, 사회적 지위, 경제적 정도 등이 비슷한 사람들 중에서 위궤양이 없는 환자 (control) 50명에 대해 아스피린의 복용여부를 조사하여 연구하는 것이다. 둘째, 특정 집단 (cohort)을 미리 정해놓고 집단의 일부는 원인에 노출시키고 (exposed group) 집단의 나머지는 노출시키지 않고 (unexposed group) 일정한 기간 동안 두 그룹의 차이를 관찰하는 것이다. 예를 들어, 새로 개발된 인플루엔자 백신이 인플루엔자의 예방에 효과가 있는지 알아보기 위해, 35명에게 백신을 투여하고, 38명에게 식염수를 투여했다. 투여 6주 후에

혈액 속에 있는 HIA (Haemagglutinin Inhibiting Antibody) 양을 조사하여 백신의 효과를 판단하는 연구다. 이는 흔히 집단 연구 (cohort study) 또는 전향적 연구 (prospective study)라 부른다. 셋째, 현황연구 (cross-sectional study)로서 지금 현재의 상황에 대해 연구한다. 예를 드어, 현재 인구 10만명 당 결핵 감염자 수 등인데 여러 시간에 걸친 현황연구의 결과는 종단연구 (longitudinal study) 분석의 기초 자료를 제공한다.

7.1.2 실험연구

실험연구는 어떤 처리 (treatment)의 효과가 있는지 알기 위해 한 그룹은 처리군 (또는 시험군 treatment group), 또 다른 그룹은 대조군 (control group)으로 놓고 두 그룹간의 차이를 여러 각도에서 살펴보는 연구다. 이런 점에서 실험연구는 관측연구의 전향적 연구와 유사하다.

7.2 임상연구의 단계

임상연구의 목적 중 가장 중요한 것으로 신약개발이며, 신약개발은 다음의 4가지 단계를 거쳐야 한다.

7.2.1 전임상시험 (preclinical trial)

질병의 치료에 효과가 있을 것으로 판단되는 시험약 (신약의 후보물질)은 바로 인간에게 실험할 수 없다. 예를 들어, 항암제를 개발하는 경우 암세포를 주입한 쥐 또는 토끼 등의 동물에게 시험약을 투여하여 항암효과가 있는지 독성 (toxicity)이 있는지 등을 연구하며 효과가 없거나 독성이 나타나면 이 시험약은 이 단계에서 탈락하게 된다.

7.2.2 1상 임상시험 (phase I clinical trial)

임상약리시험 (human pharmacology trial)이라고도 한다. 전임상단계를 통과하면 20-80명의 사람을 대상으로 시험약을 투여하고 투여된 약의 흡수, 대사, 내약성 (tolerance) 등을 평가하며 만족할 만한 평가가 되지 못하면 시험약은 이 단계에서 탈락한다.

7.2.3 2상 임상시험 (phase II clinical trial)

치료적 탐색 임상시험 (therapeutic exploratory trial)이라고도 한다. 시험약의 여러 용량을 선정하여 사람들에게 투여한 후 최적의 용량을 찾는다. 최적의 용량 (MTD : maximum tolerable dose)에서 효과가 없거나 심각한 부작용이 발생되면 이 단계에서 탈락한다.

7.2.4 3상 임상시험 (phase III clinical trial)

치료적 확증 임상시험 (therapeutic confirmatory trial)이라고도 한다. 2상에서 선택된 최적의 용량을 시험군과 대조군에 투여한 후 약의 효과를 5%의 유의성으로 판단하게 된다. 이 단계에서 유의한 결과가 나오면 심사기관의 심사를 거쳐 시판허가를 받게 된다.

7.3 임상연구의 검정

새로운 약의 효과를 , 기존 약의 효과를 라 하자. 임상 연구의 주 목적은 대부분 신약 개발에 있으며 새로 개발된 약의 효과 (μ_T)가 기존 약의 효과 (μ_C) 보다 더 치료 효과가 뛰어남을 보이는 것이 대부분이다. 이 경우에 사용되는 검정은 우월성 검정 (superior test)이다. 하지만, 새로운 약의 효과가 기존약의 효과와 동등(equivalent)함을 보이거나 더 나쁘지 않음(non-inferior)을 보이는 것을 목적으로 하는 경우도 있다. 예를 들어, 기존 약보다 치료효과는 비슷하지만 훨씬 가격이 저렴하거나, 부작용이 작거나, 복용하기 수월하다면 새로 개발된 약은 그 만한 가치가 있다고 판단되기 때문이다. 여기서 약효가 유사하다는 정도를 나타내는 값으로 마진(margin) Δ로 나타내며 이는 흔히 임상시험계획단계에서 정해져야 한다.

