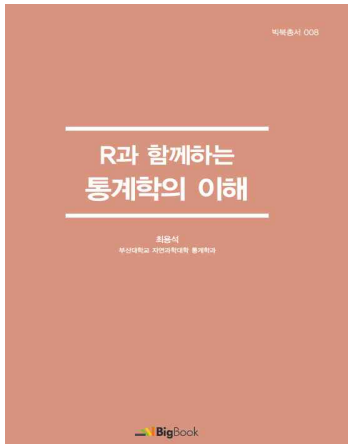
Basic Statistics with R



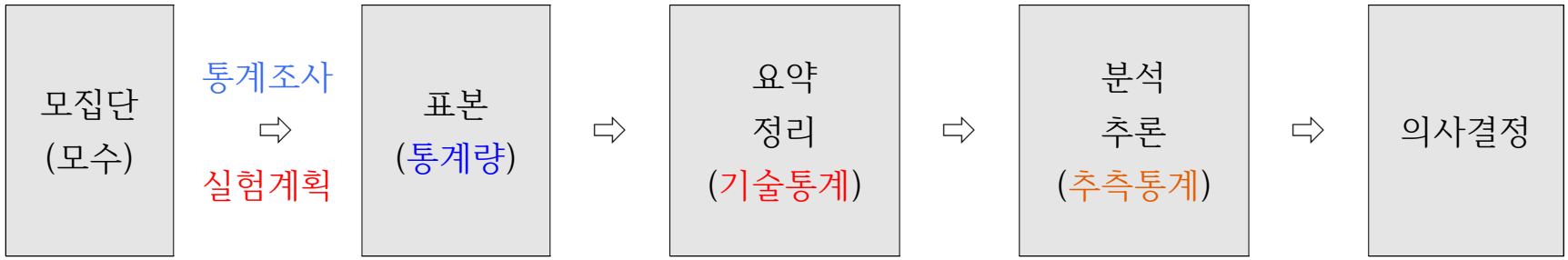
최 용 석 (부산대학교 통계학과)

yschoi.pusan.ac.kr

2019.08.21.

부산대학교 통계연구소 하계 워크숍

주제 1 : R을 활용한 통계학



통계자료 분석의 단계

1.1 자료의 이해

1.2 한 집단의 비교 : 모평균 검정

1.3 두 집단의 비교 : 독립표본, 대응표본

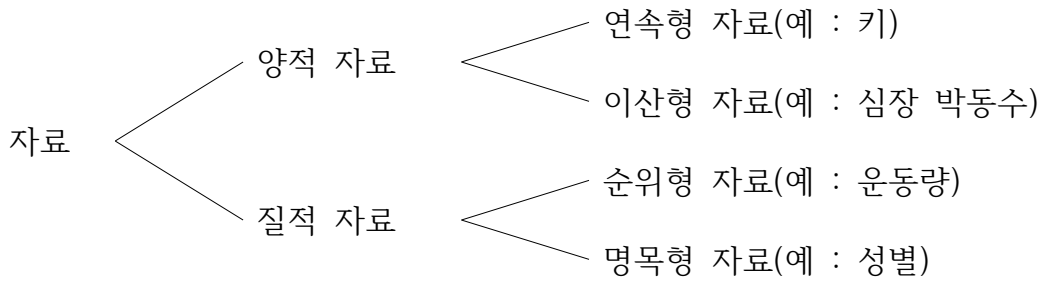
1.4 여러 집단의 비교 : 일원 분산분석

1.5 범주들의 관계 비교 : 독립성검정, 동질성 검정

1.6 선형모형 : 단순 회귀모형, 다중 회귀모형

1.1 자료의 이해

자료의 종류



[표 1.1.1] 심장 박동수 자료

학생	처음 심장 박동수	나중 심장 박동수	달리기	흡연	성별	키(cm)	몸무게(kg)	운동량
1	64	88	1	2	1	168	63.5	2
2	58	70	1	2	1	183	65.8	2
3	62	76	1	1	1	185	72.6	3
4	66	78	1	1	1	185	86.2	1
5	64	80	1	2	1	175	70.3	2
6	64	60	2	2	2	168	81.6	3
7	94	92	2	1	2	157	82.1	2
8	60	66	2	2	2	157	54.4	2
9	72	70	2	2	2	173	78.5	2
10	58	56	2	2	2	170	56.7	2

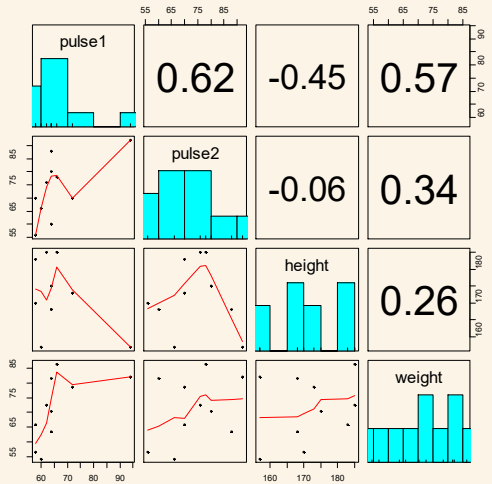
[보기 1.1.2] [표 1.1.1] 심장 박동수 자료에서 양적 자료의 요약

[표 1.1.1]에서 양적 자료는 처음 심장 박동수, 나중 심장 박동수, 키(cm), 몸무게(kg)로 이들에 대해 동시에 요약하기로 하자.

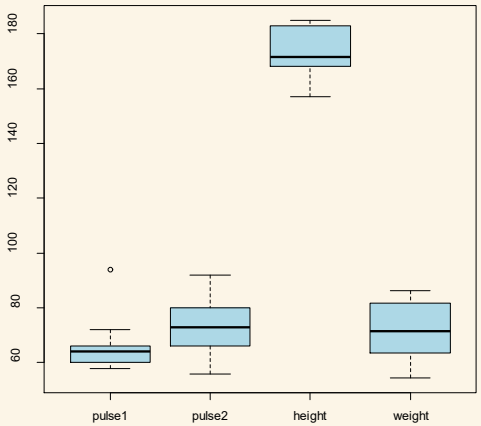
[1] 기술통계(descriptive statistics)

```
     pulse1          pulse2          height          weight
 Min.   :58.0   Min.   :56.0   Min.   :157.0   Min.   :54.40
 1st Qu.:60.5   1st Qu.:67.0   1st Qu.:168.0   1st Qu.:64.08
 Median :64.0   Median :73.0   Median :171.5   Median :71.45
 Mean   :66.2   Mean   :73.6   Mean   :172.1   Mean   :71.17
 3rd Qu.:65.5   3rd Qu.:79.5   3rd Qu.:181.0   3rd Qu.:80.83
 Max.   :94.0   Max.   :92.0   Max.   :185.0   Max.   :86.20
```

