21장 분할표 분석과 적합도 검정

21.1 분할표 분석

21.1.1 카이제곱 검정

chisq.test() 함수는 분할표 자료에 대해 적합도(goodness of fit) 검정과 독립성(independence) 및 동일성 검정(test of homogeneity)을 수행한다.

(a) 적합도 검정

적합도 검정은 각 범주에서 관측된 빈도가 특정 비율을 따르는지에 대한 검정을 수행한다.

・p= 옵션으로 확률을 지정할 수 있다.

예를 들어, 4개의 범주를 가지는 일차원 분할표에서 관측값이 25, 32, 18, 20 일 때, 각 범주에 속할 확률이 같은지에 대한 적합도 검정은 다음과 같다.

```
> chisq.test(c(25,32,18,20))      # 디폴트로 p=c(1/4,1/4,1/4,1/4) 옵션이 사용됨

        Chi-squared test for given probabilities

data:  c(25, 32, 18, 20)
X-squared = 4.9158, df = 3, p-value = 0.1781
```

(해석) 유의수준 5%에서 확률이 같다는 귀무가설을 기각할 수 없다.

[예제 1] HairEyeColor 자료를 이용하여 카이제곱 적합도 검정을 수행한다. 어느 생리학자가 눈의 색깔이 Brown 50%, Blue 25%, Hazel 15%, Green 10% 라고 주장할 때, 이 주장이 타당한지를 검정하고자 한다.

```
> HairEyeColor
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
```

```
  Blond      3   30      5      8

, , Sex = Female

       Eye
Hair     Brown Blue Hazel Green
  Black     36    9      5      2
  Brown     66   34     29     14
  Red       16    7      7      7
  Blond      4   64      5      8
```

```
> Eye <- margin.table(HairEyeColor, 2)
> Eye
Eye
Brown  Blue Hazel Green
  220   215    93    64
```

```
> chisq.test(Eye, p=c(.5, .25, .15, .1))

        Chi-squared test for given probabilities

data:  Eye
X-squared = 50.4324, df = 3, p-value = 6.462e-11
```

(해석) p-값이 거의 0이므로, 위 생리학자의 주장을 기각할 만한 통계적 근거가 충분하다고 말할 수 있다.

□

(b) 독립성 검정

독립성 검정은 이원분할표를 구성하는 두 개의 범주형 변수가 서로 독립인지에 대한 검정을 수행한다.

[예제 2] HairEyeColor 자료에서 Hair와 Eye 변수가 서로 독립인지를 검정한다.

```
> HairEye <- margin.table(HairEyeColor, c(1,2))
> HairEye
       Eye
Hair     Brown Blue Hazel Green
  Black     68   20    15     5
  Brown    119   84    54    29
```

```
  Red       26   17    14    14
  Blond      7   94    10    16

> chisq.test(HairEye)

        Pearson's Chi-squared test

data:  HairEye
X-squared = 138.2898, df = 9, p-value < 2.2e-16
```

(해석) p-값이 매우 작으므로 머리와 눈의 색깔 간에는 독립이라는 가설을 기각할 만한 근거가 충분하다고 말할 수 있다.

□

Fisher의 정확 검정

이원 분할표(two-way contingencey table)에서 칸 도수가 크지 않을 때는 카이제곱을 통한 근사적 검정보다 피셔의 정확검정(Fisher's Exact Test)이 보다 정확한 결과를 제공한다. R의 fisher.test() 함수는 피셔의 정확검정을 제공한다.

fisher.test(): 통상적으로 2×2 분할표에 국한되어 수행되나, R에서는 칸 도수가 크지 않은 경우에는 보다 큰 테이블($r\times c$ 이원 분할표)에 대해서도 적용된다.

(c) 동일성 검정

두 집단 간에 각 범주에 속할 비율이 같은지에 대한 검정을 수행한다.

[예제 3] HairEyeColor 자료에서 남성과 여성 간에 눈의 색의 비율이 같은지를 검정하고 자 한다.

```
> a <- margin.table(HairEyeColor, c(3,2))
> a
        Eye
Sex      Brown Blue Hazel Green
  Male      98  101    47    33
  Female   122  114    46    31

> chisq.test(a)   # 독립성 검정의 경우와 같음

        Pearson's Chi-squared test
```

data: a
X-squared = 1.5298, df = 3, p-value = 0.6754

(해석) p-값이 매우 크므로, 남성과 여성 간에 눈의 색깔에 차이가 없다는 귀무가설을 기각할 만한 증거가 불충분하다.