이제 세 가지 검정은 새로운 약과 기존 약의 효과의 차이, 즉 $\mu_T - \mu_C$에 근거하는데 이를 검정하기 위한 검정 통계량은 각 집단 (신약 처리군 및 기존 약 처리군)의 분포에 대한 가정에 따라 주어진다. 예를 들어, 정규성 가정하에서는 이표본 t-검정 통계량이 될 것이고 정규성 가정이 의심스러운 경우 흔히 비모수 검정통계량 (예를 들어 Wilcoxon signed rank test)이 사용된다. 대부분의 임상연구 교재에서는 이러한 검정통계량을 바탕으로 귀무가설의 기각역을 나타내고 있는데 이는 매우 혼란스럽고 직관적이지 못하여 (non-intuitive) 여기서는 매우 쉽고 직관적인 방법으로 $\mu_T - \mu_C$의 95% 신뢰구간 (L, U)을 이용한 방법을 소개한다 (임상연구에서는 대부분의 경우 유의수준을 0.05로 가정한다).

7.3.1 우월성 검정

우월성 검정의 귀무가설은 "새로운 약이 기존 약의 치료효과보다 Δ 만큼만 크다"이며 대립가설은 "새로운 약이 기존 약의 치료효과에 비해 Δ 이상이다"이다. 즉,

$$H_0 : \mu_T - \mu_C = \Delta, \;\; H_1 : \mu_T - \mu_C > \Delta$$

로 나타낼 수 있으며 $L > \Delta$이면 귀무가설을 기각한다. 즉, $\mu_T - \mu_C$의 95% 신뢰구간의 하한값이 Δ보다 크면 귀무가설을 기각한다. 대부분의 임상시험에서는 우월성 검정의 경우 $\Delta = 0$으로 가정하고 실시한다.

참고 : 우월성 검정을

$$H_0 : \mu_T - \mu_C = 0, \;\; H_1 : \mu_T - \mu_C \neq 0$$

로 표현하는 문헌도 있다. 이 경우 검정 통계량을 Z라고 하면 기각역이 $|Z| > z_{\alpha/2}$로 주어지지만 실제로 $Z > z_{\alpha/2}$인 경우에만 귀무가설을 기각해야 한다. 왜냐하면 $Z < -z_{\alpha/2}$인 경우에는 우월성이 아닌 열등성이기 때문이다. 대립가설이 이렇게 주어지는 경우 p-value 계산은 실제값의 1/2이 되어야 한다. 즉, 대립가설이 단측검정이든 양측검정이든 무관하게 결과가 같이 주어져야 하므로 가장 단순한 방법은 앞에서 언급한 것 처럼 신뢰구간을 이용하는 것이며 이를 바탕으로 p-value를 계산하는 것이 혼란의 여지를 없다.

7.3.2 동등성 검정

동등성 검정의 귀무가설은 "새로운 약이 기존 약과 치료효과가 다르다"이며 대립가설은 "새로운 약이 기존 약과 치료효과가 유사하다"이다. 즉,

$$H_0 : |\mu_T - \mu_C| > \Delta, \;\; H_1 : |\mu_T - \mu_C| \leq \Delta$$

로 나타낼 수 있으며 여기서, Δ는 미리 주어진 마진이다. 동등성 검정에서는 $-\Delta \leq L < U \leq \Delta > 0$이면 귀무가설을 기각한다. 즉, $\mu_T - \mu_C$의 95% 신뢰구간이 $(-\Delta, \Delta)$에 속하면 귀무가설을 기각한다.

7.3.3 비열등성 검정

비역등성 검정의 귀무가설은 “새로운 약이 기존 약보다 치료효과가 열등하다(나쁘다)”이며 대립가설은 “새로운 약이 기존 약보다 치료효과가 열등하지 않다”이다. 즉,

$$H_0 : \mu_T - \mu_C < -\Delta, \;\; H_1 : \mu_T - \mu_C \geq -\Delta$$

로 나타낼 수 있으며 $L > -\Delta$이면 귀무가설을 기각한다. 즉, $\mu_T - \mu_C$의 95% 신뢰구간의 하한값이 $-\Delta$보다 크면 귀무가설을 기각한다.

7.4 2 x 2 표의 연관성 분석

임상연구에서 발생한 자료는 흔히 다음과 같은 2×2 표의 형태를 갖는다.