[2] 다중 산점도(multiple scatter plot)



[3] 상자그림(box plot)



[4] 줄기-잎 그림(stem-and-leaf plot)

```
5 | 88
6 | 024446
7 | 2
8 |
9 | 4
```

pulse1

```
5 | 6
6 | 06
7 | 0068
8 | 08
9 | 2
```

pulse2

```
15 | 77
16 | 88
17 | 035
18 | 355
```

height

```
5 | 47
6 | 46
7 | 039
8 | 226
```

weight

[R-코드 1.1.1] [보기 1.1.1]과 [보기 1.1.2] 심장 박동수 자료의 표현(심장박동수자료요약.R)

```
setwd("C:/기초통계학/R-codes")
data1.1.1<-read.table("pulse.txt", header=T)
x<-data1.1.1
x
## [보기 1.1.1] 질적자료의 표현
# 운동량: 막대그림, 원도표
win.graph()
par(mfrow=c(1,2))
activity=x[,8]
freq=table(activity)
rel_freq=prop.table(freq)
cbind(freq, rel_freq)
y=round(rel_freq*100)
z=paste(lab=c("1:적음", "2:보통", "3:많음"), "(",y,"%",")")
pie(freq,labels=z)
barplot(freq, xlab="운동량(1:적음, 2:보통, 3:많음)", ylab="빈도")


## [보기 1.1.2] 양적자료의 표현 : 처음과 나중 심장 박동수, 키, 몸무게
x1<-x[, c(1,2,6,7)]

# [1] 기술통계
summary(x1)
# [2] 다중산점도
library(psych)
pairs.panels(x1, density=F, ellipses=F)
# [3] 상자그림
boxplot(x1, col="lightblue")
# [4] 줄기-잎 그림
for (j in 1:4){
stem(x1[,j])
}
```

질적 자료

- pie(), barplot() : 원도표, 막대그림
- table() : 분할표

양적 자료

- summary() : 평균(mean), 분산(variance), 등 기초통계량
- library(psych) > pairs.panels() : 다중 산점도
- boxplot() : 상자그림
- stem() : 줄기-잎 그림

1.2 한 집단의 비교 : 모평균 검정

대표본의 모평균에 대한 유의수준 α에서 $Z-$검정

귀무가설	대립가설	p-값	검정 통계량
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P[Z \geq z]$	$Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$
	$H_1 : \mu < \mu_0$	$P[Z \leq z]$	
	$H_1 : \mu \neq \mu_0$	$P[\lvert Z \rvert \geq z]$	$\approx N(0,1)$

소표본의 모평균에 대한 유의수준 α에서 $t-$검정

귀무가설	대립가설	p-값	검정 통계량
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	$P[T \geq t]$	$T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$
	$H_1 : \mu < \mu_0$	$P[T \leq t]$	
	$H_1 : \mu \neq \mu_0$	$P[\lvert T \rvert \geq t]$	

p-값(p-value)의 정의와 활용

- H_0가 참일 때 검정통계량이 표본을 통한 검정통계량의 관측값과 같거나 더 지나친 값(크거나 작은 값)을 취할 확률이다.
- p-값$< \alpha \Rightarrow$ 귀무가설 H_0를 유의수준 α에서 기각한다.

[보기 1.2.1] 대표본 묘목의 평균 크기에 대한 가설 및 검정의 이해

어린 소나무의 성장을 연구하기 위하여 1년생 붉은 소나무 묘목 40그루의 크기를 조사한 자료가 다음과 같다. $n = 40$이고 표본평균 $\bar{x} = 1.715$와 표준편차 $s = 0.475$이므로 유의수준 0.05에서

$$H_0 : \mu = 1.9,\ H_1 : \mu > 1.9$$

를 검정하기 위한 기각역을 구하고 검정통계량의 관측값을 구한 후 기각역을 만족하는 지를 알아보라. 그리고 유의확률 p–값에 의해서도 검정해보라.

2.6	1.9	1.8	1.6	1.4	2.2	1.2	1.6	1.6	1.5
1.4	1.6	2.3	1.5	1.1	1.6	2.0	1.5	1.7	1.5
1.6	2.1	2.8	1.0	1.2	1.2	1.8	1.7	0.8	1.5
2.0	2.2	1.5	1.6	2.2	2.1	3.1	1.7	1.7	1.2

1) $\overline{X}$의 분포는 근사적으로 $N(1.9, \frac{0.475^2}{40})$이 되고 이를 표준화한 검정통계량은 근사적으로 표준정규분포를 따르며 다음과 같다.

$$Z = \frac{\overline{X} - 1.9}{0.475/\sqrt{40}} \simeq N(0,1)$$

2) 유의수준 $\alpha = 0.05$에서 기각치는 $z_{0.05} = 1.645$이므로 기각역: $z > 1.645$

3) 표본평균이 $\overline{x} = 1.715$이므로 검정통계량의 관측값 :

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.715 - 1.9}{0.475/\sqrt{40}} = -2.463$$

4) $z = -2.463$가 기각역을 만족하지 못하므로 귀무가설을 기각할 수 없게 된다.

<=> 유의수준 $\alpha = 0.05$에서 $p-$값 = 0.993 > α이므로 귀무가설을 기각할 수 없게 된다.

5) 이 자료에 따르면 붉은 소나무 묘목의 키의 평균과 다른 소나무 묘목의 키의 평균과 같음을 알 수 있다.

[R-코드 1.2.1] **[보기 1.2.1] 대표본의 모평균에 대한 z-검정**

```
z.test = function(x, mu){
zeta = (mean(x) - mu)/(sqrt(var(x)/length(x)))
  return(zeta)}
x=c(2.6, 1.9, 1.8, 1.6, 1.4, 2.2, 1.2, 1.6,
    1.6, 1.5, 1.4, 1.6, 2.3, 1.5, 1.1, 1.6,
    2.0, 1.5, 1.7, 1.5, 1.6, 2.1, 2.8, 1.0,
    1.2, 1.2, 1.8, 1.7, 0.8, 1.5, 2.0, 2.2,
    1.5, 1.6, 2.2, 2.1, 3.1, 1.7, 1.7, 1.2)
z=z.test(x, 1.9)
pvalue=1-pnorm(z)
list(z, pvalue)
```

$H_0 : \mu = 1.9, \ H_1 : \mu > 1.9$

- z.test():
- z=z.test(x, 1.9, 0.2256) :
  모평균 1.9, 표본분산 $s^2 = 0.475^2 = 0.2256$
- pvalue=pnorm(z) :
  p-값 = $P[Z > -2.463]$ =1- $P[Z \leqq -2.463]$

```
> list(z, pvalue)
[[1]]
[1] -2.464421

[[2]]
[1] 0.9931383
```

[보기 1.2.2] 소표본 : 생수 세균의 수

어느 도시 위생국에서는 어떤 생수의 단위 량당 세균의 평균 숫자가 안전수준인 200 이내인지를 조사하고자 한다. 한 조사원이 10개의 표본자료를 검사한 결과 세균의 수는 다음과 같고 정규분포를 따른다.

