R을 이용한 기초 통계 분석

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-04-10

Contents

1 기본 개념

1.1 통계학이란

1. 데이터(자료)에 근거하여 자연 혹은 사회적 현상에 대한 과학적 추론과 합리적인 의사결정을 하도록 하는 학문
2. 통계학은
    1. 관심의 대상에 대한 자료를 수집하고
    2. 수집된 자료를 정리, 요약하며
    3. 주어진 자료를 토대로 불확실한 사실에 대한 과학적인 판단을 하도록 하는 제반 방법들

1.2 기본 용어

1. 모집단 (population): 관심의 대상이 되는 모든 개체들의 관측값의 집합
    - 유한모집단: 모집단의 원소의 개수가 유한개인 경우
    - 무한모집단: 모집단의 원소의 개수가 무한개인 경우
2. 표본 (sample): 모집단에서 실제 조사한 관측값의 집합
    - 랜덤표본 (random sample)

1.3 기술통계학과 추측통계학

- 기술통계학: 수집된 자료들의 특성을 파악하기 쉽도록 단순히 정리하거나 요약하는 분야
  - 표본 평균, 표본 분산 등
  - 도수 분포표, 분할표, 상자그림 등 …
- 추측통계학: 표본에 내포되어 있는 정보를 이용하여 모집단의 특성을 파악, 추론하는 분야
  - 추정과 검정
  - 예측

1.4 자료의 종류

- 질적 자료 (범주형 자료) : 명목형 또는 순서형 자료
  - 명목형 (nominal): 성별 (이진형; binary), 질환명 (multinomial)
  - 순서형 (ordinal): 5점척도(best, good, normal, bad, worst), …
- 양적 자료
  - 실수형 (real-valued)
  - 정수형 (count)
  - 비율형 (ratio, rate)

2 R을 이용한 기술통계

2.1 예제 데이터

- 통계학 수강인원 25명에 대한 성적이 다음과 같다.

```
x <- c(
  64, 84, 82, 81, 68, 85, 76, 89, 93,
  77, 66, 64, 86, 74, 64, 70, 53, 98,
  59, 79, 57, 59, 65, 67, 80)
print(x)
```

```
##  [1] 64 84 82 81 68 85 76 89 93 77 66 64 86 74 64 70 53 98 59 79 57 59 65
## [24] 67 80
```

2.2 위치를 나타내는 기술통계

- 평균(mean): 자료의 산술평균
- 중앙값(median): 자료를 크기 순서로 나열하여 찾은 중간위치의 수
- 최솟값/최대값/사분위수/백분위수

2.3 R 예제

```
mean(x)
```

```
## [1] 73.6
```

```
median(x)
```

```
## [1] 74
```

```
range(x)
```

```
## [1] 53 98
```

```
min(x);  max(x)
```

```
## [1] 53
```

```
## [1] 98
```

```r
quantile(x)
```

```
##   0%  25%  50%  75% 100%
##   53   64   74   82   98
```

```r
quantile(x, probs=c(0.05, 0.95))
```

```
##   5%  95%
## 57.4 92.2
```

2.4 산포를 나타내는 기술통계

- 범위(range): 최대값-최소값

- 사분위수 범위(IQR; interquartile range)
- 분산(variance), 표준편차(standard deviation)
- 변이계수(CV; coefficient of variation)

```r
range(x) # max(x) - min(x)
```

```
## [1] 53 98
```

```r
IQR(x)
```

```
## [1] 18
```

```r
var(x)
```

```
## [1] 143.1667
```

```r
sd(x)
```

```
## [1] 11.96523
```

```r
sd(x)/mean(x) # 변이계수
```

```
## [1] 0.162571
```

2.5 Boxplot (상자그림)

 상자그림

1. read data and ordering: $X_{(1)}, \ldots, X_{(n)}$
2. Q_1, Q_2, Q_3 계산
3. Box 그리기: Q_1, Q_3
4. IQR(interquartile range; 4분위수범위)의 계산: $IQR = Q_3 - Q_1$
5. Box밖 수염 (whisker) 그리기: 상자의 아래 혹은 위로 다음의 값까지를 선으로 연결함

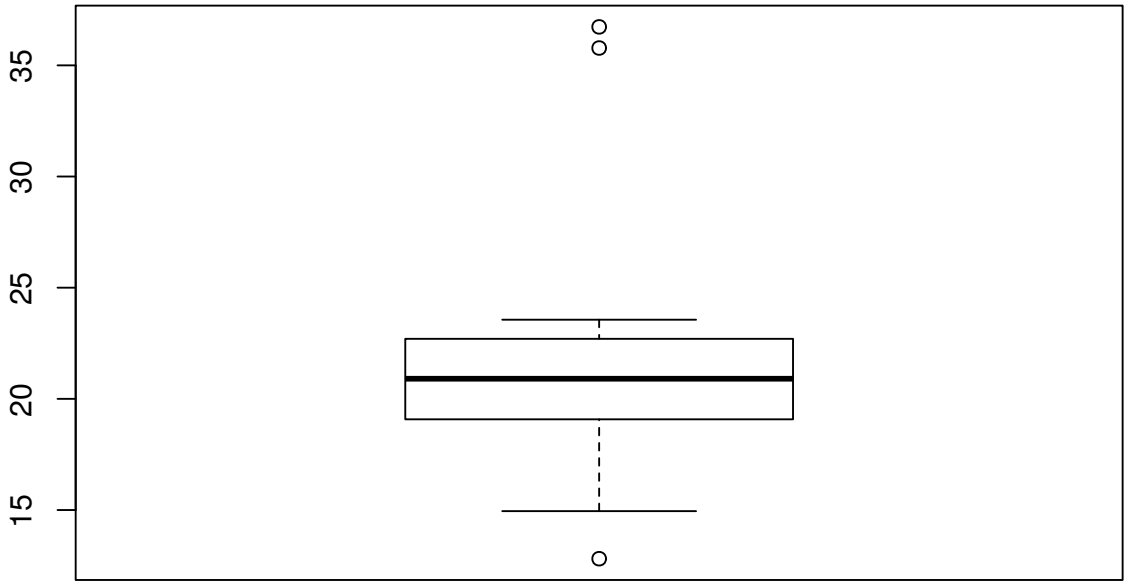
$$w_1 = Q_1 - 1.5 * IQR, \; w_2 = Q_3 + 1.5 * IQR$$

6. 이상치 판정: $X < w_1$ or $X > w_2$

2.6 R 예제 - 상자그림

 남성 20명의 폐활량 자료, 마지막 자료 35.78이 이상값?

```
x <- c(12.81, 14.95, 15.83, 15.97, 19.90,
       18.34, 19.82, 19.94, 20.62, 36.73,
       20.88, 20.93, 20.98, 21.15, 22.24,
       23.16, 22.24, 23.16, 23.56, 35.78)
boxplot(x)
```



2.7 범주형 자료의 빈도표/분할표

- 예제: 설문조사 data
  - University of Adelaide에서 237명의 학생을 대상으로 설문조사한 결과
  - 변수 및 속성

변수(요인)	설명	수준
Sex	성별	“Male”, “Female”
Exer	운동의 빈도	“Freq”, “Some”, “None”

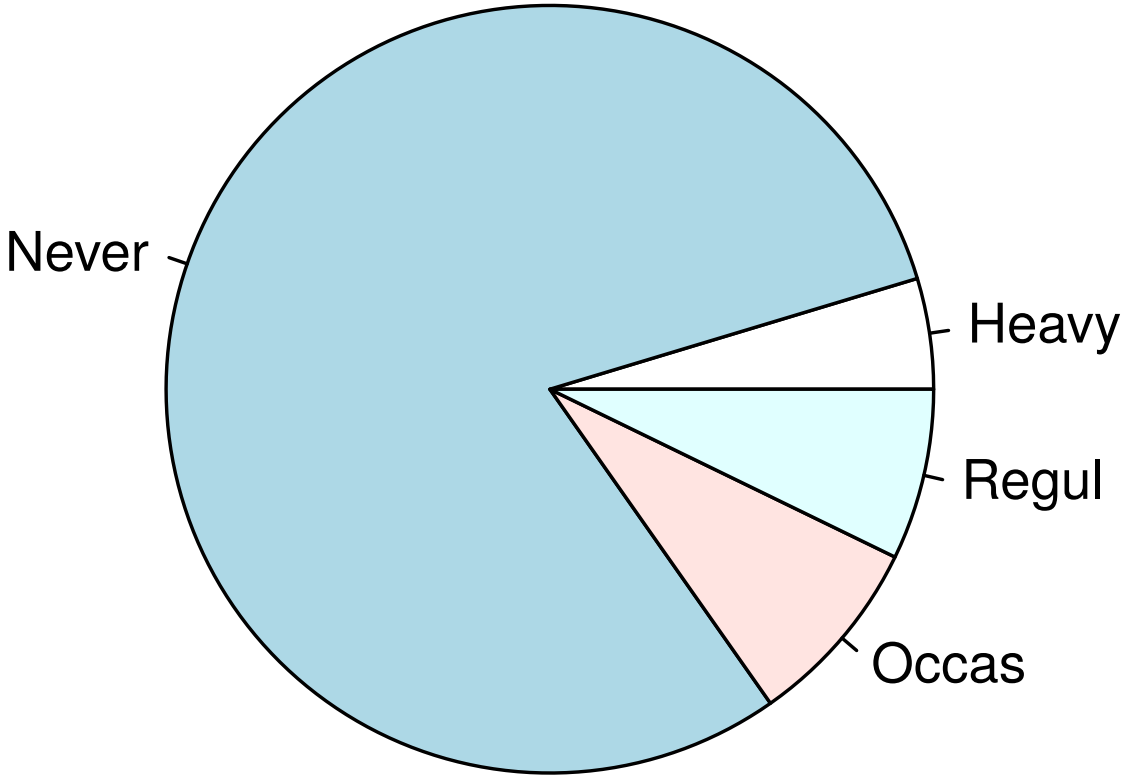
변수(요인)	설명	수준
Smoke	흡연의 정도	“Heavy”, “Regul”, “Occas”, “Never”