□

21.1.2 분할표 만들기

데이터프레임 또는 행렬(배열) 자료로부터 분할표를 작성하는데 유용한 함수는 다음과 같다.

- xtabs(formula= ~., data=parent.frame(), subset, sparse=FALSE,
  na.action, exclude=c(NA, NaN), drop.unused.levels=FALSE)
- ftable(): 평면 분할표(flat contingency table) 제공
- CrossTable{gmodels}

```
> ## 분할표 작성 예제: xtabs()와 ftable() 함수 이용
> str(esoph)
'data.frame':   88 obs. of  5 variables:
 $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1 1 1 1 1 1 1 1 ...
 $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1 1 1 1 2 2 2 2 3 3 ...
 $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4 1 2 ...
 $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 ...

> ## xtabs() 함수 적용
> xtabs(cbind(ncases, ncontrols) ~ ., data=esoph)
> ## xtabs() 함수의 결과는 "xtabs" 또는 "table" 객체임
, , tobgp = 0-9g/day,  = ncases

       alcgp
agegp   0-39g/day 40-79 80-119 120+
  25-34         0     0      0    0
  35-44         0     0      0    2
  45-54         1     6      3    4
  55-64         2     9      9    5
  65-74         5    17      6    3
  75+           1     2      1    2

, , tobgp = 10-19,  = ncases

       alcgp
agegp   0-39g/day 40-79 80-119 120+
  25-34         0     0      0    1
  35-44         1     3      0    0
  45-54         0     4      6    3
  55-64         3     6      8    6
  65-74         4     3      4    1
  75+           2     1      1    1
```

(생략)

, , tobgp = 30+, = ncontrols

agegp	alcgp 0-39g/day	40-79	80-119	120+
25-34	5	7	2	2
35-44	8	8	1	0
45-54	4	7	4	4
55-64	6	6	4	6
65-74	2	0	1	1
75+	3	1	0	0

> ## ftable(xtabs()) 함수 적용: 보다 나은 형태의 출력을 제공
> ftable(xtabs(cbind(ncases, ncontrols) ~ ., data = esoph))

			ncases	ncontrols
agegp	alcgp	tobgp		
25-34	0-39g/day	0-9g/day	0	40
		10-19	0	10
		20-29	0	6
		30+	0	5
	40-79	0-9g/day	0	27
		10-19	0	7
		20-29	0	4
		30+	0	7
	80-119	0-9g/day	0	2
		10-19	0	1
		20-29	0	0
		30+	0	2
	120+	0-9g/day	0	1
		10-19	1	1
		20-29	0	1
		30+	0	2
35-44	0-39g/day	0-9g/day	0	60
		10-19	1	14
		20-29	0	7
		30+	0	8

(생략)

75+	0-39g/day	0-9g/day	1	18
		10-19	2	6
		20-29	0	0
		30+	1	3
	40-79	0-9g/day	2	5