175, 190, 215, 198, 184, 207, 210, 193, 196, 180

귀무가설과 대립가설을 세우고 이를 유의수준 $\alpha = 0.01$에서 검정하여 보라.

1) μ를 이 생수 단위량당 세균의 평균 숫자라고 하자. 이 생수가 안전수준이라면 $\mu < 200$이고 조사원은 이 가설을 뒷받침하는 강력한 증거를 찾고자 한다. 그러므로 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \mu = 200, \quad H_1 : \mu < 200$$

2) 표본의 크기가 $n = 10$인 소표본이며 정규분포를 따르므로 t-검정을 실시할 수 있다.

$$\bar{x} = 194.8, \quad s = 13.14$$

$$t = \frac{194.8 - 200}{13.14/\sqrt{10}} = \frac{-5.2}{4.156} = -1.25$$

3) 대립가설 $H_1 : \mu < 200$에 대해 유의수준 $\alpha = 0.01$에서 자유도 $df = n - 1 = 9$인 기각치는 $-t_{0.01} = -2.821$이고

기각역: $t \leq -2.821$

4) 검정통계량의 관측치 $t = -1.25$는 기각역 $t \leq -2.821$을 만족하지 않는다. 이는 유의수준 $\alpha = 0.01$에서 귀무가설 $H_0 : \mu = 200$이 기각되지 않음을 말한다.

<=> 유의수준 $\alpha = 0.01$에서 p-값 = 0.121 > α이므로 귀무가설을 기각할 수 없다.

5) 10개의 표본자료는 모평균이 안전수준 이내에 있다는 증거가 되지 못한다.

[R-코드 1.2.2] **[보기 1.2.2] 소표본의 모평균에 대한 t-검정**

```
x=c(175, 190, 215, 198, 184, 207, 210, 193,
    196, 180)
list(mean(x),sd(x))
t.test(x, mu=200, alternative="less")
```

참고:

- $H_1 : \mu > 200$: alternative = "greater"
- $H_1 : \mu \neq 200$: alternative = "two.sided"

$H_0 : \mu = 200, \quad H_1 : \mu < 200$

- t.test():

```
> list(mean(x),sd(x))
[[1]]
[1] 194.8

[[2]]
[1] 13.13858

> t.test(x, mu=200, alternative="less")

        One Sample t-test

data:  x
t = -1.2516, df = 9, p-value = 0.1211
alternative hypothesis: true mean is less
than 200
95 percent confidence interval:
     -Inf 202.4162
sample estimates:
mean of x
    194.8
```

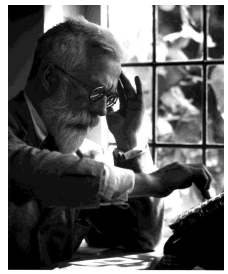
William Sealey Gosset 1876.6.13.-1937.10.16

New College Oxford에 입학 화학과 수학을 공부

- 1899년 Dublin의 Guinness 양조장에서 화학자 일함
- 술의 질을 관리 소표본에 적합한 t-test를 발견.
- 1922년 양조장 통계 컨설턴트 1934년까지 통계부서 운영.
- 통계학자와 왕래와 논쟁: 피셔
- 1935년 말 새 Guinness 양조장을 운영하기 위해 아일랜드로 떠남.
- 양조장 사업의 고된 일에도 불구하고 계속 통계학 논문을 발표하였다(Student-test).

가수 김건모 : 잘못된 만남 앨범(280 만장 판매)
강호동: 8시간 동안 쉬지 않고 악수
신세계백화점 센텀시티: 세계에서 가장 큰 백화점
2017.11 방탄소년단 : 트위터 최다 활동 남성그룹 부문

R. A. Fisher 1890.2.17-1962.7.29



- 1909년 캠브리지 대학 장학생 입학 1912년 수학의 수석 졸업
- 캠브리지에서 물리학 장학생으로 1년을 더 수학하였다

Francis Galton 822.2.16~1911.1.17

- 어릴 때부터 수학에 재능을 보였으나 지독한 근시로 머릿속으로 풀고 기하적 감각이 길러짐
- 1919년 칼 피어슨의 콜튼 연구소와 존러셀의 Rothamsted 연구소로부터 동시에 일자리를 제안 받게 되었는데 피어슨과의 논쟁으로 존러셀의 제안을 받아 들였다.
- 1930년도에 발간된 '자연도태의 유전이론'은 다원적 생각들과 멘델이론을 잘 융화
- 1933년 유전학 연구업적으로 칼 피어슨의 뒤를 이어 London University College의 교수가 됨.
-칼 피어슨의 아들인 에곤 피어슨이 통계분야를 분리해 내어 응용통계학과를 발족
- Statistics: Test, MLE, Experimental Design (ANOVA, F-distribution). P-value

1.3 두 집단의 비교: 독립표본, 대응표본

소표본과 대표본 독립표본 모평균 차의 검정 절차		
표본의 크기	소표본($n_1,\ n_2 \leq 30$)	대표본($n_1,\ n_2 > 30$)
모집단의 정규성 가정	$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$	$X \sim (\mu_1, \sigma_1^2),\ \ Y \sim (\mu_2, \sigma_2^2)$
모르는 모분산 가정	$\sigma_1^2 = \sigma_1^2 = \sigma^2$	$\sigma_1^2 \neq \sigma_1^2$
검정통계량	$T = \dfrac{\overline{X} - \overline{Y}}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$, $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$:합동표본분산	$Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}$, $s_1^2,\ s_2^2$: 표본분산
귀무가설	$H_o:\ \mu_1 - \mu_2 = 0$	
대립가설	$H_1 : \mu_1 - \mu_2 > 0,\ H_1 : \mu_1 - \mu_2 < 0,\ H_1 : \mu_1 - \mu_2 \neq 0.$	
$p-$값	$P(T \geq t), P(T \leq t),\ P(\lvert T \rvert \geq t)$	$P(Z \geq z), P(Z \leq z), P(\lvert Z \rvert \geq z)$
검정방법	$p-$값$< \alpha$ 이면 귀무가설 $H_o:\ \mu_1 - \mu_2 = 0$를 기각한다.	