2.8 R 예제 - 범주형 자료의 빈도표/분할표

```
library(MASS)
(smoke <- table(survey$Smoke))
```

```
##
## Heavy Never Occas Regul
##    11   189    19    17
```

```
pie(smoke)
```




```
(SS <- table(survey$Sex, survey$Smoke))
```

```
##
##           Heavy Never Occas Regul
##    Female     5    99     9     5
##    Male       6    89    10    12
```

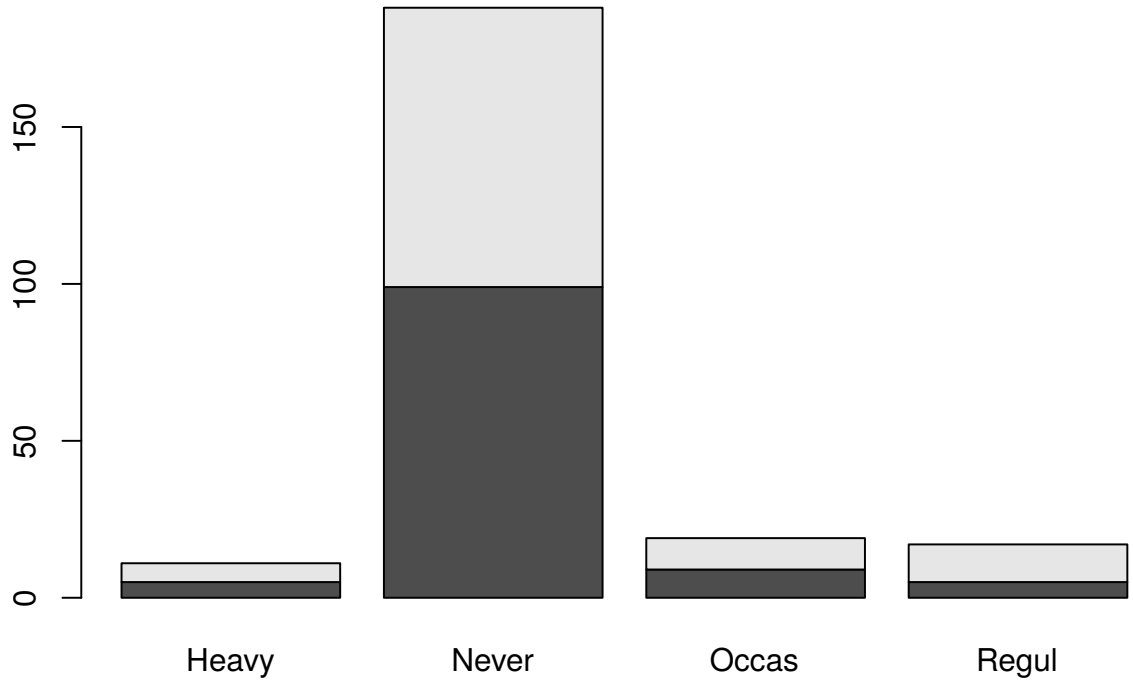
```r
prop.table(SS)
```

```
##
##                Heavy      Never      Occas      Regul
##   Female 0.02127660 0.42127660 0.03829787 0.02127660
##   Male   0.02553191 0.37872340 0.04255319 0.05106383
```

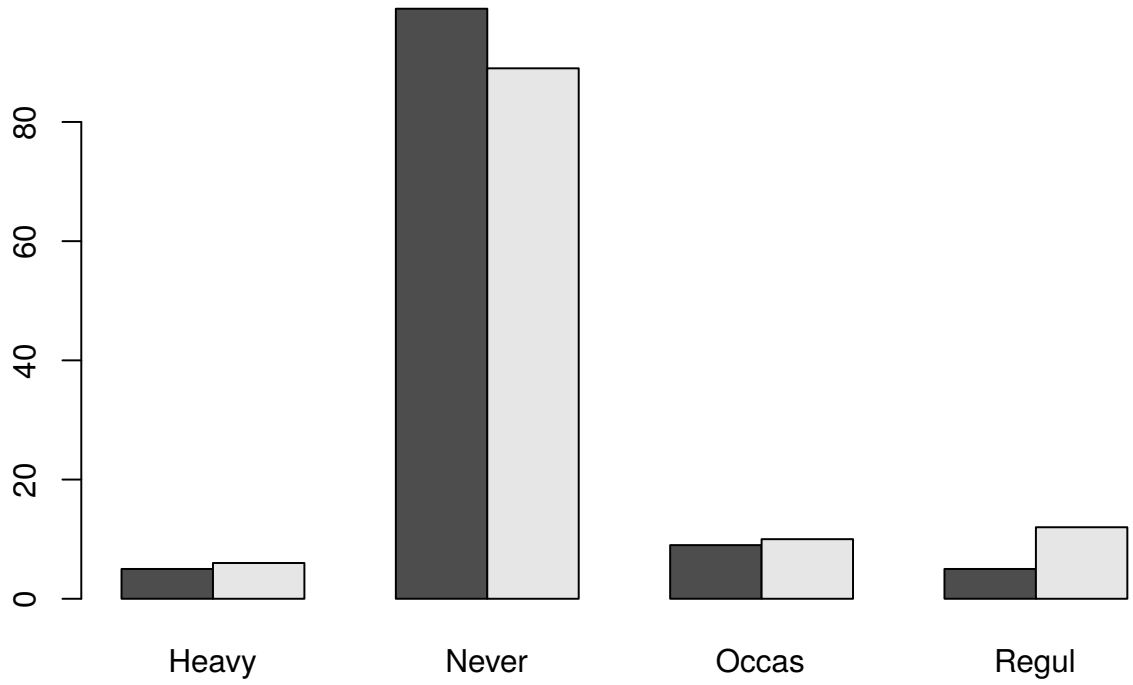
```r
ftable(Smoke~Sex, data=survey)
```

```
##        Smoke Heavy Never Occas Regul
## Sex
## Female           5    99     9     5
## Male             6    89    10    12
```

```r
barplot(SS)
```




```
barplot(SS, beside =T)
```


0
20
40
60
80
Heavy
Never
Occas
Regul

3 확률 및 확률 분포 (probability and probability distribution)

- 확률은 모집단에서 표본을 추출할 때, 표본을 바탕으로 모집단에 대한 결론을 이끌어내는 데 논리적 근거임
  - 표본공간(sample space, S): 통계적 조사 혹은 실험에서 얻을 수 있는 모든 가능한 결과들의 전체집합
  - 사건(event, A): 표본공간의 부분집합
- 확률 (probability) : 어떤 사건이 일어날 가능성을 0과 1사이의 수로 대응시킨 관계로 다음의 성질을 만족함
  1. 임의의 사건 A에 대하여, $0 \leq P(A) \leq 1$
  2. $P(\emptyset) = 0, P(S) = 1$
  3. 서로 배반인 사건열 $\{A_i\}$ 에 대하여 $P(\cup_{i=1}^{\infty} A_i) = \Sigma_{i=1}^{\infty} P(A_i)$.
- 확률변수(random variable): 표본공간 S 에서 정의된 실함수, $X : S \to R$.
- 확률분포(probability distribution): 표본공간에서 정의된 확률 P를 확률변수 X의 값에 따라 재표현한 것으로 X의 확률분포라고 함

$$\text{For all } A \subset S, X(A) = \{X(\omega) : \omega \in A\} \equiv B,$$

$$\text{then } P(A) = \Pr(B) = P(X^{-1}(A)) = P(X \in B).$$

3.1 확률분포의 표현

확률(분포)은 원래 집합함수로 정의됨, 이를 쉽게 표현하는 방법

X 의 누적분포함수 (cumulative distribution function:c.d.f.):

$$F : R \to [0, 1], \text{ that is,}$$

$$F(x) = P(X \in (-\infty, x]) \equiv P(X \leq x).$$

3.2 확률질량함수, 확률밀도함수

- 확률 분포함수가 $(-\infty, x]$ 구간에서의 확률인 반면 확률질량함수(probability mass function:p.m.f.)는 하나의 점 (point)에서의 확률를 표현한 함수

$$p(x) = P(X = x)$$

- 확률 분포함수가 미분 가능이면, 그 도함수를 확률밀도함수(probability density function: pdf)

$$f(x) = \frac{dF(x)}{dx}$$

3.3 기대값 (Expectation)

