7. 통계분석

7.1 난수 생성 및 분포함수

R은 주어진 통계 분포를 따르는 난수를 발생시키는 다양한 함수를 제공한다. 이 함수들은 난수 (random)을 뜻하는 r 뒤에 분포명을 붙인 형태다. 다음 표에 이항 분포, F 분포, 기하 분포 등 주요 분포에 대한 함수를 정리했다.

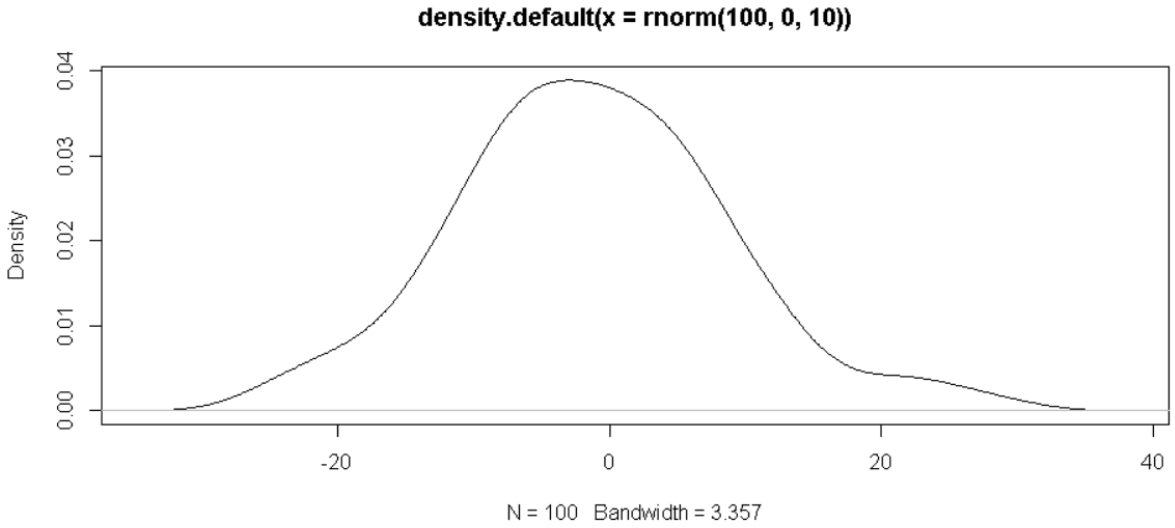
확률 분포	난수 발생 함수
이항(Binomial) 분포	rbinom
F 분포	rf
기하(Geometric) 분포	rgeom
초기하(Hypergeometric) 분포	fhyper
음 이항분포	rnbinom
정규분포	rnorm
포아송분포	rpois
t 분포	rt
연속 균등 분포	runit

이 함수들의 인자는 원하는 난수의 개수와 각 확률 분포의 파라미터다. 예를 들어, 정규 분포는 평균과 표준 편차를 인자로 받는다. 다음은 평균 0, 표준 편차 10인 정규 분포로부터 난수 100개를 뽑는 예다.

```
> head(rnorm(100,0,10))
[1]  -4.2270886   3.0467152 -19.5246396
[4]   1.3650454  -0.6421270   0.0503913
```

많은 수의 난수를 만들고 밀도 그림을 그려보면 다음 그림처럼 데이터가 정규 분포를 잘 따르고 있음을 볼 수 있다.

```
plot(density(rnorm(100,0,10)))
```



확률 밀도 함수 또는 확률 질량 함수는 d 뒤에 분포명을 적는다. 예를 들어, 정규 분포의 경우 dnorm()을 사용한다.

분포 함수는 p 뒤에 분포명을 적은 형태다. 예를 들어, 정규 분포의 분포 함수는 pnorm()이다. (누적)

분위수는 q 뒤에 분포명을 적으며, 정규 분포의 경우 qnorm()으로 구한다. qnorm()이 pnorm()의 역함수에 해당하는데, p와 q가 알파벳에서 연속된 문자임을 상기하면 기억하기 쉽다.

Note_ 분포 함수와 확률 밀도 함수, 확률 질량 함수

실험에서 나타난 결과를 실수로 표현한 변수 X를 확률 변수라 한다. 예를 들어, 동전 두 개를 던졌을 때 앞면의 개수 X, 웹 페이지의 하이퍼링크가 클릭된 수 X, 강수량 X가 확류류 변수의 예이다.