대응표본(소표본)의 검정 절차

	처리 대상					
	1	2	…	i	…	n
처리 1	X_1	X_2	…	X_i	…	X_n
처리 2	Y_1	Y_2	…	Y_i	…	Y_n
차	$D_1 = X_1 - Y_1$	$D_2 = X_2 - Y_2$	…	$D_i = X_i - Y_i$	…	$D_n = X_n - Y_n$
통계량	새로운 자료: $D_1, D_2, \ldots, D_n$ 표본평균: $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, 표본분산: $s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$					
검정 통계량	소표본 : $T = \frac{\overline{D}}{s_D/\sqrt{n}} \sim t_{(n-1)}$, 대표본 : $Z = \frac{\overline{D}}{s_D/\sqrt{n}} \approx N(0,1)$					
귀무 가설	$H_o : \mu_D = 0$					
대립 가설	$H_o : \mu_D > 0$, $H_o : \mu_D < 0$, $H_o : \mu_D \neq 0$					
$p-$값	$P(T \geq t), P(T \leq t), P(\lvert T \rvert \geq t)$					

[보기 1.3.1] 소표본 젖소용 사료의 우유 생산량 평균의 차이

젖소용 사료 X와 Y의 우유생산량이 같다는 연구결과가 있다. 이를 입증하기 위하여 정규분포를 따르며 분산이 동일한 두 모집단으로부터 각각 추출한 13마리에는 사료 X를 공급하고 나머지 12마리에는 사료 Y를 공급하였다. 3주일 후 젖소의 우유 생산량 검사를 해본 결과 다음의 자료를 얻었다. 알려진 연구 결과와는 달리 사료 X의 우유생산량이 사료 Y의 생산량보다 더 많다고 할 수 있는 지를 유의수준 5%에서 검정하라.

모집단	자료 크기	평균	표준편차
사료 X	$n_1 = 13$	45.15	7.998
사료 Y	$n_2 = 12$	42.25	8.740

1) 사료 X와 Y의 우유생산량의 모평균을 각각 μ_1과 μ_2라 하면 가설 : $H_o : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$

2) 표본의 크기가 13과 12로 소표본에 해당하고 모분산이 동일하며 정규분포를 따르므로 합동표본분산 :

$$s_p = \sqrt{\frac{12(7.998)^2 + 11(8.740)^2}{23}} = 8.36$$

검정통계량의 관측값 계산 : $$t = \frac{45.15 - 42.25}{8.36\sqrt{\frac{1}{13} + \frac{1}{12}}} = \frac{2.90}{3.347} = 0.87$$

3) 대립가설이 단측인 $H_1 : \mu_1 - \mu_2 > 0$이므로 유의수준 5%에서 자유도 $df = n_1 + n_2 - 2 = 23$인 기각치는 $t_{.05} = 1.714$.

기각역 : $t \geqq 1.714$.

4) 검정통계량의 관측값 $t = 0.87$이 기각역을 벗어나므로 귀무가설 $H_o : \mu_1 - \mu_2 = 0$을 기각하지 못한다.

<=> 유의수준 $\alpha = 0.05$에서 $p-$값 = 0.197 > α이므로 귀무가설을 기각할 수 없다.

5) 이 연구에 따르자면 사료 X의 우유생산량이 사료 Y의 생산량보다 더 많다고 할 수 없다.

[R-코드 1.3.1] **[보기 1.3.1] 소표본의 두 집단의 모평균 t-검정**

```
x=c(44, 44, 56, 46, 47, 38, 58, 53,
49, 35, 46, 30, 41)
y=c(35, 47, 55, 29, 40, 39, 32, 41,
42, 57, 51, 39)
t.test(x, y, var.equal=T,
    alt="greater")
```

$H_o : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$

```
> t.test(x, y, var.equal=T, alt="greater")

        Two Sample t-test

data:  x and y
t = 0.8676, df = 23, p-value = 0.1973
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -2.83279      Inf
sample estimates:
mean of x mean of y
 45.15385  42.25000
```

[보기 1.3.2] 소표본 두 집단의 분산이 다른 경우

새롭게 개발된 농약이 과실의 수확량을 높인다는 연구가 있다. 실제로 과수원에서 9그루 나무에 기존의 농약을 사용하였고, 나머지 9그루 나무에 새 농약을 사용하여 수확한 과실의 수확량에 대한 다음의 자료를 얻었다. 연구결과와 같이 새로운 농약 X가 기존의 농약 Y보다 과실의 평균 수확량을 더 많이 낸다고 할 수 있는 지를 유의수준 1%에서 검정하라.

모집단	표본의 크기	평균 수확량	표준편차
농약 X	9	249	19
농약 Y	9	233	45

1) 새 농약 X와 기존의 농약 Y의 수확량의 모평균을 각각 μ_1과 μ_2라 하면 다음과 같은 가설을 세울 수 있다.

$$H_o : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$$

2) 두 표본의 크기 9는 소표본에 해당하며 두 표준편차의 비가 $19/45 = 0.42$로 두 모평균이 같다고 보기 어렵다. 따라서 수정된 자유도

$$df = \frac{\left[\frac{19^2}{9} + \frac{45^2}{9}\right]^2}{\frac{1}{8}\left[\frac{19^2}{9}\right]^2 + \frac{1}{8}\left[\frac{45^2}{9}\right]^2} = 10.765$$

검정통계량의 관측값:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{249 - 233}{\sqrt{\frac{19^2}{9} + \frac{(45)^2}{9}}} = 0.983$$

3) 대립가설이 단측인 $H_1 : \mu_1 - \mu_2 > 0$에 대해 유의수준 1%에서 자유도 $df = 11$인 기각치 $t_{.01} = 2.718$이다.

기각역: $t \geqq 2.718$.

검정통계량의 관측값 $t = 9.98$가 기각역을 만족하지 않으므로 귀무가설 $H_o : \mu_1 - \mu_2 = 0$을 기각하지 못한다.

<=> 유의수준 $\alpha = 0.05$에서 $p-$값 = 0.174 > α이므로 귀무가설을 기각할 수 없다.

4) 새로 개발된 농약 X가 기존 농약 Y의 농작물 평균 수확량보다 더 많이 낸다고 할 수 없다.

참고: 수정된 자유도 계산 대신에 두 표본의 크기로부터 $n_1 - 1$과 $n_2 - 1$중 작은 값을 자유도로 정하는 방법에 따르면 두 표본의 크기가 9로 같으므로 자유도는 8이 된다. 이 때 기각치는 $t_{.01} = 2.896$이며 검정의 결과는 동일하다.

<=> 유의수준 $\alpha = 0.05$에서 $p-$값 = 0.171 > α이므로 귀무가설을 기각할 수 없다.