	질병	정상	합
실험군	n_{11}	n_{12}	$n_{1.}$
대조군	n_{21}	n_{22}	$n_{2.}$
합	$n_{.1}$	$n_{.2}$	$n_{..}$

이제 질병 여부가 실험군과 대조군에 동일하게 나타나는지 다르게 나타나는지를 알고 싶은데 이를 연관성 분석 (association analysis)이라 하며 이를 나타내는 측도로 상대 위험도와 오즈비를 소개한다.

(1) 상대 위험도 (relative risk)

상대 위험도(RR)는 관측연구에서 전향적 연구 (코호트 연구) 또는 실험연구에 사용되는 측도로서 실헝군의 질병율과 대조군의 질병율의 비 (ratio)로 정의된다. 즉,

$RR = \frac{n_{11}/n_{1.}}{n_{21}/n_{2.}}$

으로 나타낼 수 있으며 만약 RR이 1보다 크면 실험군의 질병율이 대조군의 질병율보다 크다는 것이고 1보다 작으면 그 반대로 해석한다. 또한, RR에 대한 95% 신뢰구간 (L, U)이 1 보다 크거나 (즉, $L > 1$) 1보다 작으면 (즉, $U < 1$) p-value는 0.05보다 작아진다.

(예) 아스피린 복용 여부가 심근경색 (myocardial infarction) 발생 여부에 영향을 미치는지 알아보기 위해 11,037명에게 아스피린을 복용 (325 mg/day)하게 하고 11,034명에게는 위약 (placebo)을 복용하게 한후 60개월 동안 추적 관찰하였더니 다음과 같은 결과를 얻었다.

	심근경색 발생함	심근경색 발생 안함	합
아스피린 복용	139	10,898	11,037
위약 복용	239	10,795	11,034
합	378	21,693	22,071

위 자료에 대한 상대위험도는

$$RR = \frac{139/11{,}037}{239/11{,}034} = .581$$

로서 아스피린의 복용이 심근경색 발생율을 42% 정도 감소시켜준다고 해석할 수 있다. 또한, 95% 신뢰구간은 $(0.473, 0.715)$로 주어지며 p-value 역시 0.05보다 작은 값으로 주어진다.

(2) 오즈 비 (odds ratio)

오즈 비(OR)는 관측연구에서 후향적 연구 (사례-대조 연구)에 사용되는 측도로서 사례군과 대조군의 오즈 비를 나타낸다. 여기서 오즈 (odds)란 질병에 걸릴 확률과 질병에 걸리지 않을 확률의 비를 나타낸다. 따라서, 오즈 비는 사례군의 오즈와 대조군의 오즈의 비를 나타내며 다음과 같이 정의된다.

$$OR = \frac{(n_{11}/n_{1.})/(n_{12}/n_{1.})}{(n_{21}/n_{2.})/(n_{22}/n_{2.})} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

으로 나타낼 수 있으며 만약 OR이 1보다 크면 사례군의 오즈 비가 대조군의 오즈 비보다 크다는 것이고 1보다 작으면 그 반대로 해석한다. 또한, OR에 대한 95% 신뢰구간 (L, U)이 1 보다 크거나 (즉, $L > 1$) 1보다 작으면 (즉, $U < 1$) p-value는 0.05보다 작아진다.

(예) 약물남용 여부가 심작발작에 미치는 영향을 관찰하였더니 다음과 같은 결과를 얻었다.

	약물남용 함	약물남용 안함	합
심장발작	73	141	214
심장발작 안함	18	196	214

	약물남용 함	약물남용 안함	합
합	91	337	428

위 자료에 대한 오즈 비는

$OR = \frac{73x196}{18x141} = 5.64$

로서 약물남용이 심장발작을 일으킬 위험이 약물남용을 하지 않는 집단에 비해 5.64베 높다고 해석할 수 있다. 또한, 95% 신뢰구간은 $(3.222, 9.863)$로 주어지며 p-value 역시 0.05보다 작은 값으로 주어진다.

2×2 분할표에서 두 변수의 연관성을 나타내는 측도로 상대 위험도와 오즈 비를 소개하였는데 두 변수 간의 연관성을 검정하는 방법으로 이미 소개한 카이제곱 검정, 피셔의 정확검정 (Fisher's exact test) 등이 있다.