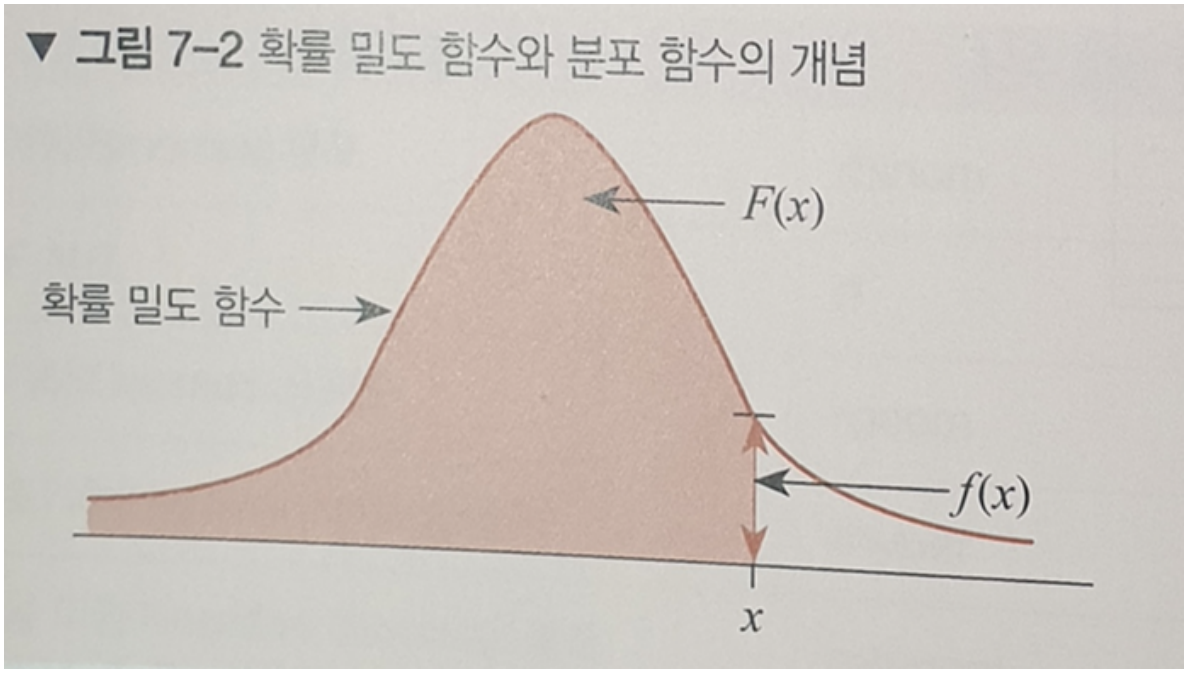
분포함수는 확률 변수의 누적 분포를 기술하는 함수로, 보통 대문자 F를 써서 F(x)로 표현한다. **누적 분포 함수 F(x)**는 확률 변수 X가 x이하의 값을 가질 확률을 뜻한다. 즉, **F(x) = P(X <= x)** 이다.

확률 밀도 함수는 연속형 데이터의 확률을 표현한다. 강수량, 키, 몸무게 등의 값이 연속형 데이터의 예이다. **확률 밀도 함수는 f(x)** 처럼 소문자로 표시한다. 그러나 연속형 데이터의 경우 x를 하나의 정확한 값으로 정하기가 곤란하다. 이에 f(x)는 구간에 대해 정의하며, 함수를 특정 구간에 대해 적분 한 값이 확률 변수 X가 그 구간에 속할 확률이 되는 함수다. 다시말해 $F(x) = \int_{-\infty}^{x} f(x)$ 이다.

x가 이산형 값들만 가진다면 특정 x에 대한 확률을 계산할 수 있다. 예를 들어, 동전 한 개를 던졌을 때 앞면의 수를 X라고 하면 P(X=1) 은 앞면, 뒷면의 두 가지 가능성 중 하나에 해당하므로 1/2이다. 이러한 P(X=x) 를 확률 질량 함수라고 하며, 흔히 보기 쉽게 p(x)로 적는다. 확률 질량 함수에서는 확률 밀도 함수와 달리 p(0), p(1) 처럼 구간이 아닌 특정 값에 대한 확률을 이야기 할 수 있다.

R에서 d로 시작하는 함수 예를 들면 (dnorm 함수)는 f(x), p(x)를 계산하며, q로 시작하는 함수는 F(x)를 계산한다.

다음으림은 확률 밀도 함수와 분포 함수의 개념도를 그린 것이다. 확률 밀도 함수의 그림이 주어졌을 때 f(x)는 x 값에 대한 확률 밀도 함수의 값 자체가 되고, F(x)는 X <= x인 모든 f(x)의 적분 값 (즉, 확률 밀도 함수 영역의 넓이)이 된다.