[R-코드 1.3.2] [보기 1.3.2] 소표본의 두 집단(분산이다름)의 모평균 $t-$검정

```
#평균,표준편차
mx=249;sx=19
my=233;sy=45

# t-검정통계량 : 소표본 분산이 다를때
n1=9
n2=9
#t-test : 소표본
t=(mx-my)/sqrt(sx^2/n1 + sy^2/n2)
wdf=((sx^2/n1 +
sy^2/n2)^2)/((sx^2/n1)^2/(n1-1) +
(sy^2/n2)^2/(n2-1))

#p-value : 단측검정
pvalue1=1-pt(t, df=wdf)

#p-value : df=min(n1-1, n2-1)
pvalue2=1-pt(t, df=15)

list(t, wdf, pvalue1, pvalue2)
```

$H_o : \mu_1 - \mu_2 = 0,\ H_1 : \mu_1 - \mu_2 > 0$

```
> list(t, wdf, pvalue1, pvalue2)
[[1]]
[1] 0.9826662

[[2]]
[1] 10.76449

[[3]]
[1] 0.1736712

[[4]]
[1] 0.1706785
```

[보 기 1.3.3] 두 회사 등산화 마모 차이

두 아웃도어 회사의 등산화 밑창의 마모 정도를 비교하기 위하여 10명의 학생들을 대상으로 X회사 제품을 왼발에 착용하였고 오른발에는 Y회사 제품을 각각 착용하게 하여 마모를 측정한 결과 다음의 자료를 얻었다. X회사 제품의 등산화의 마모율이 Y회사 제품과 차이가 있다고 할 수 있는 지를 유의수준 5%에서 검정하라.

모집단	처리 대상									
제품 X (왼발)	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
제품 Y (오른발)	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
차 D	0.8	0.6	0.3	−0.1	1.1	−0.2	0.3	0.5	0.5	0.3

[1] 수술 전 X와 수술 후 Y의 S-perpendicular to the PNS의 거리 변화 차이의 모평균 μ_D에 대한 귀무가설과 대립가설:

$$H_0 : \mu_D = 0, \quad H_1 : \mu_D \neq 0$$

[2] 대응표본의 크기 10은 소표본:

$$\overline{D} = \frac{1}{10}\sum_{i=1}^{n} D_i = 0.41, \qquad s_D^2 = \frac{1}{10-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = 0.0149$$

[3] 귀무가설 H_o : $\mu_D = 0$하에 검정통계량의 관측값:

$$t = \frac{0.41}{0.122/\sqrt{10}} = 3.349$$

3) 대립가설이 양측인 $H_1 : \mu_D \neq 0$에 대해 유의수준 5%에서 자유도 $df = n-1 = 9$에서 기각치 $t_{.0.025} = 2.262$이다.

기각역: $|t| \geq 2.262$

검정통계량의 관측값의 $|t| = 3.349$가 기각역을 만족하므로 귀무가설 $H_o : \mu_D = 0$을 기각한다.

<=> 유의수준 $\alpha = 0.05$에서 $p-$값 = 0.0091 < α이므로 귀무가설을 기각한다.

4) 등산화 제품 X와 제품 Y의 마모율은 차이가 있다고 볼 수 있다.

[R-코드 1.3.3] **[보기 1.3.3] 대응표본에 대한 검정 t-검정**

```
x=c(14.0, 8.8, 11.2, 14.2, 11.8, 6.4,
9.8, 11.3, 9.3, 13.6)
y=c(13.2, 8.2, 10.9, 14.3, 10.7, 6.6,
9.5, 10.8, 8.8, 13.3)
t.test(x,y,paired=T,
alternative="two.sided")
```

$H_0 : \mu_D = 0,\ H_1 : \mu_D \neq 0$

```
> t.test(x, y, paired=T, alternative="two.sided")

        Paired t-test

data:  x and y
t = 3.3489, df = 9, p-value = 0.008539
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.1330461 0.6869539
sample estimates:
mean of the differences
                   0.41
```

[R-코드 1.3.4] [보기 1.3.4] 대응표본에 대한 t-검정

```
## 독감 예방 백신에 의한 항체율 비교
## 방법 1
z.prop = function(x1,x2,n1,n2){
  numerator = (x1/n1) - (x2/n2)
  p.common = (x1+x2) / (n1+n2)
  denominator = sqrt(p.common * (1-p.common)
* (1/n1 + 1/n2))
  z.prop.ris = numerator / denominator
return(z.prop.ris)}
n1 <- 113
n2 <- 139
x <- 34
y <- 54
z=z.prop(x, y, n1, n2)
z
pvalue1=pnorm(z)
pvalue1

## 방법 2
# Chisquare test
chi=z^2
pvalue2=(1-pchisq(chi, df=1))/2
list(chi, pvalue2)

## 방법 3
#  prop.test()
prop.test(x=c(34, 54), n=c(113,
139),alternative="less", correct=F)
```

$$H_o : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0$$

```
> list(z, pvalue1)
[[1]]
[1] -1.450804

[[2]]
[1] 0.07341719

> list(chi, pvalue2)
[[1]]
[1] 2.104833

[[2]]
[1] 0.07341719

> prop.test(x=c(34, 54), n=c(113, 139),alternative="less", correct=F)

  2-sample test for equality of proportions without continuity
correction

data:  c(34, 54) out of c(113, 139)
X-squared = 2.1048, df = 1, p-value = 0.07342
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01068364
sample estimates:
   prop 1    prop 2
0.3008850 0.3884892
```

1.4 여러 집단의 비교 : 일원 분산분석 - Fisher(1935) The Design of Experiments

일원 분산분석을 적용할 g개 처리(그룹)의 독립표본의 구조

처 리(그룹)	1	2	…	g
자 료	y_{11} y_{12} ⋮ y_{1n_1}	y_{21} y_{22} ⋮ y_{2n_2}	… … …	y_{g1} y_{g2} ⋮ y_{gn_g}
평 균	$\bar{y}_1$	$\bar{y}_2$	…	$\bar{y}_g$
총평균	$\bar{y} = \frac{n_1\bar{y}_1 + \cdots + n_g\bar{y}_g}{n}$			

일원 분산분석 모형 : $y_{ij} = \mu_i + e_{ij},\ j = 1,...,n_i,\ i = 1,...,g$

μ_i : i번째 처리에 대한 효과(effect)로 평균 반응

e_{ij} : 서로 독립이며 정규분포 $N(0,\sigma^2)$을 따르는 오차항

$\Rightarrow$ 귀무가설 : $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$(처리의 효과가 동일하다)

- 변동의 분해

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{g} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

$$\Leftrightarrow \quad TSS = SST + SSE$$

일원 분산분석표

요 인 (Source)	자유도 (df)	제곱합 (Sum of Squares)	평균제곱합 (Mean Square)	분산비 (F value)	$p-$값
처 리 (Model)	$g-1$	SST	$MST = SST/(g-1)$	$F = \frac{MST}{MSE}$	$P[F \geq f]$
오 차 (Error)	$n-g$	SSE	$MSE = SSE/(n-g)$		
전 체 (Total)	$n-1$	TSS			

$$F = \frac{MST}{MSE} = \frac{SST/(g-1)}{SSE/(n-g)} > F_{(\alpha)}(g-1, n-g)$$

$\Leftrightarrow p-$값 $< \alpha \Rightarrow$ Reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ at significant level α

[보기 1.4.1] 4가지 수영 강습법의 효과 비교

스포츠 센터의 수영 강습생중 수준이 비슷한 24명을 임의 추출하여 42 수준으로 나눈 후 네 가지 방법을 적용하여 강습하였다. 강습법에 차이가 있는지, 어느 것이 효과적인지에 관심이 있다. 1개월의 과정이 끝난 후 이들 수료자들에게 25m를 접영으로 수영하게 하여 걸린 시간을 기록한 자료는 다음과 같다.