- 확률변수 x의 기대값 ($f(x)$는 확률밀도함수)

$$E(X^r) = \int x^r f(x) dx$$

- 주어진(관측된) 자료를 이용한 기대값의 근사

$$E(X^r) \simeq \frac{1}{n} \sum_{i=1}^{n} x_i^r.$$

  - 표본 평균: $r = 1$인 경우

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

  - 표본분산 : $r = 2$인 경우

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right\} - \frac{n}{n-1} \bar{X}^2.$$

3.4 확률 분포의 종류

- 이산확률분포 (discrete distribution)
  - 베르누이 (Bernoulli)

- 이항분포 (Binomial), 다항분포 (Multinomial)
- 포아송 (Poisson)
- ...

- 연속확률분포 (continuous distribution)
  - 정규분포 (Normal, Gaussian)
  - 지수분포 (Exponential)
  - 감마분포 (Gamma)
  - 베타분포 (Beta)
  - ...

4 통계적 추정

- 통계적 추론: 연구자가 관심을 가지는 모집단의 특성치 (모수, parameter)에 대한 판단을 하기 위하여, 수집된 데이터를 기초로 통계 이론을 이용한 일련의 과정

  - 추정 (estimation) :

    1. 점추정 : 관측된 표본을 이용하여 모수 (θ)를 하나의 값으로 추정
    2. 구간추정: 관측된 표본을 이용하여 모수가 포함되리라고 예상되는 범위를 추정

    3. 신뢰수준 (confidence level): $1-\alpha$ 또는 $(1-\alpha)\times 100\%$

$$P(L(X) \leq \theta \leq U(X)) = 1-\alpha$$

  - 가설검정 (hypothesis test) : 모수 또는 모집단 분포에 대한 가정 (가설)을 세우고, 표본을 기초로 가정의 참 거짓을 판단하는 방법

    1. 대립가설(alternative hypothesis): 통상적으로 연구자가 입증하려는 가설로, 표본을 토대로 확실한 근거가 있어야 받아들임

    2. 귀무가설(null hypothesis): 대립가설과 상반되는 가설
    3. 보통 귀무가설을 H_0, 대립가설을 H_1으로 표시함

    4. 검정통계량 (test statistic): 가설검정을 하기 위해 이용하는 통계량

5. 유의수준 (significance level):
6. 기각역 (critical region): 귀무가설을 기각시키기 위한 검정통계량 관측값의 영역
7. P값 (유의확률; P-value): 귀무가설이 참이라는 전제하에서 검정통계량이 관측값을 벗어날 확률

$$P(T > t|H_0\, true)$$

4.1 일표본 (단변량)에서 모평균 추정

- 랜덤표본 : $X_1, X_2, \ldots, X_n \sim \ iid\, N(\mu, \sigma^2)$,
- 모평균 (μ)의 추정

$$\hat{\mu} = \frac{\Sigma_{i=1}^{n} X_i}{n} = \bar{X}$$

- 모분산 (σ^2)의 추정

$$\hat{\sigma^2} = \frac{\Sigma_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = S^2$$

- 모평균 (μ)의 구간추정 (신뢰구간, confidence Interval)
  - 신뢰수준 (confidence level) $1-\alpha$
  - 모분산이 알려진 경우 (σ^2), 또는 표본의 수가 많은 경우 ($\hat{\sigma}^2$)

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

  - 모분산을 모르는 경우 ($\hat{\sigma}^2$)

$$(\bar{X} - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}})$$

  - $z_{\alpha/2}$: 표준정규분포에서 $\alpha/2$-백분위수
  - $t_{\alpha/2}$: 자유도가 $n-1$인 t분포의 $\alpha/2$-백분위수

4.2 R 예제: 모평균 추정

- 수면제의 효과 측정 데이터 (sleep) : 10명의 환자에게 수면제를 투여 후, 초과 수면시간 측정

- Objectives: 모평균의 점/구간 추정

```
x <- sleep[1:10, c(3,1)]
head(x)
```

```
##   ID extra
## 1  1   0.7
## 2  2  -1.6
## 3  3  -0.2
## 4  4  -1.2
## 5  5  -0.1
## 6  6   3.4
```

```
#install.packages("Rmisc")
library(Rmisc)
c1 <- CI(x[,2], ci=0.95)[c(3,1)]
c2 <- CI(x[,2], ci=0.99)[c(3,1)]
```

```r
o <- rbind(c1, c2); row.names(o)<-NULL
o <- data.frame("level"=c(0.95, 0.99), o, mean=c(mean(x[,2]), NA), se=c(sd(x[,2]/sqrt(10)), NA
o
```

```
##   level      lower    upper mean        se
## 1  0.95 -0.5297804 2.029780 0.75 0.5657345
## 2  0.99 -1.0885442 2.588544   NA        NA
```

4.3 일표본 (단변량)에서 모비율 추정

- 랜덤표본 : $X_1, X_2, \ldots, X_n \sim \ iid\, Bernoulli(p),$
- $\Sigma_{i=1}^{n} X_i = X_1 + \ldots + X_n$ 의 분포는 (n, p)를 모수로 하는 이항분포를 따름 ($Bin(n, p)$)
- 표본비율: 모비율 (p)의 추정

$$\hat{p} = \frac{\Sigma_{i=1}^{n} X_i}{n} = \bar{X}$$

- 표본비율의 정규근사 (중심극한정리): $np \geq 5, n(1-p) \geq 5$

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

- 모비율의 구간추정 (신뢰수준: $1 - \alpha$)

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

4.4 R 예제: 모비율 추정

- 대학생의 흡연율을 파악하기 위해 100명을 랜덤 추출, 흡연여부를 조사하여 아래의 결과를 얻었다.

흡연	비흡연	합계
30	70	100

- Objectives: 모비율의 점추정치 및 95% 신뢰구간

```
x <- 30; n <- 100; p <- x/n;
se <- sqrt(p*(1-p)/n)

z_alpha <- qnorm(0.975)
lb <- p - z_alpha * se
ub <- p + z_alpha * se

o <- t(c(p, se, lb, ub))
colnames(o)<- c("mean", "se", "lower", "upper")
o
```

```
##      mean         se     lower     upper
```

```
## [1,]  0.3 0.04582576 0.2101832 0.3898168
```

5 가설검정

5.1 가설검정에서 오류

	H_0 is TRUE	H_0 is FALSE
Reject H_0	Type I Error	True Positive
Accept H_0	True Negative	Type II error

- Type I Error: 귀무가설이 참인데도 불구하고 귀무가설을 기각할 오류
- 유의수준 (significance level): Type I Error를 범할 확률의 최대값

$$\alpha = P(Reject\ H_0 | H_0\ is\ TRUE)$$

- Type II Error: 귀무가설이 거짓인데도 불구하고 귀무가설을 채택하는 오류

- 검정력 (Power) : 귀무가설이 거짓일 때, 귀무가설을 기각할 확률

$$1 - \beta = P(Reject\ H_0 | H_0\ is\ FALSE)$$

5.2 t검정

1. 일표본 t검정
2. 이표본 t검정
3. 대응비교

5.3 일표본 t검정

- 표본자료: $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- 모평균이 μ_0 (특정값)인지를 관측된 표본을 이용하여 검증
- 가설: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- 검정통계량 :

1. 모집단 분산(σ^2)이 알려진 경우

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$:표준정규분포

2. 모집단 분산을 모르는 경우: 자유도(degree of freedom)가 $n-1$인 t분포

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1),$$

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

- 기각역: 유의수준 α인 양측검정에서

$$|T| > t_{\alpha/2,n-1}$$

- p-value(유의 확률): 검정통계량의 관측값이 t_0일 때,

$$p - value = \Pr(|T| > t_0)$$

5.4 R 예제: 일표본 t검정

- 수면제의 효과 측정 데이터 (sleep) : 10명의 환자에게 수면제를 투여 후, 초과 수면시간 측정

- 가설 : 초과 수면시간(μ)이 0보다 큰가? ($H_0 : \mu = 0$ vs, $H_1 : \mu > 0$)

```
x <- sleep[1:10, c(3,1)]
head(x)
```