Note_ 분포함수와 분위수 함수의 관계

분포 함수 F(x)는 확률 밀도 함수 f(x)가 주어졌을 때 $F(x) = \int_{-\infty}^{x} f(x)$ 로 정의된다. 즉, X가 x이하의 값을 가질 확률을 뜻한다. 따라서 F(p) = q라면 x가 p보다 작은 비율이 q임을 뜻한다.

분위수 함수는 q를 주었을 때 F(p) = q인 p를 찾는다. 따라서 분포 함수의 역함수 $F^{-1}()$ 로 이해하면 된다.

이 함수들을 실제 코드로 연습해보자. 포아송 분포의 확률 질량 함수는 다음과 같다.

7.2 기초 통계량

이 절에서 설명할 통계량은 표본 평균, 분산, 표준 편차, 다섯 수치 요약, 최빈값 등이다.

7.2.1 표본 평균, 표본 분산, 표본 표준 편차

표본 평균, 표본 분산, 표본 표준 편차는 표본 $X_1, X_1, \cdots, X_n$의 n개 표본이 있을 때 다음과 같이 계산한다.

기초 통계량의 계산

통계량	수식
표본 평균	$\bar{X} = \frac{1}{n}\sum X_i$
표본 분산	$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$
표본 표준 편차	$\sqrt{s^2}$

분산의 계산에서 분모에 n이 아니라 n-1을 사용하고 있다는 점에 유의하기 바란다. R에서 기본적으로 계산하는 분산과 표준 편차는 전체 데이터 중 일부를 샘플로 추출한 뒤 이에 대해 분산과 표준 편차를 계산하는 표본 분산과 표본 표준 편차다. 따라서 n-1을 분모로 사용한다.

다음은 c(1,2,3,4,5)의 평균, 표본 분산, 표본 표준 편차를 계산한 예이다.

```
> mean(1:5)
[1] 3
> var(1:5)
[1] 2.5
> sum((1:5-mean(1:5))^2)/(5-1)
[1] 2.5
> sd(1:5)
[1] 1.581139
```

다섯 수치 요약

다섯 수치 요약은 데이터를 최솟값, 제 1 사분위수, 중앙값, 제 3 사분위수, 최댓값으로 요약한다. 다섯수치 요약을 구하는 함수는 fivenum()이다. summary()는 fivenum()과 유사하지만 다섯 수치 요약에 더해 평균까지 계산해준다.

fivenum : 다섯 수치 요약을 구한다.

다음은 c(1,2,3, ..., 11)의 다섯 수치 요약을 계산한 결과다. 최솟값이 1, 최댓값이 11, 중앙값 등을 알 수 있다.

```
> fivenum(1:11)
[1]  1.0  3.5  6.0  8.5 11.0
> summary(1:11)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     3.5     6.0     6.0     8.5    11.0
```

fivenum()과 summary()는 데이터의 크기가 홀수일 경우에는 위의 예처럼 동일한 결과를 보이지만, 짝수일 때는 다소 다른 결과를 출력한다. 다음은 c(1,2,3,4) 에 대한 다섯 수치 요약과 summary()의 출력 결과를 비교한 예다. 제 1사분위수와 제 3 사분위수가 다르다.

```
> fivenum(1:4)
[1] 1.0 1.5 2.5 3.5 4.0
> summary(1:4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.75    2.50    2.50    3.25    4.00
```

최빈값

최빈값은 데이터에서 가장 자주 나타난 값을 말한다. 최빈값은 table()을 사용해 각 데이터가 출현한 횟수를 센 분할표(Contingency Table)를 만들고, which_max()를 사용해 최대 빈도가 저장된 색인을 찾는 방법으로 구할 수 있다.

table : 분할표를 작성한다.

which.max : 최댓값이 저장된 위치의 색인을 반환한다.

다음 코드는 문자열 벡터 c("a", "b", "c", "c","c", "d","d") 로부터 가장 자주 출현한 문자열 "c"를 찾는 예다. 최빈값 자체를 구할 때는 names()를 사용해 분할표에서 각 셀의 이름을 구한 다음, 최댓값이 저장된 셀을 선택했다.

```
> (x <- factor(c("a", "b", "c", "c","c", "d","d")))
[1] a b c c c d d
Levels: a b c d
> table(x)
x
a b c d
```

```
1 1 3 2
> which.max(x)
[1] 6
> which.max(table(x))
c
3
> names(table(x))[3]
[1] "c"
```

7.3 표본 추출

현대의 데이터는 기하급수적으로 증가하고 있다. 아무리 분산 컴퓨팅 능력이 받쳐준다 할지라도 낭비에 가깝다. 그보다는 특정 기간에 있었던 질의만 분석한다던가, 특정 조건을 만족하는 질의만 분석하는 것이 효율적이다. 이처럼 전체 데이터(모집단 population) 중 일부를 표본 (샘플 sample) 으로 추출하는 작업은 데이터 분석에서 필수다.