강습법	시간(초)					
강의 및 비디오 촬영	21.4	20.1	21.1	19.6	21.8	19.0
강의	17.8	19.3	19.1	18.8	18.3	19.0
비디오 촬영	18.9	20.3	19.1	19.6	20.0	20.1
강의 및 비디오 촬영 없이	19.9	18.4	18.0	17.9	20.2	19.5

1) 네 가지 수영 강습법 A, B, C, D는 처리가 t=4개임을 말하며 강습법의 효과가 같다는 다음의 귀무가설을 세울 수 있다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

2) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$을 F-검정하기 위한 일원 분산분석표는 다음과 같다.

요 인	자유도	제곱합	평균제곱합	분산비
처 리	3	11.42	3.81	5.31
오 차	20	14.35	0.72	
전 체	23	25.77		

3) 분산분석표로부터 검정통계량 관측값 $f = 5.31$은 자유도 (3, 20)을 따르며 유의수준 0.05에서 기각치 $F_{0.05}(3, 20) = 3.10$ 보다 크므로 기각역을 만족한다. 즉,

$$f = 5.31 > F_{0.05}(3, 20) = 3.10$$

<=> 유의수준 $\alpha = 0.05$에서 $p-$값 = 0.007 < α이므로 귀무가설을 기각한다.

4) 분산분석의 결과 귀무가설은 기각되고 수영 강습법의 효과에는 차이가 있다고 볼 수 있다.

[R-코드 1.4.1] [보기 1.4.1] 4가지 수영 강습법의 효과 비교

```
time=c(21.4, 20.1, 21.1, 19.6, 21.8, 19.0, 17.8, 19.3, 19.1, 18.8,
 18.3, 19.0, 18.9, 20.3, 19.1, 19.6, 20.0, 20.1, 19.9, 18.4, 18.0,
 17.9, 20.2, 19.5)
lecture = c(rep("A",6), rep("B",6), rep("C",6), rep("D",6))
swim = data.frame(time,lecture)
summary(unstack(swim))
results = aov(time ~ lecture, data=swim)
anova(results)
```

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

```
> summary(unstack(swim))
       A               B               C               D
 Min.   :19.00   Min.   :17.80   Min.   :18.90   Min.   :17.90
 1st Qu.:19.73   1st Qu.:18.43   1st Qu.:19.23   1st Qu.:18.10
 Median :20.60   Median :18.90   Median :19.80   Median :18.95
 Mean   :20.50   Mean   :18.72   Mean   :19.67   Mean   :18.98
 3rd Qu.:21.32   3rd Qu.:19.07   3rd Qu.:20.07   3rd Qu.:19.80
 Max.   :21.80   Max.   :19.30   Max.   :20.30   Max.   :20.20
> results = aov(time ~ lecture, data=swim)
> anova(results)
Analysis of Variance Table

Response: time
           Df Sum Sq Mean Sq F value    Pr(>F)
lecture     3 11.423  3.8078   5.307 0.007437 **
Residuals 20 14.350  0.7175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

1.5 범주들의 관계 비교 : 독립성 검정, 동질성 검정

[보기 1.5.1] 독립성 검정

어느 제약회사에서 489명의 고객들이 어떤 종류의 약을 선호하는지를 알기 위해 설문조사를 실시하여 고객들의 나이와 먹는 약의 종류라는 2개의 범주형 변수에 따라 자료를 표로 정리하였다. 나이는 20세 이상 30세 미만, 30세 이상 50세 미만, 50세 이상인 세 범주로 구성되어 있고 먹는 약의 종류는 캡슐형과 정제형인 두 범주로 구성되어 있다.

나 이	먹는 약의 종류		합 계
	캡술형	정제형	
20세 이상 30세 미만	38	79	117
30세 이상 50세 미만	87	118	205
50세 이상	78	89	167
합 계	203	286	489

[보기 1.5.2] 동질성 검정

두 가지 식이요법 A와 B의 효과를 비교하기 위해 150명의 환자를 대상으로 임의 추출된 80명에게는 식이요법 A를 나머지 70명에게는 식이요법 B를 적용하였다. 얼마간의 시간이 흐른 뒤에 식이요법과 각 환자의 건강상태인 2개의 범주형 변수에 따라 자료를 표로 정리하였다. 식이요법은 범주 A와 B로, 환자의 건강상태는 양호, 보통, 불량으로 세 범주로 구성되어 있다.

식이요법	환자의 건강상태			합 계
	양 호	보 통	불 량	
A	37	24	19	80
B	17	33	20	70
합 계	54	57	39	150

$r \times c$ 분할표의 행과 열의 독립성과 동질성 검정

- 피어슨의 카이제곱 검정 통계량 :

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\text{관찰도수} - \text{기대도수})^2}{\text{기대도수}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

$$\text{여기서 } E_{ij} = \frac{i\text{번째 행합계} \times j\text{번째 열합계}}{\text{총합계}} = \frac{O_{i+} \times O_{+j}}{N}$$

- 자유도 : $df = (r-1)(c-1)$

- 유의확률 : $p-$값$= P[X^2 > \chi^2] < \alpha \Rightarrow$ 귀무가설 H_0 : 서로 독립(서로 동질)을 유의수준 α에서 기각한다.

[R-코드 1.5.1] **[보기 1.5.1] 고객들의 나이와 약의 종류에 대한 독립성 검정**

```
row1 = c(38,79)
row2 = c(87,118)
row3 = c(78,89)
data.table = rbind(row1,row2,row3)
chisq.test(data.table)
```

H_0: 먹는 약의 종류의 선호도와 나이는 관련이 없다

```
    Pearson's Chi-squared test

data:  data.table
X-squared = 5.8608, df = 2, p-value = 0.05338
```

[R-코드 1.5.2] **[보기 1.5.2] 두 식이요법 간의 동질성 검정**