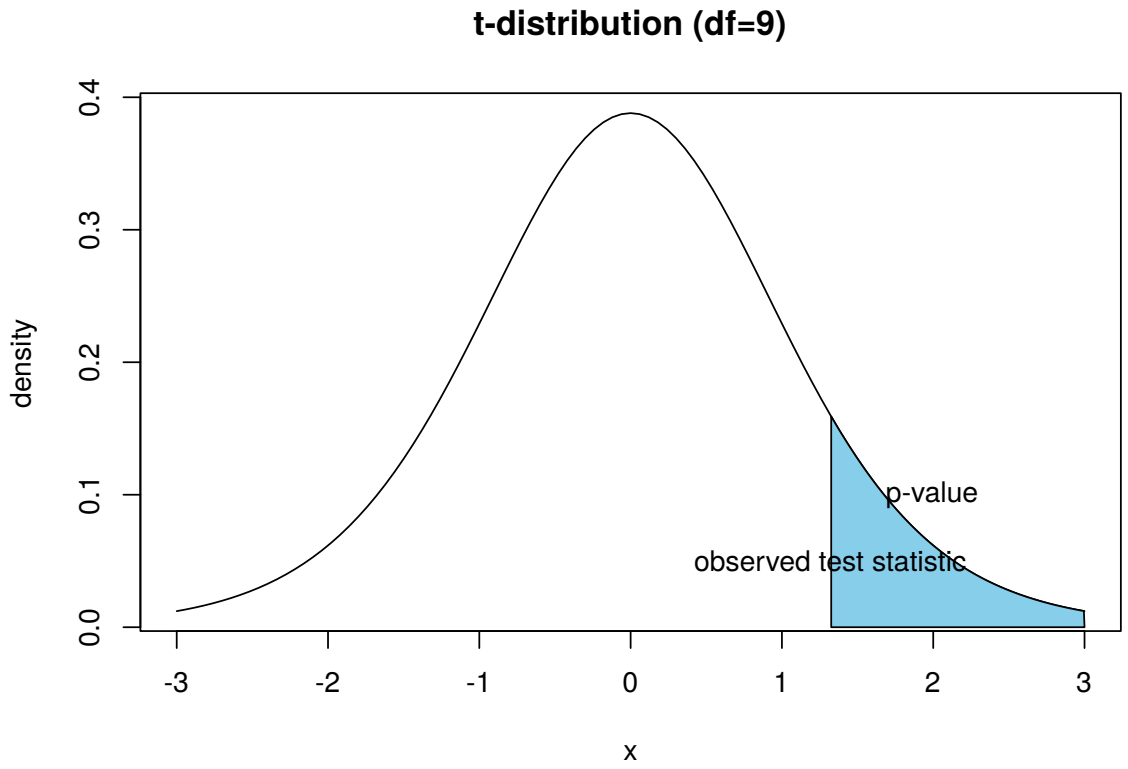
```
##   ID extra
## 1  1   0.7
## 2  2  -1.6
## 3  3  -0.2
## 4  4  -1.2
## 5  5  -0.1
## 6  6   3.4
```

- alternative="greater": $H_1 : \mu > 0$

```r
o <- t.test(x=x$extra, alternative="greater")
o
```

```
##
##  One Sample t-test
##
## data:  x$extra
## t = 1.3257, df = 9, p-value = 0.1088
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2870553        Inf
## sample estimates:
## mean of x
##      0.75
```

```r
cord.x <- c(o$statistic, seq(o$statistic, 3, 0.01), 3)
cord.y <- c(0, dt(seq(o$statistic, 3, 0.01), 9), 0)
curve(dt(x, 9), xlim=c(-3,3), ylab="density", main='t-distribution (df=9)')
polygon(cord.x,cord.y,col='skyblue')
text(o$statistic, 0.05, "observed test statistic")
text(2, 0.1, "p-value")
```


t-distribution (df=9)
density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x
p-value
observed test statistic

5.5 이표본(독립표본) t검정 (Two Sample t-test) : 두 모집단의 평균의 차이 검정

- 표본자료

$$\begin{aligned} X_1, \ldots, X_m &\sim N(\mu_1, \sigma_1^2) \\ Y_1, \ldots, Y_n &\sim N(\mu_2, \sigma_2^2) \end{aligned}$$

- 가설 : $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

- 검정통계량:

$$T = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}, \text{ where } s_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

- 검정통계량의 분포

1. 분산이 서로 같은 경우 : $t(n+m-2)$
2. 다른 경우: Welch t 검정이며 자유도는

$$df = \frac{(s_x^2/n + s_y^2/m)^2}{(s_x^2/m)^2/(m-1) + (s_y^2/n)^2/(n-1)}$$

5.6 R 예제: 독립표본 t검정

- 수면제의 효과 측정 데이터 (sleep) : 수면제 A, B를 각 10명의 환자에게 투여 후, 초과 수면시간 측정

- 가설 : A의 초과 수면시간(μ_1) 보다 B(μ_2)가 큰가? ($H_0 : \mu_1 = \mu_2$ vs, $H_1 : \mu_1 < \mu_2$)

```
A <- sleep[1:10, 1]
B <- sleep[11:20, 1]
x <- data.frame(n=1:10, A, B)
x
```

```
##     n    A    B
## 1   1  0.7  1.9
## 2   2 -1.6  0.8
## 3   3 -0.2  1.1
## 4   4 -1.2  0.1
## 5   5 -0.1 -0.1
## 6   6  3.4  4.4
## 7   7  3.7  5.5
## 8   8  0.8  1.6
## 9   9  0.0  4.6
## 10 10  2.0  3.4
```

 이분산: var.equal=FALSE

```
t.test(x=x$A, y=x$B, paired=F, alternative="less", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x$A and x$B
## t = -1.8608, df = 17.776, p-value = 0.0397
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1066185
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

 등분산: var.equal=TRUE

```
t.test(x=x$A, y=x$B, paired=F, alternative="less", var.equal=TRUE)
```

```
##
```

```
##  Two Sample t-test
##
## data:  x$A and x$B
## t = -1.8608, df = 18, p-value = 0.03959
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1076222
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

5.7 등분산 검정

- 가설 : $H_0 : \sigma_1^2/sigma_2^2 = 1$ vs. H_1: not H_0
- 검정통계량 : $F = s_x^2/s_y^2 \sim F(m-1, n-1)$

```
var.test(A, B)
```

```
##
##  F test to compare two variances
##
## data:  A and B
## F = 0.79834, num df = 9, denom df = 9, p-value = 0.7427
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.198297 3.214123
## sample estimates:
## ratio of variances
##          0.7983426
```

- 분산이 다르지 않음 : 등분산 가정의 t검정 결과를 이용함

5.8 대응비교(짝비교, Paired t-test)

- 동일한 대상에 대하여 서로 다른 처리를 한 후 처리 효과의 차이를 비교할 때
- 표본자료:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim N\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \Sigma\right)$$

Then

$$D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n \sim N(\mu_1 - \mu_2, \sigma_D^2)$$

- 검정통계량:

$$T = \frac{\bar{D}}{s_D/\sqrt{n}} \sim t(n-1), s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

5.9 R 예제: 대응비교 t검정

- 운동선수 트레이닝 방법 : 100m 단거리 육상선수 10명에게 새로운 훈련방법을 도입하여 전/후 효과를 측정
- 가설 : 훈련 후 (μ_2)가 훈련 전 (μ_1)보다 효과가 있는가? ($H_0 : \mu_1 = \mu_2$ vs, $H_1 : \mu_1 < \mu_2$)

```
pre = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
post = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
x <- data.frame(id=1:10, pre, post, d=post-pre)
x
```

```
##    id  pre post    d
## 1   1 12.9 12.7 -0.2
## 2   2 13.5 13.6  0.1
## 3   3 12.8 12.0 -0.8
## 4   4 15.6 15.2 -0.4
## 5   5 17.2 16.8 -0.4
## 6   6 19.2 20.0  0.8
## 7   7 12.6 12.0 -0.6
## 8   8 15.3 15.9  0.6
## 9   9 14.4 16.0  1.6
## 10 10 11.3 11.1 -0.2
```

```r
t.test(post-pre, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  post - pre
## t = 0.21331, df = 9, p-value = 0.4179
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.3796859        Inf
## sample estimates:
## mean of x
##      0.05
```

5.10 비율검정 - one sample 비율검정

$$X_1, \ldots, X_n \sim \text{iid } Ber(p) \Rightarrow X = \sum X_i \sim Bin(n, p)$$

* Hyperthesis:

$$H_0 : p = p_0, \text{ vs } H_1 : \text{ not } H_0$$

* Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1) \text{ under } H_0$$

5.11 R 예제

- 가설 : 대학생의 흡연율이 40%보다 작은가?
  - $H_0 : p = p_0,$ vs $H_1 :$ not H_0
- 결과

```
x <- 30; n <- 100; hatp <- x/n; p0 <-0.4
z <- (hatp - p0)/sqrt(p0*(1-p0)/n)
prop.test(x, n=n, p=p0, correct=F, alternative="l")
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability p0
## X-squared = 4.1667, df = 1, p-value = 0.02061
## alternative hypothesis: true p is less than 0.4
## 95 percent confidence interval:
##  0.0000000 0.3798321
## sample estimates:
##   p
## 0.3
```