표본 추출 (샘플링) 은 훈련 데이터와 데스트 데이터의 분리에서도 중요하다. 전체 데이터 중 80%를 훈련 데이터, 20%를 테스트 데이터로 분리한 뒤 데이터에 대한 모델링은 훈련 데이터로만 수행하고, 모델의 성능은 테스트 데이터로 평가하면 모델의 성능을 가장 적절히 평가할 수 있다. 데이터를 분리하지 않고 전체 데이터를 모델링과 모델 평가에 사용하게 되면 데이터에 내재하는 실제적 특징 외에 데이터에 우연히 포함된 노이즈까지 반영한 모델을 만들게 될 위험이 있다. 이를 과적합 이라고 하며, 과적합된 모델은 예측력이 떨어지므로 반드시 경계해야 할 대상이다.

이 절에서는 전체 데이터로부터 표본을 추출하는 방법으로 단순 임의 추출, 층화 임의 추출, 계통 추출에 대해 알아본다.

단순 임의 추출

단순 임의 추출(단순 무작위 추출)은 전체 데이터에서 각 데이터를 추출할 확률을 동일하게 하여 표본을 추출하는 방법이다. 예를 들어, 항아리에 빨란색 공 30개와 파란색 70개를 섞어놓은 뒤 항아리를 보지 않으면서 공 10개를 꺼내는 경우를 생각해볼 수 있다. 이렇게 공을 꺼내면 각 공이 뽑힐 확률이 동일하므로 단순 임의 추출에 해당한다.

데이터를 추출하는 방법에는 복원추출과 비복원추출이 있다. 복원 추출은 한 번 추출된 표본을 다시 선택하는 것이 가능한 경우를 뜻하며, 비복원 추출은 한 번 추출한 표본은 다시 선택할 수 없는 경우를 망한다. 항아리에서 공을 뽑을 때, 공을 하나 뽑아 그 색깔을 확인한 후 다시 항아리에 넣은 다음 공을 뽑는 것을 반복하는 경우가 복원 추출이다.

단순임의추출은 sample() 함수를 사용한다. 1~10까지의 수에서 5개를 비복원 추출로 뽑아보자.

```
> sample(1:10,5)
[1]  2  5 10  7  3
> sample(1:10,5, replace = T)
[1] 2 4 3 5 3
```

1~10까지의 수에서 복원 추출로 5개의 표본을 뽑아보자. 복원 추출이므로 같은 값이 여러번 뽑힐 수 있다. 위 예에서는 3이 두번 뽑혔다.

가중치를 고려한 표본 추출

각 데이터의 중요도나 발생 빈도가 다르다면 이를 고려하여 표본을 추출해야 한다. 예를 들어, 인터넷 쇼핑몰에서 제품과 제품별 판매량이 다음 표와 같이 정리되어 있다고 가정해보자.

제품명	판매량
A	23,428
B	104
C	3,392,201
D	1,392,485
...	...
Y	3,007
Z	32,037

그리고 이들 중 5개 제품을 선택해서 판매 데이터를 분석하기를 원한다고 가정해보자. 어떤 제품들을 분석해야 할까? 이 문제에 단순 임의 추출을 적용하면 알파벳 A~Z 중 5개가 모두 같은 확률로 선택될 것이다. 그러나 이렇게 표본을 추출하면 우연히 판매량이 적은 데이터만 선택될 가능성이 있다. 따라서 이런 경우에는 제품 5개를 선택될 가능성이 있다. 또, 인터넷 쇼핑몰 운영자 입장에서도 판매량이 많은 제품에 가중치를 두어 분석해보고 싶을 것이다. 따라서 이런 경우에는 제품 5개를 선택할 때 판매량에 비례하여 표본을 추출해야 한다.

sample()에는 이런 경우를 위해 각 데이터가 뽑힌 가중치를 지정하는 prob 파라미터가 있다. prob 라는 이름은 확률을 의미하지만 실제로 이 인자에 확률을 지정할 필요는 없다. 다만 가중치에 비례하는 음이 아닌 값을 지정하면 된다.