```
row1 = c(37,24, 19)
row2 = c(17, 33, 20)

data.table = rbind(row1,row2)
chisq.test(data.table)
```

H_0: 두 식이요법 간에는 차이가 없다

```
        Pearson's Chi-squared test

data:  data.table
X-squared = 8.224, df = 2, p-value = 0.01638
```

Karl Pearson 1857.3.27-1936.4.27	History of Life
  Egon Pearson:1895.8.11-1980.6.12	He established the discipline of mathematical statistics. In 1911 he founded the world's first university statistics department at University College London. He was a proponent of eugenics, and a protégé and biographer of Sir Francis Galton. He was also a socialist and finally adopted Karl - supposedly also after Karl Marx. Egon Pearson became an eminent statistician himself, and created Neyman-Pearson statistics. He succeeded his father as head of the Applied Statistics Department at University College London. Statistics: Mean, S.D, Correlation coefficient, Chi-square distribution, Parameter

- Fisher(1920). Statistical Methods for Research Workers- Pearson에게 chi-square 표를 다시 작성하는 것에 대한 허락을 구했지만 거절당함 : 출판과 기금 문제
- Fisher의 Tables for Statisticians and Biometricians의 판매 효과를 두려함.

1.6 선형모형: 단순회귀모형, 다중회귀모형

단순회귀모형(simple regression model)

Francis Galton: 1822-1911

$$y = \beta_0 + \beta_1 x + \epsilon$$

y: 반응변수(종속변수)

x: 설명변수(독립변수)

ϵ: 오차항으로 평균이 0이고 분산이 σ^2인 정규분포를 따르는 확률변수

다윈의 조카로 유전학

- 정규분포의 혼합이 다시 정규분포
- 부자의 키에 대한 선형 관계 :Regression Toward Mediocrity in Hereditary Stature

적합된 회귀식(fitted regression equation)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$: 기울기 β_1의 최소제곱추정량

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$: 절편β_0의 최소제곱추정량

단순회귀모형의 타당성 검정을 위한 분산분석표

요인 (Source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	F 값 (F value)	$p-$값 ($p-$value)
모형 (Model)	1	SSR	$MSR = SSR$	$F = MSR/MSE$	$P(F \geq f)$
잔차 (Error)	$n-2$	SSE	$MSE = SSE/(n-2)$		
전체 (Total)	$n-1$	SST			

단순회귀모형의 타당성 검정

$y = \beta_0 + \beta_1 x + \epsilon$에서 $H_0 : \beta_1 = 0$의 검정 :

$f > F_\alpha(1, n-2) \Rightarrow$ 귀무가설을 기각하여 단순회귀모형이 적합함을 나타낸다.

여기서, f는 $F = MSR/MSE$의 관측 값이다.

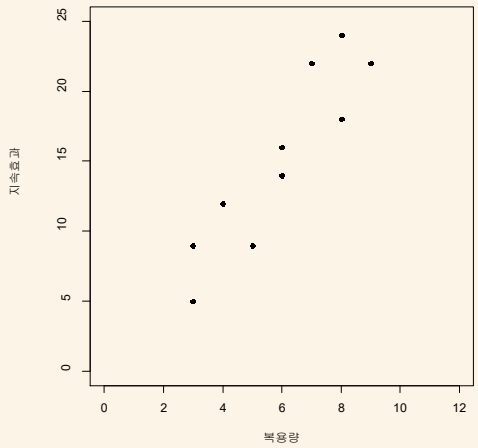
$\Leftrightarrow$ $p-$값 $< \alpha = 0.05$ $\Rightarrow$ 유의수준 5%에서 귀무가설을 기각한다.

[보기 1.6.1] 약의 복용량(x)과 지속효과(y)

10명의 환자에 대한 약의 복용량(x)과 지속효과(y)에 대한 자료의 적합된 회귀식 $\hat{y} = -1.07 + 2.74x$의 유의성 검정을 위한 분산분석표를 작성하고 해석하라.

환 자	1	2	3	4	5	6	7	8	9	10
약의 복용량 x	3	3	4	5	6	6	7	8	8	9
약의 지속효과 y	9	5	12	9	14	16	22	18	24	22

1) 약의 복용량(x)과 약의 지속효과 기간(y)에 대한 산점도 $\Rightarrow$ 단순회귀모형 $y = \beta_0 + \beta_1 x + \varepsilon$을 가정



환자	x	y	x^2	y^2	xy	$\hat{y}$
1	3	9	9	81	27	7.15
2	3	5	9	25	15	7.15
3	4	12	16	144	48	9.89
4	5	9	25	81	45	12.63
5	6	14	36	196	84	15.37
6	6	16	36	256	96	15.37
7	7	22	49	484	154	18.11
8	8	18	64	324	144	20.85
9	8	24	64	576	192	20.85
10	9	22	81	484	198	23.59
합계	59	151	389	2651	1003	

2) 다음의 절차에 따라 적합된 회귀식을 계산할 수 있다.	3) β_0와 β_1의 최소제곱추정량과 적합된 회귀식을 구한다.
$\bar{x} = 5.9,\ \bar{y} = 15.1$ $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 389 - 10(5.9)^2 = 40.9$ $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 2651 - 10(15.1)^2 = 370.9$ $S_{xy} = \sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y} = 1003 - 10(5.9 \times 15.1) = 112.1$	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{112.1}{40.9} = 2.74,$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 15.1 - 2.74 \times 5.9 = -1.07,$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -1.07 + 2.74x$

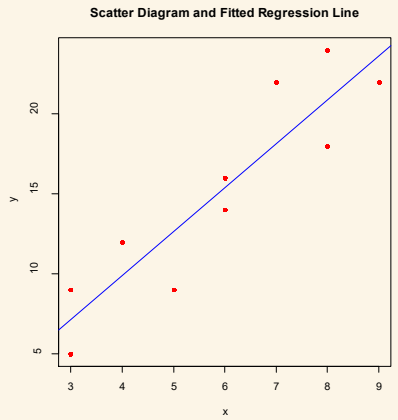
4) 단순회귀모형의 타당성 검정을 위한 분산분석표

요인 (Source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	F 값 (F value)
모형 (Model)	1	307.25	307.25	38.62
잔차 (Error)	8	63.65	7.96	
전체 (Total)	9	970.90		

5) F = 38.62 > $F_{0.05}(1,8) = 5.32 \Longrightarrow H_0 : \beta_1 = 0$를 유의수준 0.05에서 기각. $\Longrightarrow$ 적합된 회귀직선은 유의하다.

[R-코드 1.6.1] [보기 1.6.1] 약의 복용량(x)과 지속효과(y)