5.12 비율검정 (이표본 비교)

- Data

	group 1	group 2
success	X_1	X_2
trials	n_1	n_2
success probability	$\hat{p}_1 = X_1/n_1$	$\hat{p}_2 = X_2/n_2$

- Hyperthesis:

$$H_0 : p_1 = p_2, \text{ vs } H_1 : \text{ not } H_0$$

- Test statistic:

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$

5.13 R 예제

- 특정 바이러스에 대한 남, 여 항체보유율을 비교,

	남	여	합계
대상자	100	150	250
항체보유자	24	64	58

- 가설: $H_0 : p_1 = p_2$ vs $H_1 : p_1 < p_2$

```
x <- c(24, 64); n <- c(100, 150);
p <- x/n;
prop.test(x, n=n, correct=F, alternative="l")
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 9.1657, df = 1, p-value = 0.001233
## alternative hypothesis: less
## 95 percent confidence interval:
```

```
##  -1.000000 -0.089986
## sample estimates:
##    prop 1    prop 2
## 0.2400000 0.4266667
```

 여자의 항체보유율이 남자에 비해 높다고 판단

6 분산분석(ANOVA, Analysis of Variance)

6.1 분산분석(ANOVA, Analysis of Variance)

1. 여러(3이상) 모집단의 평균비교

6.2 기본 용어

1. 요인 (factor): 실험에 고려한 설명변수

    요인의 수가 하나이면 일원분산분석 (one-way ANOVA)

2. 수준 (level): 요인이 취하는 값
3. 처리 (treatment): 요인과 수준의 조합

6.3 통계 모형

		요인의 수준(처리)			
		1	2	⋯	p
반복수	1	Y_{11}	Y_{21}	⋯	Y_{p1}
	2	Y_{12}	Y_{22}	⋯	Y_{p2}
	⋮	⋮	⋮	⋮	⋮
		Y_{1n_1}	Y_{2n_2}	⋮	Y_{pn_p}

- 자료의 형태

- 모형

$$Y_{ij} = \mu + a_i + \epsilon_{ij}, i = 1, \ldots, p, j = 1, \ldots, n_i$$

  - μ : 전체 평균
  - a_i : i번째 처리효과, $\Sigma_{i=1}^{p} a_i = 0$

6.4 통계적 추론

- 처리효과(a_i)의 추정

- 처리효과(a_i)의 차이 검정: $H_0 : a_1 = \ldots = a_p = 0$

6.5 변동의 분해 및 분산분석표

- 분산분석표

	자유도	제곱합	평균제곱	F
처리	p-1	SStr	MStr	MStr/MSE
오차	n-p	SSE	MSE	
전체	n-1	SST		

- 세개이상의 모평균차에 대한 검정법: F-검정

$$F = \frac{\sum_{i=1}^{p} n_i (\overline{Y}_i - \overline{Y})^2 / (p-1)}{\sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 / \sum_{i=1}^{p} (n_i - 1)} \sim F_{p-1, N-p}$$

Figure 1: F test

6.6 R 예제: InsectSprays

- 6종의 살충제를 뿌리고 죽은 해충의 수를 조사

A	B	C	D	E	F
10	11	0	3	3	11
7	17	1	5	5	9
20	21	7	12	3	15
14	11	2	6	5	22
14	16	3	4	3	15
12	14	1	3	6	16
10	17	2	5	1	13
23	17	1	5	1	10
17	19	3	5	3	26

A	B	C	D	E	F
20	21	0	5	2	26
14	7	1	2	6	24
13	13	4	4	4	13

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

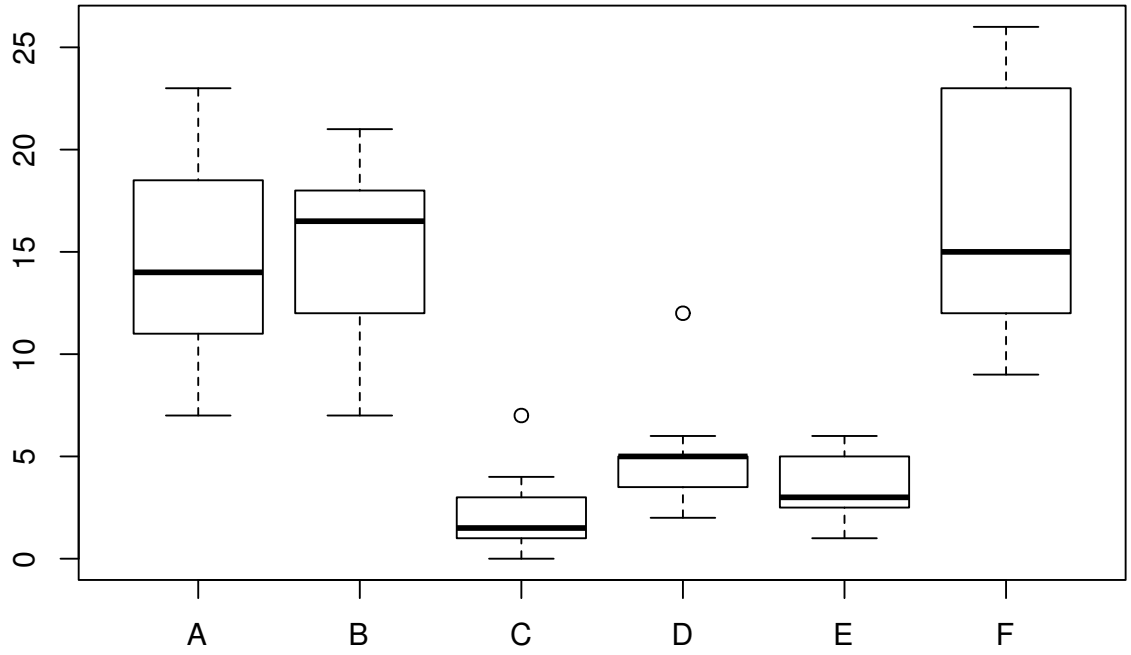
- Descriptive statistics: Mean, variance, number of elements in each cell

```
m <- tapply(count, spray, mean)
s <- tapply(count, spray, sd)
rbind(mean=m, sd=s)
```

```
##               A         B        C        D        E         F
## mean 14.500000 15.333333 2.083333 4.916667 3.500000 16.666667
## sd    4.719399  4.271115 1.975225 2.503028 1.732051  6.213378
```

- Visualise the data → boxplot; look at distribution, look for outliers

```
boxplot(count ~ spray)
```



- ANOVA table

```r
o <- aov(count ~ spray)
summary(o)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 가설검정 :
  - p-value < 0.001 : Reject $H_0 : a_1 = \ldots = a_p = 0$

6.7 효과의 추정

```r
o <- lm(count ~ spray)
summary(o)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  14.5000000   1.132156 12.8074279 1.470512e-19
## sprayB        0.8333333   1.601110  0.5204724 6.044761e-01
## sprayC      -12.4166667   1.601110 -7.7550382 7.266893e-11
## sprayD       -9.5833333   1.601110 -5.9854322 9.816910e-08
## sprayE      -11.0000000   1.601110 -6.8702352 2.753922e-09
## sprayF        2.1666667   1.601110  1.3532281 1.805998e-01
```

- Estmate for each effect: $\mu_j - \mu_A, j \in \{B, C, D, E, F\}$
  - $\mu_B - \mu_A = 0.83$
  - $\mu_C - \mu_A = -12.42$

6.8 Post Hoc tests (multiple comparison, 다중비교)

- Tukey HSD(Honestly Significant Difference)

```
o<-aov(count ~ spray)
TukeyHSD(o)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = count ~ spray)
##
## $spray
##             diff        lwr       upr     p adj
## B-A   0.8333333  -3.866075  5.532742 0.9951810
## C-A -12.4166667 -17.116075 -7.717258 0.0000000
## D-A  -9.5833333 -14.282742 -4.883925 0.0000014
## E-A -11.0000000 -15.699409 -6.300591 0.0000000
## F-A   2.1666667  -2.532742  6.866075 0.7542147
## C-B -13.2500000 -17.949409 -8.550591 0.0000000
## D-B -10.4166667 -15.116075 -5.717258 0.0000002
## E-B -11.8333333 -16.532742 -7.133925 0.0000000
```