예를 들어, 1에서 10까지의 수에 각각 1에서 10까지 가중치를 주어 복원 추출을 해보자. 다음 코드를 보면 가중치가 큰 표본이 더 많이 뽑히는 경향을 쉽게 확인할 수 있다.

```
> sample(1:10,5, replace = T, prob = 1:10)
[1] 10 10  9  7 10
```

층화 임의 추출

데이터가 중첩 없이 분할될 수 있는 경우(즉, 서로 교집합이 없는 집합들로 나뉠 수 있는 경우) 그리고 각 분할의 성격이 명확히 다른 경우 층화 임의 추출(Stratified Random Sampling) 을 수행하여 더 정확한 분석 결과를 얻을 수 있다.

예를 들어, 남성 20%, 여성 80%로 구성된 집단이 있을 때 이 집단의 표본을 뽑아 키를 측정한 뒤 전체 집단의 평균 키를 에상한다고 가정해보자. 성별에 따라 키의 차이가 존재할 것이므로 , 성별을 고려하여 표본을 추출하는 것이 중요할 것이다. 그런데 단순 임의 추출을 이 데이터에 적용하면 전체 데이터로부터 각 데이터를 같은 확률로 뽑는다. 그 결과 남성이 우연히 20%보다 많이 추출되거나 적게 추출될 수 있다. 이렇게 뽑힌 표본에서 평균 키를 계산하면 평균 키가 집단의 실제 평균보다 우연히 작거나 크게 추정될 위험이 있다.

이 데이터의경우 남성 데이터와 여성데이터를떼놓고 각각으로부터 표본을 추출함으로써 남성과 여성의 표본 비율을 20% : 80%로 유지할 수 있다. 이를 층화 임의 추출이라 하며 데이터를 남성, 여성으로 분할한 것을 '층' 이라고 부른다.

층화 임의 추출을 하면, 뽑힌 남성의 표본을 사용해 남성 키의 평균을 추정할 수 있고, 여성 표본을 사용해 여성 키의 평균을 추정할 수 있다는 장점이 있다. 다시 말해 전체 평균 뿐만 아니라 층별 (성별) 평균 키의 추정이 가능해진다.

층화 임의 추출은 sampling::strata() 함수를 사용한다.

sampling::strata : 주어진 데이터에 대해 층화 임의 추출을 수행한다.

sampling :: getdata : 표본 추출을 수행한 뒤 데이터 프레임으로부터 값을 추출한다.

포아송 추출

각 데이터의 추출 확률을 다르게 설정하고 각 데이터를 독립적으로 추출하는 경우다. 예를 들어, N개의 데이터를 뽑을 때 각 데이터가 뽑힐 확률을 $\lambda_1, \lambda_2, \cdots . \lambda_n$ 으로 놓아 각각 독립적으로 추출하는 경우를 말한다.

다음은 iris 데이터로부터 비복원 단순 임의 추출을 사용해 각 종별로 3개의 표본을 추출하는 예다.

```
> (x<- strata(c("Species"),size = c(3,3,3),
+             method = "srswor", data = iris) )
       Species ID_unit Prob Stratum
12      setosa      12 0.06       1
17      setosa      17 0.06       1
29      setosa      29 0.06       1
51  versicolor      51 0.06       2
81  versicolor      81 0.06       2
93  versicolor      93 0.06       2
128  virginica     128 0.06       3
133  virginica     133 0.06       3
135  virginica     135 0.06       3

> getdata(iris, x)
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species ID_unit Prob
Stratum
12           4.8         3.4          1.6         0.2     setosa      12 0.06
1
17           5.4         3.9          1.3         0.4     setosa      17 0.06
1
29           5.2         3.4          1.4         0.2     setosa      29 0.06
1
51           7.0         3.2          4.7         1.4 versicolor      51 0.06
2
81           5.5         2.4          3.8         1.1 versicolor      81 0.06
2
93           5.8         2.6          4.0         1.2 versicolor      93 0.06
2
128          6.1         3.0          4.9         1.8  virginica     128 0.06
3
133          6.4         2.8          5.6         2.2  virginica     133 0.06
3
135          6.1         2.6          5.6         1.4  virginica     135 0.06
3
```

strata() 함수가 편리한 점은 층별로 다른 수의 표본을 추출할 수 있다는 점이다. 다음 예에서는 setosa 종에서 3개의 표본을 추출하고, 나머지 종에서는 1개씩 표본을 추출한다.

```
#srswor : 비복원 단순 임의 추출(Simple Random Sampling Without Replacement)
#srswr : 복원 단순 임의 추출(Simple Random Sampling With Replacement)
#spoison : 포아송 추출
#systematic : 계통 추출

> strata(c("Species"),size = c(3,1,1),
+        method = "srswr", data = iris)
       Species ID_unit     Prob Stratum
25      setosa      25 0.058808       1
27      setosa      27 0.058808       1
46      setosa      46 0.058808       1
91  versicolor      91 0.020000       2
127  virginica     127 0.020000       3
```