	10-19	1	3
	20-29	0	3
	30+	1	1
80-119	0-9g/day	1	1
	10-19	1	1
	20-29	0	0
	30+	0	0
120+	0-9g/day	2	2
	10-19	1	1
	20-29	0	0
	30+	0	0

```
> ## ftable(xtab()) 함수: 더 작은 요인에 대해 적용
> ftable(xtabs(cbind(ncases, ncontrols) ~ agegp, data=esoph))
       ncases ncontrols
agegp
25-34       1        116
35-44       9        199
45-54      46        213
55-64      76        242
65-74      55        161
75+        13         44

> ## xtabs() 함수 적용 예: 배열 형태의 자료에 대해 적용
> DF <- as.data.frame(UCBAdmissions)
> DF
      Admit Gender Dept Freq
1  Admitted   Male    A  512
2  Rejected   Male    A  313
3  Admitted Female    A   89
4  Rejected Female    A   19
5  Admitted   Male    B  353
(생략)
23 Admitted Female    F   24
24 Rejected Female    F  317

> xtabs(Freq ~ Gender + Admit, DF)
         Admit
Gender    Admitted Rejected
  Male        1198     1493
  Female       557     1278
```

```
> summary(xtabs(Freq ~ ., DF))     # 독립성 검정 결과 제공
Call: xtabs(formula=Freq ~ ., data=DF)
Number of cases in table: 4526
Number of factors: 3
Test for independence of all factors:
        Chisq = 2000.3, df = 16, p-value = 0


> ## 분할표 작성 예제: CrossTable{gmodels} 함수 이용
> str(infert)
# infert 자료는 인공유산(induced), 자연유산(spontaneous) 후의 출산아수(parity)를 대응 사례-대조(matched case-control) 연구로 조사한 자료임
'data.frame':   248 obs. of  8 variables:
 $ education     : Factor w/ 3 levels "0-5yrs","6-11yrs",..: 1 1 1 1 2 2 2 2 2 2 ...
 $ age           : num  26 42 39 34 35 36 23 32 21 28 ...
 $ parity        : num  6 1 6 4 3 4 1 2 1 2 ...
 $ induced       : num  1 1 2 2 1 2 0 0 0 0 ...
 $ case          : num  1 1 1 1 1 1 1 1 1 1 ...
 $ spontaneous   : num  2 0 0 0 1 1 0 0 1 0 ...
 $ stratum       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ pooled.stratum: num  3 1 4 2 32 36 6 22 5 19 ...

> library(gmodels)
> data(infert, package="datasets")
>  CrossTable(infert$education, infert$induced, expected=TRUE, dnn=c("Education", "Induced"))     # dnn= 옵션은 결과에서 차원이름을 정해줌(dimension names)

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

Total Observations in Table:  248

             | Induced
   Education |         0 |         1 |         2 | Row Total |
-------------|-----------|-----------|-----------|-----------|
      0-5yrs |         4 |         2 |         6 |        12 |
             |     6.919 |     3.290 |     1.790 |           |
             |     1.232 |     0.506 |     9.898 |           |
```

```
             |     0.333 |     0.167 |     0.500 |     0.048 |
             |     0.028 |     0.029 |     0.162 |           |
             |     0.016 |     0.008 |     0.024 |           |
-------------|-----------|-----------|-----------|-----------|
     6-11yrs |        78 |        27 |        15 |       120 |
             |    69.194 |    32.903 |    17.903 |           |
             |     1.121 |     1.059 |     0.471 |           |
             |     0.650 |     0.225 |     0.125 |     0.484 |
             |     0.545 |     0.397 |     0.405 |           |
             |     0.315 |     0.109 |     0.060 |           |
-------------|-----------|-----------|-----------|-----------|
     12+ yrs |        61 |        39 |        16 |       116 |
             |    66.887 |    31.806 |    17.306 |           |
             |     0.518 |     1.627 |     0.099 |           |
             |     0.526 |     0.336 |     0.138 |     0.468 |
             |     0.427 |     0.574 |     0.432 |           |
             |     0.246 |     0.157 |     0.065 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |       143 |        68 |        37 |       248 |
             |     0.577 |     0.274 |     0.149 |           |
-------------|-----------|-----------|-----------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  16.53059     d.f. =  4     p =  0.002383898

Warning message:
In chisq.test(t, correct = FALSE, ...) :
  Chi-squared approximation may be incorrect
```

```
> ## 아래 명령어도 유사한 결과를 제공함
> CrossTable(infert$education, infert$induced, expected=TRUE, format="SAS")
> CrossTable(infert$education, infert$induced, expected=TRUE, format="SPSS")
```

21.2 분포에 대한 적합도 검정

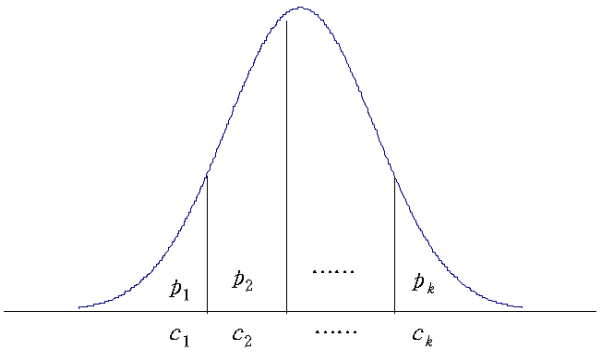
분포함수에 대한 적합도 검정은 데이터가 특정한 분포로 나왔는지(일표본 적합도 검정), 또는 두 자료 집단의 분포함수가 같은지를 검정한다(이표본 적합도 검정). 가장 널리 사용되는 적합도 검정법으로는 카이제곱 적합도(Chi-square Goodness of Fit) 검정(모수적 방법)과 콜모고로프-스미르노프(Kolmogorov-Smirnov) 검정(비모수적 방법)이 있다.