7.5 시험 대상자 수 계산

임상시험의 단계중 신약의 효과를 검정하기 위한 제3상 임상시험에서 시험 대상자를 미리 정하는 문제는 매우 중요하다.이는 임상시험의 제1차 목적에 사용되는 검정통계량에 근거하되 원하는 마진과 검정력이 충족될 수 있도록 설정된다. 임상 대상자 수를 결정하는 문제는 임상시험의 최종 목적에 따라 다르므로 그에 맞는 계산식을 유도하여야 하며 여기서는 가장 일반적인 경우 한 가지만 소개한다.

$X_1, \cdots, X_{n_1}$ 은 평균 μ_1, 분산 σ_1^2을 갖는 모집단 1로 부터의 표본으로 신약의 치료효과라 하고, $Y_1, \cdots, Y_{n_2}$ 는 평균 μ_2, 분산 σ_2^2을 갖는 모집단 2로 부터의 표본으로 기존 약의 치료효과라 하자. 두 모집단은 서로 독립이라는 가정하에 우월성 검정

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

을 위한 시험자 대상수를 계산해 보자. 유의수준 α (흔히 0.05로 주어짐)로 주어졌을 때 $\delta = \mu_1 - \mu_2$에서의 검정력을 $1-\beta$라고 하자. 검정력은 흔히 0.80, 즉 $\beta = 0.2$로 주어지는 경우가 많다. 이제 두 모집단이 독립이므로 두 집단의 표본평균 차이인 $\bar{X} - \bar{Y}$의 분산은

$\sigma_\delta^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$가 되므로 귀무가설 하에서

$$\alpha = P(\frac{\bar{X} - \bar{Y}}{\sigma_\delta} > z_\alpha)$$

이고, 대립가설 하에서

$$\beta = P(\frac{\bar{X} - \bar{Y}}{\sigma_\delta} < z_\alpha) = P(\frac{\bar{X} - \bar{Y} - \delta}{\sigma_\delta} < z_\alpha - \frac{\delta}{\sigma_\delta})$$

로 주어진다. 따라서,

$$z_\alpha + z_\beta = \frac{\delta}{\sigma_\delta}$$

이 성립함을 보일 수 있고 만약 $n_1 = rn_2$이면

$$n_2 = (\frac{\sigma_1^2}{r} + \sigma_2^2)\ \frac{(z_\alpha + z_\beta)^2}{\delta^2}$$

으로 주어진다. 단, 여기서 r은 배분율 (allocation rate)을 나타낸다.

예제 : 두 집단의 평균차이가 $\delta = 1$이고, 표준편차가 $\sigma_1 = \sigma_2 = 1$이라 가정하자. (1) 각 집단의 표본크기가 $n_1 = n_2 = 10$일때 유의수준 $\alpha = 0.05$하에서 검정력(power)을 계산하자. (2) 검정력 80%일때 요구되는 표본크기를 계산하자.

```
## (1) calc power
power.t.test(delta=1, sd=1, n=10, sig.level=0.05)
#>
#>      Two-sample t test power calculation
#>
#>               n = 10
#>           delta = 1
#>              sd = 1
#>       sig.level = 0.05
#>           power = 0.5619846
#>     alternative = two.sided
```

```
#>
#> NOTE: n is number in *each* group

## (2) calc sample size
power.t.test(delta=1, sd=1, power=0.8, sig.level=0.05)
#>
#>      Two-sample t test power calculation
#>
#>               n = 16.71477
#>           delta = 1
#>              sd = 1
#>       sig.level = 0.05
#>           power = 0.8
#>     alternative = two.sided
#>
#> NOTE: n is number in *each* group
```

예제 : 두 집단의 비율이 $p_1 = 0.5$, $p_2 = 0.75$라 가정하자. (1) 각 집단의 표본크기를 $n = 10$으로 할때 유의수준 $\alpha = 0.05$하에서 검정력(power)을 계산하자. (2) 검정력 80%일때 요구되는 표본크기를 계산하자.

```
## (1) calc power
power.prop.test(p1=0.5, p2=0.75, n=10, sig.level=0.05)
#>
#>      Two-sample comparison of proportions power calculation
#>
#>               n = 10
#>              p1 = 0.5
#>              p2 = 0.75
#>       sig.level = 0.05
#>           power = 0.2022738
#>     alternative = two.sided
#>
#> NOTE: n is number in *each* group

## (2) calc sample size
power.prop.test(p1=0.5, p2=0.75, power=0.8, sig.level=0.05)
#>
#>      Two-sample comparison of proportions power calculation
#>
#>               n = 57.67344
#>              p1 = 0.5
#>              p2 = 0.75
#>       sig.level = 0.05
#>           power = 0.8
#>     alternative = two.sided
#>
#> NOTE: n is number in *each* group
```