```
## F-B   1.3333333  -3.366075  6.032742 0.9603075
## D-C   2.8333333  -1.866075  7.532742 0.4920707
## E-C   1.4166667  -3.282742  6.116075 0.9488669
## F-C  14.5833333   9.883925 19.282742 0.0000000
## E-D  -1.4166667  -6.116075  3.282742 0.9488669
## F-D  11.7500000   7.050591 16.449409 0.0000000
## F-E  13.1666667   8.467258 17.866075 0.0000000
```

 A-B-F 와 C-D-F에 대하여 그룹화 가능함 : 그룹화 후 재 분석 필요

X \ Y	b_1	b_2	...	b_c	계 (total)
a_1	O_{11}	O_{12}	...	O_{1c}	$O_{1\cdot}$
a_2	O_{21}	O_{22}	...	O_{2c}	$O_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
a_r	O_{r1}	O_{r2}	...	O_{rc}	$O_{r\cdot}$
계 (total)	$O_{\cdot 1}$	$O_{\cdot 2}$	...	$O_{\cdot c}$	n

Figure 2: 2원 분할표

7 범주형자료의 분석

- 질적자료 또는 범주화된 양적자료의 분석
- 분할표 (contingency table): 범주형 자료의 분석에 사용하는 테이블 형태의 자료
  - 열 또는 행은 요인(범주형 변수)의 수준
  - 셀은 요인의 각 수준에 해당되는 관측치의 빈도

7.1 범주형자료의 분석

(1) 적합도 검정(goodness of fit test) : 분할표의 값들이 특정 분포를 따르고 있는지를 검정

(2) 독립성 검정(test of independence) : 분할표에서 두 범주형 변수(요인)들이 서로 독립인지를 검정

- $H_0 : p_{ij} = p_i p_j, i = 1, \ldots, n, j = 1, \ldots, m$

(3) 동질성 검정(test of homogeneity) : 서로 다른 부모집단(subpopulation)에서 범주형 변수의 확률분포가 서로 동일한지를 검정

- $H_0 : p_{1j} = p_{2j} = \ldots p_{nj}, j = 1, \ldots, m$

- Test-statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi^2((r-1)(c-1))$$

7.2 R 예제: 설문조사 data (survey)

- University of Adelaide에서 237명의 학생을 대상으로 설문조사한 결과
- 변수 및 속성

변수	설명	변수값(범주형)
Sex	성별	“Male”, “Female”
Wr.Hnd	글쓰는 손의 한뼘의 길이 (단위: Cm)	
NW.Hnd	반대쪽 손의 한뼘 길이	
W.Hnd	글쓰는 손의 위치	“Left”, “Right”
Fold	팔을 접었을 때 양손의 위치	“R on L”, “L on R”, “Neither”
Pulse	혈압 (beats per minute)	
Clap	박수칠 때 위로 올라가는 손	“Right”, “Left”, “Neither”
Exer	운동의 빈도	“Freq”, “Some”, “None”
Smoke	흡연의 정도	“Heavy”, “Regul”, “Occas”, “Never”
Height	키 (Cm)	
Age	나이	

1. 독립성 검정

- 운동의 빈도(Exer)와 흡연정도(Smoke)에 대한 분할표 (contingency table)

```r
library(MASS)
(x <- table(survey$Exer, survey$Smoke))
```

```
##
##        Heavy Never Occas Regul
##   Freq     7    87    12     9
##   None     1    18     3     1
##   Some     3    84     4     7
```

- Pearson's Chi-squared test: 운동의 빈도(Exer)와 흡연정도(Smoke)가 서로 독립인가?

```r
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

2. 동일성 검정

- 성별(Sex)과 흡연정도(Smoke)에 대한 분할표 (contingency table)

```r
library(MASS)
(o <- table(survey$Sex, survey$Smoke))
```

```
##
##          Heavy Never Occas Regul
##   Female     5    99     9     5
##   Male       6    89    10    12
```

- Pearson's Chi-squared test: 남,여 그룹에서 흡연정도(Smoke)의 비율이 서로 동일한가?

```r
  chisq.test(o)
```

```
##
##  Pearson's Chi-squared test
##
## data:  o
## X-squared = 3.5536, df = 3, p-value = 0.3139
```

8 상관분석 (corelation analysis)

1. 공분산 : 두 연속형 변수간의 산포의 정도를 측정

$$Cov(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big] = E[XY] - E[X]E[Y]$$

2. 상관계수 (correlation coefficient) : 연속형 변수간의 선형적 관계의 정도

   - population

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

   - estimate

$$r = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}}$$

   - property & interpretation
     - $-1 \leq \rho \leq 1$
     - 절대값이 1에 가까울수록 강한 직선관계
     - 부호에 따라 양(음)의 상관관계

- Test for correlation: $H_0 : \rho = 0$
- Test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2) \text{ under } H_0 \text{ is true}$$