21.2.1 카이제곱 적합도 검정

카이제곱 적합도 검정은 적당히 나누어진 k개의 각 구간에 포함되는 자료 수(O_i)와 특정 귀무분포하에서의 기대도수의 추정치($\hat{E}_i$)와의 차이에 기초한 검정 방법이다. 피어슨(Pearson)에 의한 카이제곱 검정통계량은 다음과 같다.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

카이제곱 검정은 대표본 검정이므로, 표본의 수가 적을 때는 결과에 대한 신뢰도가 떨어짐에 유의하여야 한다.



범주	C_1	C_2	…	C_k	합계
관측도수	O_1	O_2	…	O_k	n
기대도수	E_1	E_2	…	E_k	n

[그림 21.1] 카이제곱 적합도 검정

카이제곱 적합도 검정은 gofTest{EnvStats} 함수를 이용한다. 이 함수의 일반 형식은 다음과 같다.

```
gofTest(y, x=NULL, test=, distribution="norm", alternative="two.sided",
        param.list=NULL, estimate.params=, n.param.est=NULL,
        exact=, correct=, ...)
```

• test= "sw"(x가 생략 시 디폴트임: Shapiro-Wilk), "chisq"(카이제곱 검정),
"ks"(Kolmogorov-Smirnov 검정), ...
• distribution= "norm"(디폴트), "lnorm", "gamma" 등 지정 가능
• param.list= 지정된 분포의 모수값을 지정. 지정 방법은 도움말을 이용할 것
• estimate.params= 모수의 추정을 할 것인지의 여부를 지정. TRUE(param.list=NULL인 경우) 또는 FALSE.
• correct= FALSE(디폴트). TRUE는 "chisq" 검정 시 연속성 수정의 결과를 제시함

[예제 4] gofTest() 함수를 이용하여 적합도 검정을 수행한다.

```
> library(EnvStats)
> set.seed(1020)
> x <- rexp(50, rate=1.0)   # 지수분포로부터 난수 발생
> gofTest(x, test="chisq")

Results of Goodness-of-Fit Test
-------------------------------
Test Method:                     Chi-square GOF
Hypothesized Distribution:       Normal
Estimated Parameter(s):          mean = 1.383150
                                 sd   = 1.301921
Estimation Method:               mvue
Data:                            x
Sample Size:                     50
Test Statistic:                  Chi-square = 20.8
Test Statistic Parameter:        df = 7
P-value:                         0.004077716
Alternative Hypothesis:          True cdf does not equal the
                                 Normal Distribution.
-------------------------------------------------
```

(해석) p-값이 매우 작으므로 데이터가 정규분포를 따른다는 귀무가설을 기각한다.

□

참고 위의 결과에서 제공되는 적합분포의 모수 추정에는 fitdistr{MASS} 함수를 이용할 수도 있다. 이 함수는 다양한 분포에 대해 추정치, 표준오차, 분산-공분산 행렬, 로그-가능도 등을 제공해 준다.

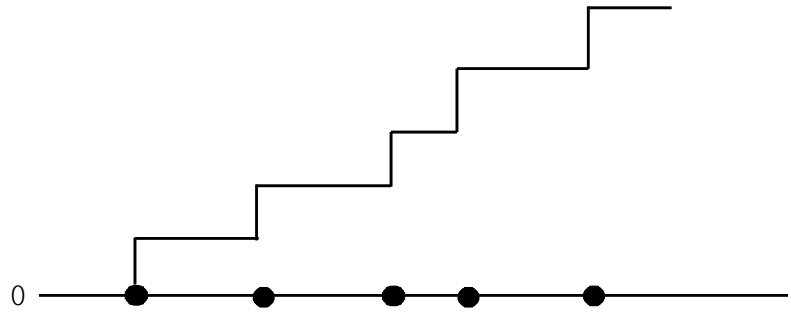
```
> library(MASS)
> fitdistr(x, "normal")
```

```
     mean          sd
 1.3831500    1.2888358     # 위의 결과와 유사함
(0.1822689) (0.1288836)
```

21.2.2 콜모고로프-스미르노프 검정

경험분포함수

경험분포함수(empirical distribution function)는 크기 n인 자료에서 각 관측값에 균등한 확률($1/n$)을 부여한 뒤, 이 확률분포로부터 누적분포함수를 구한 것이다. 경험분포함수를 그림으로 나타내면 다음과 같다.