또, strata()는 다수의 층을 기준으로 데이터를 추출할 수 있다. 다음은 species2 라는 이름으로 또 다른 층을 만들고, (Species, Species2) 의 각 층마다 1개씩 표본을 추출하는 예다.

```
> iris$Species2 <- rep(1:2, 75)
> strata(c("Species", "Species2"), size = c(1,1,1,1,1,1), method = "srswr", data
= iris)
       Species Species2 ID_unit Prob Stratum
21      setosa        1      21 0.04       1
50      setosa        2      50 0.04       2
71  versicolor        1      71 0.04       3
88  versicolor        2      88 0.04       4
107  virginica        1     107 0.04       5
124  virginica        2     124 0.04       6
```

계통 추출

아침부터 밤까지 특정한 지역을 지나간 차량의 번호를 모두 조사한 뒤 이들로부터 조사 대상을 뽑는 경우를 가정해보자. 가장 간단한 단순 임의 추출을 적용하여 차량 번호를 뽑는다면 우연히 아침 시간에 지나간 차량을 더 많이 뽑거나, 저녁 시간에 지나간 차량을 더 많이 뽑는 편향이 발생할 수 있다. 계통 추출은 이런 상황에서 해결책이 될 수 있다.

계통 추출은 모집단의 임의 위치에서 시작해 매 k번째 항목을 표본으로 추출하는 방법이다. 예를 들어, $1, 2, 3, \cdots, 10$의 수에서 3개의 표본을 뽑는다고 가정해보자. 10/3 = 3.3333... 이므로 k =3 으로 놓는다. 표본 추출 시작 위치를 잡기 위해 1~k 사이의 수 하나를 뽑는다. 이 수가 2라하자. 나머지 두 수를 뽑기 위해 2+k에 해당하는 5를 뽑는다. 다음, 5+k에 해당하는 8을 뽑는다. 그러면 최종적으로 표본 2,5,8을 얻는다.

매우 단순한 방법이지만 데이터가 임의로 분포된 경우에는 단순 임의 추출 방법과 동일한 효과를 보이고, 데이터가 순서대로 나열된 순서 모집단(예를 들면, 1,2,3,..., 10과 같이 순서대로 나열된 모집단)의 경우 단순 임의 추출보다 좋은 표본을 추출한다. 하지만 데이터에 일종의 주기성이 존재한다면 (예를 들면, 데이터가 1,2,3,1,2,3,...과 같이 주기적으로 반복되는 경우) 편향된 추정값을 얻게 된다.

계통 추출은 doBy 패키지의 sampleBy(formula, frac = 0.1, replace = FALSE, data = parent.frame(), systematic = FALSE) 함수를 사용하여 수행할 수 잇다. 계통 추출을 하려면 systematic 인자에 TRUE를 지정한다.

다음 예는 1:10을 저장한 데이터 프레임에서 3개의 표본을 계통 추출로 뽑는 예다. 코드에서 sampleBy() 의 첫 번째 인자는 '~1' 이다. 그 이유는 첫 번째 인자가 표본을 추출할 그룹을 지정하는 표물러기 때문이다. 만약 그룹별로 데이터를 뽑느 층화 임의 추출이라면 그룹을 뜻하는 표현을 적어여 하지만, 여기서는 그룹의 구분이 없음로 상수 1을 사용했다. 실행 결과 '1,4,7' 3개의 표본이 뽑혔다.

```
> (x<-data.frame(x=1:10))
    x
```

```
1   1
2   2
3   3
4   4
5   5
6   6
7   7
8   8
9   9
10 10
> sampleBy(~1, frac = .3, data = x, systematic = T)
  [,1] [,2] [,3]
1    1    4    7
  [,1] [,2] [,3] [,4]
1    1    3    6    8
> sampleBy(~1, frac = .5, data = x, systematic = T)
  [,1] [,2] [,3] [,4] [,5]
1    1    3    5    7    9
```

7.4 분할표

분할표는 명목형 또는 순서형 데이터의 도수를 표 형태로 기록한 것이다. 분할표가 작성되면 카이 제곱 검정으로 변수 간에 의존 관계를 있는지를 독립성 검정으로, 도수가 특정 분포를 따르는지를 적합도 검정으로 살펴볼 수 있다.