```
x=c(3, 3, 4, 5, 6, 6, 7, 8, 8, 9)
y=c(9, 5, 12, 9, 14, 16, 22, 18, 24, 22)
plot(x, y, col="blue", pch=16)
lsm=lm(y~x)
summary(lsm)
plot(x, y, col="red", pch=16, main="Scatter Diagram and Fitted Regression Line")
abline(coef(lsm), col="blue")
anova(lsm)
```




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0709     2.7509  -0.389 0.707219
x             2.7408     0.4411   6.214 0.000255 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.821 on 8 degrees of freedom
Multiple R-squared:  0.8284,    Adjusted R-squared:  0.8069
F-statistic: 38.62 on 1 and 8 DF,  p-value: 0.0002555
```

```
> anova(lsm)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 307.247 307.247  38.615 0.0002555 ***
Residuals  8  63.653   7.957
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

다중회귀모형(multiple regression model)

$$y = b_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon$$

y: 반응변수(종속변수)

$x_1, x_2, \ldots, x_p$: 설명변수(독립변수)

ϵ: $N(0, \sigma^2)$를 따르는 확률변수

적합된 회귀식(fitted regression equation)과 제곱합 분해

$$y = \hat{b_0} + \hat{b_1} x_1 + \cdots + \hat{b_p} x_p$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

$$\Leftrightarrow \quad SST = SSR + SSE$$

총제곱합 회귀제곱합 잔차제곱합

다중회귀모형의 타당성 검정을 위한 분산분석표

요인 (Source)	자유도 (df)	제곱합 (SS)	평균제곱 (MS)	F 값 (F value)	$p-$값 ($p-$value)
모형 (Model)	p	SSR	$MSR = SSR/p$	$F = MSR/MSE$	$P(F \geq f)$
잔차 (Error)	$n-p-1$	SSE	$MSE = SSE/(n-p-1)$		
전체 (Total)	$n-1$	SST			

다중회귀모형의 타당성 검정

$H_0 : b_1 = \cdots = b_p = 0$ vs. $H_1 : b_1 \neq 0$, 또는, ..., $b_p \neq 0$:

$f > F_\alpha(p, n-p-1) \Rightarrow$ 귀무가설을 기각하여 다중회귀모형이 적합.

여기서, f는 $F = MSR/MSE$의 관측 값.

$\Leftrightarrow$ $p-$값 $< \alpha = 0.05$ $\Rightarrow$ 유의수준 5%에서 귀무가설을 기각.

[보기 1.6.2] 심장 박동수 자료의 다중 회귀분석

[표 1.1.1](pulse.txt)은 50미터 달리기(1=달리기를 한 사람, 2= 달리기를 하지 않은 사람) 전후 심장 박동수가 흡연(1=흡연자, 2=비흡연자), 성별(1=남자, 2=여자), 키(cm)와 몸무게(kg), 운동량(1=적음, 2=보통, 3=많음)과 어떤 관계가 있는 지에 대한 10명의 성인을 대상으로 측정된 자료이다. 여기에서는 달리기 전후 심장 박동수의 차이 y에 영향을 주는 요인이 무엇인지를 알아보고자 다중회귀모형을 설정하고 적합하라. (심장박동수자료-다중회귀모형.R)의 실행결과를 참고 하라.

[1] 다중회귀분석의 해석을 쉽게 하도록 질적변수의 입력 값을 조정하고 달리기 전후 심장 박동수에 대한 변수를 재설정하자.

- 달리기, 흡연, 성별의 자료 값이 2로 되어 있는 것을 0으로 하자:
  달리기(1=달리기를 한 사람, 0= 달리기를 하지 않은 사람)
  흡연(1=흡연자, 0=비흡연자), 성별(1=남자, 0=여자),
- 운동량은 적음과 보통을 0으로 많음을 1로 하자.
- y = 심장박동수 차이(심장박동수 후 - 심장박동수 전)

[2] 다중회귀모형의 설정:

$$y = b_0 + b_1\text{달리기} + b_2\text{흡연} + b_3\text{성별} + b_4\text{운동량} + b_5\text{키} + b_6\text{몸무게} + \varepsilon$$

[3] 모형의 유의성: $R^2 = 0.976$이고 검정 통계량의 관측값 $f = 20.37$는 유의확률 $p-$값 = 0.01577에 의해서 유의수준 5%에서 모형이 매우 적합함을 보여 준다. [결과]의 F-statistic, p-value를 참고 바람.

[4] 적합된 다중회귀모형과 해석:

$$\hat{y} = 98.023 + 18.587\text{달리기} + 0.870\text{흡연} + 6.520\text{성별} + 2.044\text{운동량} - 0.518\text{키} - 0.215\text{몸무게}$$

- 달리기를 한 경우(1)가 하지 않은 경우(0)에 비해 $\hat{b_1}$=18.587만큼 더 큰 심장 박동수 차이를 보였다. 이 변수는 귀무가설 $H_0 : b_1 = 0$를 유의수준 5%에서 기각하기에 충분한 $p-$값 = 0.0117로 모형 [2]에서 통계적으로 유의하다.

- 흡연자(1)가 비흡연자(0)보다 $\hat{b_2}$ = 0.870 만큼 큰 반응을 보였으나 $p-$값 0.7451로 통계적으로 유의하지 않다.

- 남자(성별=1)는 여자(성별=0)에 비해 $\hat{b_3}$ = 6.520 만큼 더 큰 반응을 보였으나 $p-$값 = 0.1867로 통계적으로 유의하지 않다.

- 운동량이 많은 경우(1)는 적은 경우(0)에 비해 $\hat{b_4}$ = 2.044 만큼 더 큰 반응을 보였고 $p-$값 = 0.4603으로 통계적으로 유의하지 않다.

- 키가 크고 몸무게가 많이 나갈수록 반응이 작아지고 특히, 키는 통계적으로 유의하다.

[R-코드 1.6.2] [보기 1.6.2] 심장 박동수 자료의 다중회귀분석

```
##[보기 1.6.2] 심장 박동수 자료의 다중회귀분석
setwd("c:/기초통계학/R-codes")
data1.6.2<-read.table("pulse.txt", header=T)
x<-data1.6.2

## 변수 변환
y=x$pulse2-x$pulse1
y
키=x$height
몸무게=x$weight
달리기=as.numeric(x$ran<=1)
흡연=as.numeric(x$smokes<=1)
성별=as.numeric(x$gender<=1)
운동량=as.numeric(x$activity>=3)

## 다중 회귀모형의 적합
model1=lm(y~달리기+흡연+성별+운동량+키+몸무게)
summary(model1)
model2=lm(y~달리기+키)
summary(model2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.0231    25.2427   3.883   0.0303 *
달리기       18.5873     3.3654   5.523   0.0117 *
흡연          0.8700     2.4408   0.356   0.7451
성별          6.5201     3.8187   1.707   0.1863
운동량        2.0435     2.4190   0.845   0.4603
키           -0.5182     0.1514  -3.422   0.0418 *
몸무게       -0.2153     0.1118  -1.927   0.1496
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.579 on 3 degrees of freedom
Multiple R-squared:  0.976,     Adjusted R-squared:  0.9281
F-statistic: 20.37 on 6 and 3 DF,  p-value: 0.01577
```

참고문헌

최용석(2014). R과 함께하는 통계학의 이해, 교보문고.

최용석(2018). R과 함께하는 다변량 자료분석, 경문사.

김용일 ◦ 최용석(2018). 치의학을 위한 통계적 방법과 응용 - R과 함께하는 의학통계-, 교우사.

Johnson, R.A. and Bhattacharyya, G.K.(2006). *Statistics*, John Wiley & Sons, Inc., Danvers.