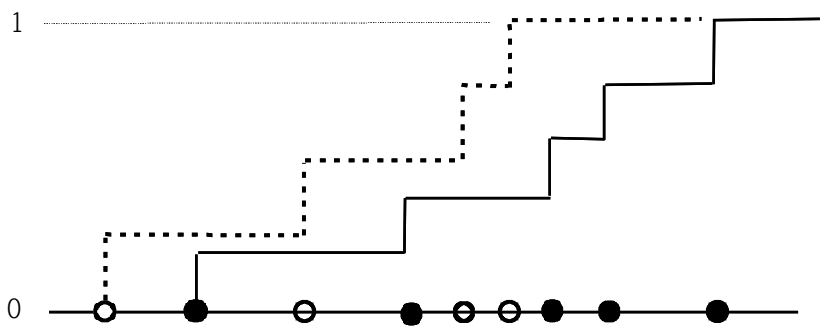


[그림 21.2] 경험분포함수(일표본의 경우)

콜모고로프-스미르노프(Kolmogorov-Smirnov) 검정은 일표본과 이표본 문제에 적용이 가능하며, 일표본의 경우는 양측(two-sided) 및 단측(one-sided) 검정이 모두 가능하나, 이표본의 경우는 양측 검정만이 가능하다(카이제곱검정은 일표본 양측검정만 가능). 콜모고로프-스미르노프 검정통계량은 두 경험분포함수의 차이(이표본의 경우) 또는 경험분포함수와 특정 모분포함수의 차이(일표본의 경우)의 최댓값인

$$T = \sup_x |F_1(x) - F_2(x)|$$

의 형태를 취한다. 위에서 F_1, F_2는 모분포함수 또는 경험분포함수이다. 예를 들어, 이표본의 경우 두 자료셋으로부터 구해지는 경험분포함수는 다음 그림과 같다.



[그림 21.3] 경험분포함수(이표본의 경우)

일표본 문제에서 모분포의 디폴트는 정규분포이며, 평균과 분산은 각각 mean(z)과 sqrt(var(z))으로 추정된다.

ks.test() 함수는 일표본과 이표본의 콜모고로프-스미르노프 검정을 수행한다. ks.test() 함수의 일반형식은 다음과 같다.

```
ks.test(x, y, ...,
        alternative=c("two.sided", "less", "greater"),
        exact=NULL)
```

[예제 5] ks.test() 함수를 이용하여 분포함수에 대한 검정을 수행한다.

```
> require(graphics)
> x <- rnorm(50)
> y <- runif(30)
```

(i) x와 y가 같은 분포로부터 나왔는지를 검정

```
> ks.test(x, y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.48, p-value = 0.0002033
alternative hypothesis: two-sided
------------------------------------
```

(해석) p-값이 매우 작으므로 분포가 같다는 귀무가설을 기각할만한 충분한 근거가 있다고 말할 수 있다.

(ii) x+2가 Gamma(3, 2) 분포로부터 나왔는지를 검정

```
> ## 양측검정; 정확한 p-값 계산
> ks.test(x+2, "pgamma", 3, 2)
# pgamma는 감마분포의 누적분포함수를 나타냄. 유사한 방법으로 연속형 분포를 지정할 수 있음.

        One-sample Kolmogorov-Smirnov test

data:  x + 2
D = 0.3292, p-value = 2.507e-05
alternative hypothesis: two-sided
--------------------------------
```

(해석) p-값이 매우 작으므로 데이터가 Gamma(3, 2) 분포를 따른다는 귀무가설을 기각할만한 충분한 근거가 있다.

```
> ## 양측 검정; 근사적 p-값 제시
> ks.test(x+2, "pgamma", 3, 2, exact=FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x + 2
D = 0.3292, p-value = 3.94e-05
alternative hypothesis: two-sided
-----------------------------------
```

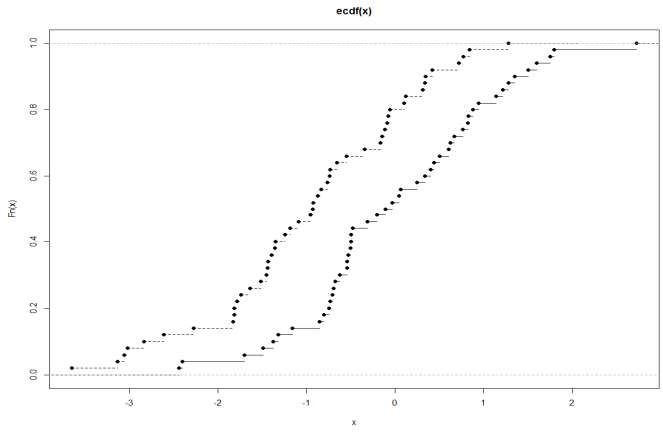
(해석) exact=TRUE(디폴트)가 사용된 위의 결과와 유사함

```
> ## 단측 검정
> ks.test(x+2, "pgamma", 3, 2, alternative="gr")

        One-sample Kolmogorov-Smirnov test

data:  x + 2
D^+ = 0.067, p-value = 0.6112
alternative hypothesis: the CDF of x lies above the null hypothesis
-----------------------------------------------------------
```