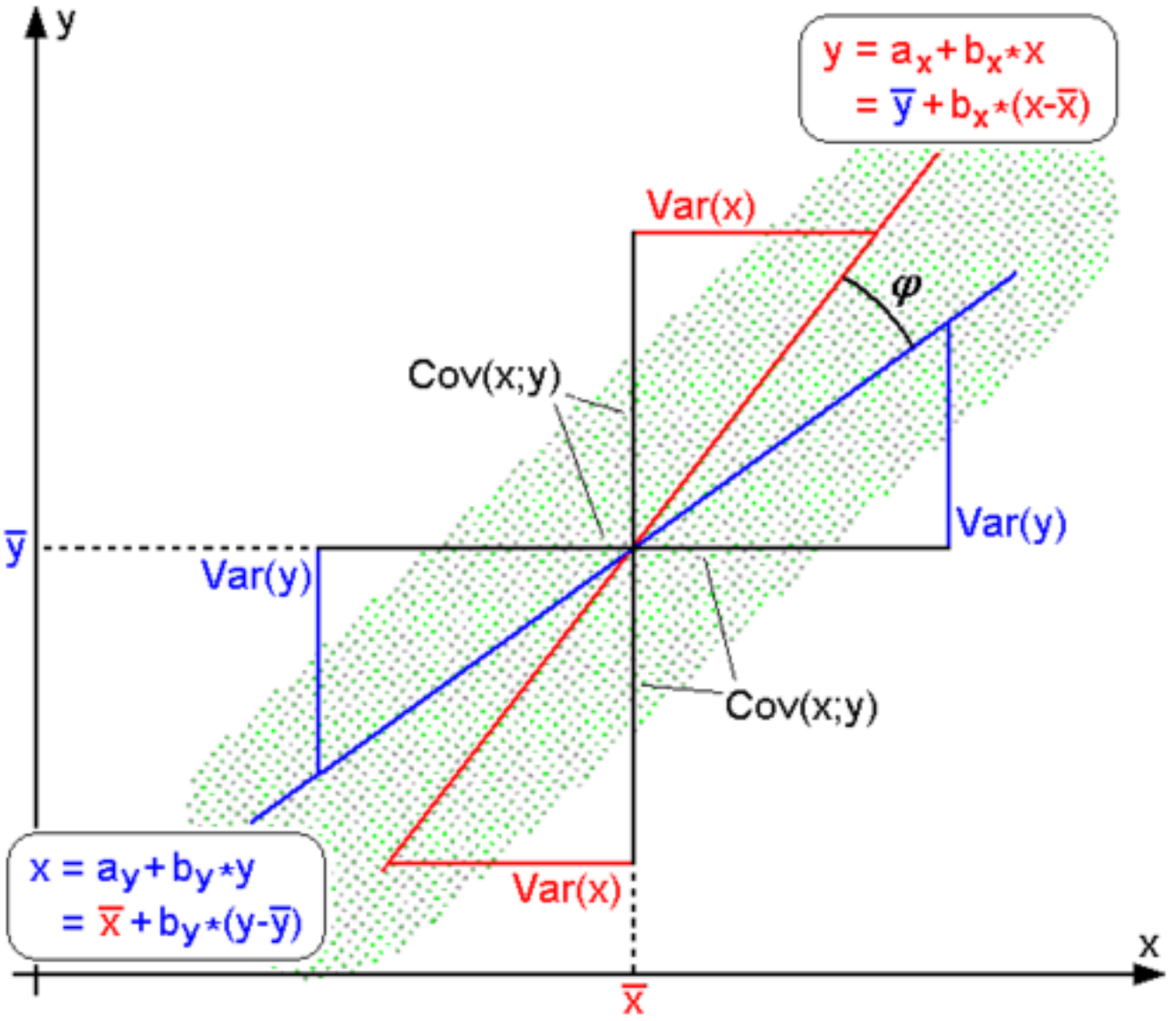


Figure 3: 기하학적 의미

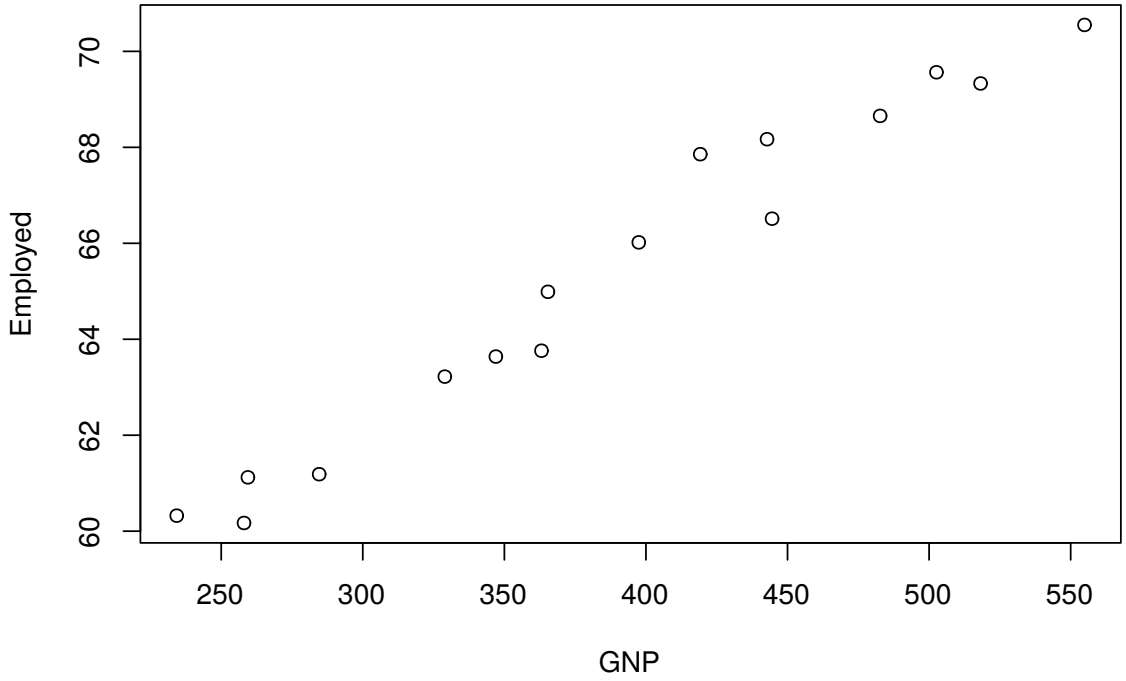
8.1 R 예제: 상관분석

- longley data : 7가지 경제지표 (1947~1962)
- 연도별 GNP와 취업률의 관계

```
x <- longley[, c("GNP", "Employed")]
x
```

```
##          GNP Employed
## 1947 234.289   60.323
## 1948 259.426   61.122
## 1949 258.054   60.171
## 1950 284.599   61.187
## 1951 328.975   63.221
## 1952 346.999   63.639
## 1953 365.385   64.989
## 1954 363.112   63.761
## 1955 397.469   66.019
## 1956 419.180   67.857
## 1957 442.769   68.169
## 1958 444.546   66.513
```

```
## 1959 482.704   68.655
## 1960 502.601   69.564
## 1961 518.173   69.331
## 1962 554.894   70.551
```



- covariance matrix

```r
cov(x)
```

```
##                GNP  Employed
## GNP      9879.3537 343.33021
## Employed  343.3302  12.33392
```

- correlation and test

```r
cor.test(x[,1], x[,2])
```

```
##
##  Pearson's product-moment correlation
##
## data:  x[, 1] and x[, 2]
## t = 20.374, df = 14, p-value = 8.363e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9519914 0.9944238
## sample estimates:
##       cor
## 0.9835516
```

9 선형회귀분석 (linear regression analysis)

- 확률변수간의 함수관계를 추정하는 통계분석 방법

1. 설명변수 (독립변수):
2. 반응변수 (종속변수):
3. 회귀계수

9.1 회귀모형

- $x_1, \ldots, x_p$: 설명변수, y: 반응변수

$$y = f(x_1, \ldots, x_p) + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

- Let $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and $x = (x_1, \ldots, x_p)^T$

$$y = f(x) = \beta_0 + \beta^T x.$$

9.2 통계적 추론

- 회귀계수(β_j)의 추정: 최소제곱법
  - 아래의 식을 최소화

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}))^2$$

  - 결과 : $\hat{\beta} = (X^T X)^{-1} X^T y$
- 회귀계수에 대한 t 검정: ($H_0 : \beta_j = 0$)

$$t = \frac{\hat{\beta}_j}{s.e(\beta_j)} \sim t(n-p)$$

- 적합도 평가(goodness of fit):
  - 결정계수(Coefficient of determination)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}$$

  - 수정결정계수

$$R^2_{adj} = 1 - \frac{SSE/n-p}{SST/n-1} = 1 - \left(\frac{n-p}{n-1}\right)\frac{SSE}{SST}$$

- F 검정

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \sim F(p, n-p-1).$$

9.3 Variable selection (변수선택법)

- 다중회귀모형에서는 (의미가 있던 또는 없던지) 설명변수가 많이 포함될 수록 R^2이 커짐 (과적합)
- 변수선택법 : 유의한 설명 변수를 찾는 방법
  - All possible subset regression
  - 전진선택법 (Forward selection)
  - 후진소거법 (Backward elimination)
  - 단계적선택법 (Stepwise selection)
- 변수선택의 판정기준 (Selection Criterion)
  - F-test
  - Akaike Information Ceriterion (AIC)
  - Bayesian Information Ceriterion (BIC)

9.4 R 예제: 회귀분석 (UScrime)

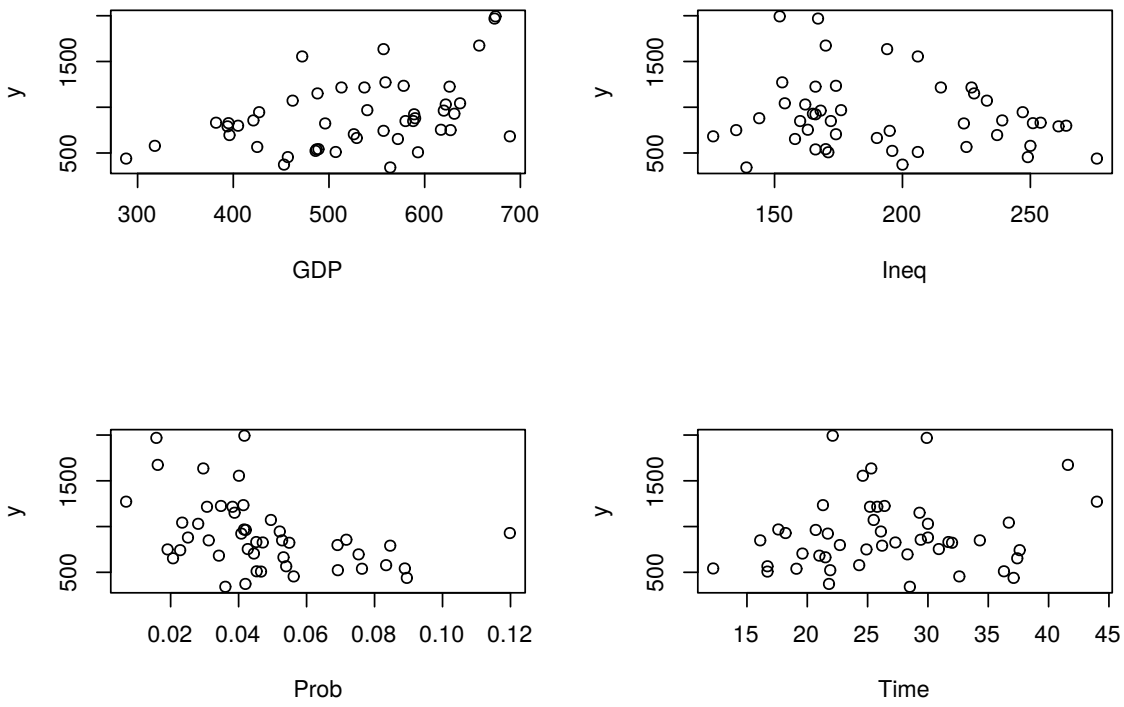
- objective: 처벌정책이 범죄율에 미치는 영향 연구
- data: 1960년 미국 47개주의 데이터
- 변수소개: 설명변수(15), 반응변수(y)

변수명	변수설명
M	percentage of males aged 14-24
So	indicator variable for a southern state
Ed	mean years of schooling
Po1	police expenditure in 1960
Po2	police expenditure in 1959
LF	labour force participation rate
M.F	number of males per 1000 females
Pop	state population
NW	number of nonwhites per 1000 people
U1	unemployment rate of urban males 14-24
U2	unemployment rate of urban males 35-39
GDP	gross domestic product per head
Ineq	income inequality
Prob	probability of imprisonment

변수명	변수설명
Time	average time served in state prisons
y	rate of crimes in a particular category per head of population

 scatter plot

```
library(MASS)
par(mfrow=c(2,2))
for(i in 12:15)
plot(UScrime[, c(i,16)])
```