Note 명목형 순서형 데이터

명목형 데이터는 가능한 값이 제한되어 있고 종종 고정되어 있는 변수를 의미한다. 예를 들면, 국가명, 혈액형 등이다. 순서형 데이터는 값의 순서를 숫자로 저장한 변수다. 설문 조사에서 제품 만족도를 조사하면서 응답을 매우 만족, 만족, 보통, 불만족, 매우 불만족으로 받을 수 있다. 이들 응답은 각각5,4,3,2,1으로 저장할 수 있는데, 이때 큰 값은 더 큰 만족을 의미한다. 하지만 이 값들 간에 비례적 관계는 존재하지 않는다.

분할표가 사용되는 한 가지 경우는 기계 학습으로 데이터의 양성, 음성으로 예측할 때다. 예를 들어, 이메일 텍스트를 보고 해당 이메일이 스팸인지 아닌지를 예측하는 경우를 생각해보자. 이때 두 가지 변수는 예측값(모델로 스팸인지를 판단한 결과)과 실제 값(실제로 해당 이메일이 스팸인지 여부)이다. 이런 실험에서 분할표의 모양은 다음과 같다.

	예측 - 스팸	**예측 - 스팸아님**
실제-스팸	a	b
실제-스팸 아님	c	d

표에서 a는 주어진 이메일이 실제로 스팸일 때 모델의 예측 결과도 스팸인 경우의 수다. b는 실제로 스팸인데 예측은 스팸이 아니라고 된 경우다.

분할표 작성

분할표를 작성하는 함수에는 table(), xtabs()가 있다. 여기서는 xtabs()를 주로 살펴보겠다.

xtabs : 포뮬러를 사용해 분할표를 작성한다.

다음은 table()을 사용해 주어진 벡터에서 a,b,c의 출현 횟수를 세는 간단한 예다.

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

a b c d
1 3 2 1
```

xtab()은 포뮬러를 사용해 데이터를 지정할 수 있다. 예를 들어, x,y라는 두 변수가 있고(x,y)에 대한 도수가 num에 저장되어 있을 때, 이 데이터로부터 분할표를 만드는 포뮬러는 num ~ x+y이다. 다음은 이 포뮬러를 사용해 분할표를 작성하는 예다.

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

만약 도수를 나타내는 컬럼이 따로 없고, 각 관찰 결과가 서로 다른 행으로 표현되어 있다면 '~번수 + 변수 ...' 형태로 포뮬러를 작성한다. 다음 코드는 x값이 A 또는 B인 각 경우의 수를 세는 예를 보여준다.

```
> (d2<- data.frame(x=c("A","A","A","B","B")))
  x
1 A
2 A
3 A
4 B
5 B
> (xtabs(~x,d2))
x
A B
3 2
```

합, 비율의 계산

여러 변수가 있을 때 한 변수만을 기준으로 총계를 구하는 경우를 생각해보자. 예를 들어, 다음은 변수 A,B에 대한 분할표다. 각 변수는 True, False 값을 가질 수 있다. 이 표에는 변수 B가 True, False인 경우의 합 (100과 90)을 표시한 컬럼과, 변수 A가 True, False 일 때 합을 표시한 총계 행 (80과 110)이 표시되어 있다. 또, 전체 도수의 합 역시 표시되어 있다. 이러한 총계 컬럼 또는 행을 주변 합이라고 한다. '주변'이란 데이터가 표시된 바깥쪽 행 또는 열에 값이 기록되기에 붙여진 이름이다.

	변수 A - True	**변수 A - False**	**총계**
변수 B - True	30	70	100
변수 B - False	50	40	90
총계	80	110	190

주변 합이 계산되고 나면 이를 기준으로 비율을 계산할 수 있다. 예를 들어, 다음은 위에 보인 표에서 변수 B의 합을 기준으로 비율을 계산한 결과다.

	변수 A - True	변수 A - False
변수 B - True	0.3	0.7
변수 B - False	0.56	0.44

마찬가지로 변수 A의 합을 기준으로 비율을 구하거나 전체 데이터의 총 빈도 (190)를 기준으로 비율을 계산할 수도 있다.

주변 합과 주변 비율은 margin.table(), prop.table()로 계산한다.

margin.table : 분할표의 주변 합을 구한다.

prop.table : 분할표의 주변 비율을 구한다.