(해석) 유의수준 5%에서 데이터로부터 구한 경험분포함수가 Gamma(3, 2) 분포의 누적분포함수보다 함수적으로 위쪽에 놓여있다고 말할 수 있다. 여기서 주의할 점은 분포함수가 위쪽에 있다는 말은 확률적으로는 작다는 것(stochastically less than)과 같은 의미이다.

(iii) x가 x2보다 확률적으로 큰지(stochastically larger than)를 검정

```
> x2 <- rnorm(50, -1)
> plot(ecdf(x), xlim=range(c(x, x2)))
> plot(ecdf(x2), add=TRUE, lty="dashed")
```




```
> ## 모수적 검정: t-검정
> t.test(x, x2, alternative="g")
```

```
        Welch Two Sample t-test

data:  x and x2
t = 5.1152, df = 97.871, p-value = 7.802e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6458135       Inf
sample estimates:
  mean of x   mean of y
-0.01226031 -0.96850472
--------------------------------
```

(해석) 유의수준 5%에서 귀무가설을 기각하므로, x가 x2보다 확률적으로 크다고 할 수 있다.

```
> ## 비모수적 검정: 윌콕슨 순위합 검정
> wilcox.test(x, x2, alternative="g")

        Wilcoxon rank sum test with continuity correction

data:  x and x2
W = 1922, p-value = 1.835e-06
alternative hypothesis: true location shift is greater than 0
-----------------------------------------------
```

(해석) 유의수준 5%에서 귀무가설을 기각하므로, x가 x2보다 확률적으로 크다고 할 수 있다.
이 경우 모수적 검정(t-검정)과 비모수적 검정(윌콕슨의 순위합 검정)의 결과가 유사하다.

```
> ks.test(x, x2, alternative="l")

        Two-sample Kolmogorov-Smirnov test

data:  x and x2
D^- = 0.46, p-value = 2.542e-05
alternative hypothesis: the CDF of x lies below that of y
-----------------------------------------------
```

(해석) 유의수준 5%에서 x의 분포함수가 x2의 분포함수보다 아래쪽에 있다고 말할 수 있다.

□

21장 연습문제

1. 어느 도시에서 일 년 동안 요일에 따라 발생한 중범죄 건수가 다음과 같이 조사되었다.

월	화	수	목	금	토	일
62	71	56	65	67	82	73

(a) 범죄가 모든 요일에 관계없이 같은 비율로 일어났는지를 검정하여라.

(b) 범죄가 주중(월~금)에 같은 비율로 일어났는지를 검정하여라.

2. 다음 표는 부모의 안전벨트 착용 여부와 자녀의 착용여부에 대한 조사표이다. 두 변수가 서로 독립인지에 대한 검정을 수행하여라.

	어린이	
부모	착용	미착용
착용	72	18
미착용	6	32

3. 20명의 학생을 임의로 10명씩 나누어, 한 그룹을 시력교정 프로그램에 참여시킨 후 시력의 상태를 조사하였다. 두 그룹의 분포가 같은지에 대한 검정을 수행하여라.

프로그램	많이 나빠짐	나빠짐	같음	좋아짐	많이 좋아짐
비참가	0	2	8	2	0
참가	0	2	3	5	2

4. morley 자료는 빛의 속도를 측정한 자료이다. 총 100개의 빛의 속도 자료가 정규분포를 따르는지를 알아보고자 한다.

(a) 카이제곱 적합도 검정을 수행하여라.

(b) 콜모고로프-스미르노프 검정을 수행하여라.

5. ToothGrowth 자료에서 비타민 C의 종류에 따른 이빨의 길이가 같은 분포를 따르는지를 알아보고자 한다.

(a) 카이제곱 적합도 검정을 수행하여라.

(b) 콜모고로프-스미르노프 검정을 수행하여라.