상관계수

```
x <- UScrime[,12:16]
cor(x)
```

```
##                GDP       Ineq       Prob          Time          y
## GDP   1.0000000000 -0.8839973 -0.5553347  0.0006485587  0.4413199
## Ineq -0.8839972758  1.0000000  0.4653219  0.1018228182 -0.1790237
```

```
## Prob -0.5553347075  0.4653219  1.0000000 -0.4362462614 -0.4274222
## Time  0.0006485587  0.1018228 -0.4362463  1.0000000000  0.1498661
## y     0.4413199490 -0.1790237 -0.4274222  0.1498660617  1.0000000
```

 AVOVA table

```
m2 <- lm(y~GDP+Ineq+Prob+Time, data=x)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## GDP        1 1340152 1340152 14.5517 0.0004409 ***
## Ineq       1 1403081 1403081 15.2350 0.0003377 ***
## Prob       1  248260  248260  2.6957 0.1080897
## Time       1   21396   21396  0.2323 0.6323064
## Residuals 42 3868038   92096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

 회귀계수의 추정 및 검정 / 모형의 적합도

```r
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ GDP + Ineq + Prob + Time, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -587.30 -175.27  -11.71  116.08  757.78
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3008.562   1028.933  -2.924 0.005549 **
## GDP             4.590      1.058   4.338 8.84e-05 ***
## Ineq            9.362      2.463   3.801 0.000459 ***
## Prob        -4596.040   2783.448  -1.651 0.106156
## Time           -3.661      7.595  -0.482 0.632306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 303.5 on 42 degrees of freedom
```

```
## Multiple R-squared:  0.4379, Adjusted R-squared:  0.3843
## F-statistic: 8.179 on 4 and 42 DF,  p-value: 5.69e-05
```

- 단계적선택법 (AIC)

```
s <- step(m2, trace=0)
coef(summary(s))
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -3087.688952 1006.644979 -3.067307 3.728496e-03
## GDP             4.587014    1.048568  4.374553 7.611052e-05
## Ineq            9.104082    2.382697  3.820915 4.230197e-04
## Prob        -3893.644780 2350.238295 -1.656702 1.048598e-01
```

- Model comparisons
  - initial model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3008.562230	1028.933034	-2.9239631	0.0055488
GDP	4.589500	1.058067	4.3376256	0.0000884

	Estimate	Std. Error	t value	Pr(>|t|)
Ineq	9.361878	2.463027	3.8009650	0.0004594
Prob	-4596.040204	2783.448432	-1.6512036	0.1061559
Time	-3.660756	7.594888	-0.4820026	0.6323064

 final model (stepwise selection)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3087.688952	1006.644979	-3.067307	0.0037285
GDP	4.587014	1.048568	4.374553	0.0000761
Ineq	9.104082	2.382697	3.820915	0.0004230
Prob	-3893.644780	2350.238295	-1.656702	0.1048598

Measures	initial	final
Multiple R-squared	0.4379	0.4348
Adjusted R-squared	0.3843	0.3953
F-statistic	8.179	11.02
DF	(4, 42)	(3, 43)
p-value	5.69e-05	1.7e-05

9.5 범주형 변수가 포함된 경우의 회귀분석

- 가변수(dummy variable)의 이용
- 범주의 수가 K개인 범주형 변수: $(K-1)$개의 가변수(dummy variable) $z_1, \ldots, z_{K-1}$로 코딩
  - $K = 3$인 경우

범주	z_1	z_2
1	1	0
2	0	1
3	0	0

- 가변수를 이용한 선형모형

$$y = \beta_0 + \sum_{k=1}^{K-1} \beta_k z_k + \epsilon$$

9.6 Advanced Regression methods

- penalized regression :
  - lasso, glmnet, elastic net, … : 변수선택과 계수추정을 동시에 하는 방법들
- non-linear or non-parametric models
  - random forest, svm, deep learning (multi-layer neural network)

10 Logistic regression (로지스틱회귀)

- Regression model : 반응변수가 연속형인 경우

$$E(y|x_1, x_2, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

- Logistic regression model : 반응변수가 이진형인 경우 ($y \in \{0, 1\}$)

$$\log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

$$P(Y=1|x) = \frac{\exp(\beta_0 + \Sigma_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \Sigma_{j=1}^{p} \beta_j x_j)}$$

10.1 로지스틱 회귀에서 오즈비 (odds ratio)

	smoking(x=1)	non-smoking(x=0)
lung cancer(y=1)	a	c
non (y=0)	b	d

- Risk Ratio(RR): 흡연의 폐암 발생률을 비흡연의 폐암 발생률로 나눈 값

$$RR = \frac{P(Y=1|x=1)}{P(Y=1|x=0)} = \frac{a/(a+b)}{c/(c+d)}$$

- 오즈:
  - 흡연(비흡연)의 폐암 확률을 흡연(비흡연)의 폐암이 발생하지 않을 확률로 나눈 값
  - 특정 리스크에 노출될 경우의 위험도로 해석

$$odds(x=1) = \frac{P(Y=1|x=1)}{P(Y=0|x=1)} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$$

- 오즈비
  - 설명변수 $x=1$에서의 오즈와 $x=0$에서의 오즈의 비
  - x가 한 단위 증가할 때 $y=1$일 위험과 $y=0$일 위험의 비의 증가율

- 특정 리스크에 노출될 경우, 그렇지 않은 경우에 대한 상대적 위험도

$$\frac{P(Y=1|x=1)/P(Y=0|x=1)}{P(Y=1|x=0)/P(Y=0|x=0)} = \exp(\beta_1).$$

- RR vs OR
  - $P(Y=1|x=0) \approx 0$
  - 즉, 리스크에 노출되지 않을 경우 질병에 걸릴 확률)이 아주 작으면 (희귀성의 가정)
  - $RR \approx OR$

- 로그 오즈비 : 오즈비에 로그를 취한 값으로 회귀계수와 일치

- 예: x는 흡연 유무이고 y는 폐질환 여부 (1, 0)
  - $\hat{\beta} = 3.72 \rightarrow$ odds ratio $= \exp(3.72) = 42$
  - 흡연자의 폐질환에 대한 위험이 비흡연자의 위험에 비해 42배 증가하는 것으로 해석

10.2 R 예제: 전립선암 (Prostate Cancer) 양성 여부

- 림프절이 전립선암에 대해 양성인지 여부
- 53명의 환자자료
- 변수 설명

변수명	변수 설명
aged	환자의 연령
stage	질병 단계: 질병이 얼마나 진행되어 있는지 나타내는 측도
grade	종양의 등급: 진행의 정도
xray	X-선 결과
acid	혈청인산염(serum acid phosphatase) 특정한 부위에 종양이 전이되었을 때 상승되는 혈청의 인산염값
r	전립선암 양성(1) 여부

```
library(boot)
x<-nodal[,-1]
head(x, 20)
```

```
##    r aged stage grade xray acid
```

```
## 1  1    0     1     1    1    1
## 2  1    0     1     1    1    1
## 3  1    0     1     1    1    1
## 4  1    0     1     1    1    1
## 5  1    0     1     1    1    1
## 6  0    0     1     1    1    1
## 7  1    0     0     0    0    1
## 8  0    0     0     0    0    1
## 9  0    0     0     0    0    1
## 10 0    0     0     0    0    1
## 11 0    0     0     0    0    1
## 12 0    0     0     0    0    1
## 13 0    1     1     1    0    0
## 14 0    1     1     1    0    0
## 15 0    1     1     1    0    0
## 16 0    1     1     1    0    0
## 17 1    1     1     0    0    1
## 18 1    1     1     0    0    1
## 19 0    1     1     0    0    1
## 20 0    1     1     0    0    1
```

 initial model

```r
gfit = glm(r~., data=x, family="binomial")
coef(summary(gfit))
```

```
##               Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -3.0793806  0.9867696 -3.1206684 0.001804411
## aged        -0.2917427  0.7540054 -0.3869239 0.698812567
## stage        1.3729295  0.7838488  1.7515235 0.079855775
## grade        0.8719723  0.8155785  1.0691457 0.285004012
## xray         1.8008141  0.8104165  2.2220847 0.026277583
## acid         1.6839295  0.7914741  2.1275863 0.033371400
```

- stepwise selection

```r
m <- step(gfit, trace =0)
coef(summary(m))
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -3.051787  0.8420409 -3.624273 0.0002897749
## stage        1.645346  0.7296744  2.254905 0.0241392977
## xray         1.911626  0.7771336  2.459842 0.0138998172
## acid         1.637778  0.7539433  2.172283 0.0298343277
```

- 결과
  - 유의한 변수들(유의확률: $p < 0.05$)은 전립선암에 영향을 줌
    - 질병의 단계(stage)가 심화
    - X-선 결과(xray)가 좋지 않을수록
    - 혈청인산염 값(acid)이 높을수록
  - stage(질병의 단계)의 오즈비
    - $\exp(1.645346) = 5.18280$
    - 질병의 진행단계가 심화된 그룹은 그렇지 않은 그룹에 비해 전립선암에 노출 위험이 약 5.2배 정도 높음