```
> xt
   y
x   A B
  1 3 7
  2 8 5
> margin.table(xt,1) # 3 +7 = 10, 8 + 5 = 13
x
 1  2
10 13
> margin.table(xt,2) # 1 + 10 = 11, 7 + 5 = 12
y
 A  B
11 12
> margin.table(xt) 3 + 7 + 8 + 5 = 23
[1] 23
```

prop.table()은 분할표로부터 각 셀의 비율을 계산한다.

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

변수 간의 독립성 검정에는 카이 제곱을 사용하며, 이때 사용되는 통계량은 다음과 같다.

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

위 식에서 c는 열의 수, r은 행의 수를 의미한다. O_{ij} 는 분할표의 (i,j) 셀에 기록되어 있는 값이며, 분할표를 보면 바로 알 수 있는 값이다. E_{ij} 는 분할표의 두 변수가 독립일 때, (i,j) 셀에 대한 기댓값이다.

$cf)\ \ E_{ij} = n \times P(i,j) = P(i) \times P(j)$

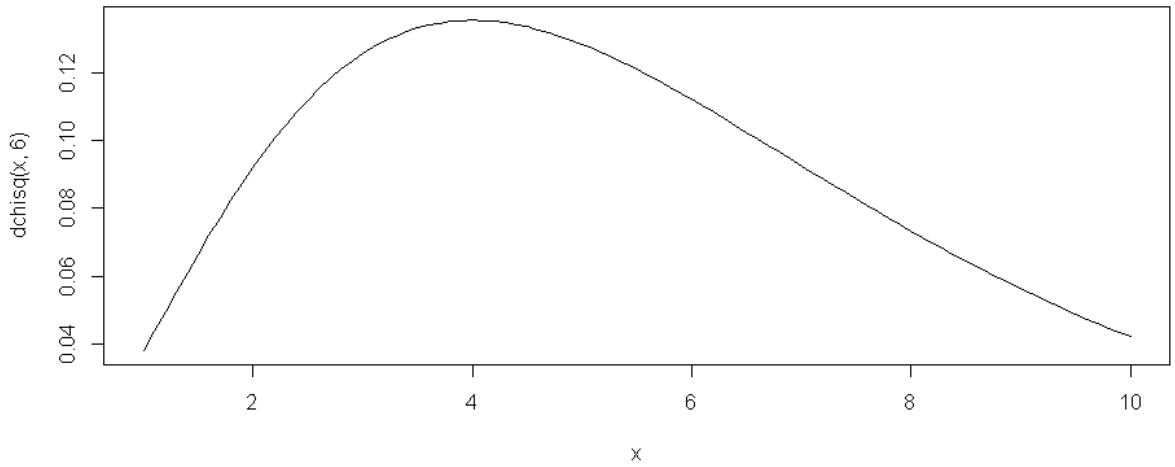
Note_ ~$\chi^2(r-1)(c-1)$ 의 의미

통계학에서 ~기호는 ~좌측에 있는 확률 변수가 ~우측에 있는 확률 분포를 따름을 의미한다. 따라서

~$\chi^2(r-1)(c-1)$ 은 ~좌측의 식이 카이 제곱 분포를 따른다는 의미이다. 카이 제곱 분포는 자유도라는 하나의 파라미터를 가지며, 이 파라미터에 따라 분포의 모양이 달라진다.

예를 들어, 분할표가 4행 3열이라면 (r-1)(c-1) = (4-1)(3-1) = 3 x 2 = 6이므로 $\chi^2(6)$ 분포를 따른다는 의미다. 이때 6은 카이 제곱 분포의 자유도를 뜻한다. 다음 그림에서 $\chi^2(6)$를 보였다. 다음그림에서 x축은 통계량 $\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, y축은 해당 통계량을 가질 확률이다.

```
> x <- seq(1,10,.1)
> plot(x,dchisq(x,6),type = 'l')
```



학생 설문 조사 데이터를 담고 있는 MASS::survey를 사용해 학생들의 성별에 따른 운동량에 차이가 있는지 독립성 검정을 해보자. 다음은 survey 데이터의 모양을 보여준다.

```
>  library(MASS)
> data(survey)
> str(survey)
'data.frame':   237 obs. of  12 variables:
 $ Sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
 $ Wr.Hnd: num  18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
 $ NW.Hnd: num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
 $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold  : Factor w/ 3 levels "L on R","Neither",..: 3 3 1 3 2 1 1 3 3 3 ...
 $ Pulse : int  92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap  : Factor w/ 3 levels "Left","Neither",..: 1 1 2 2 3 3 3 3 3 3 ...
 $ Exer  : Factor w/ 3 levels "Freq","None",..: 3 2 2 2 3 3 1 1 3 3 ...
 $ Smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
 $ Height: num  173 178 NA 160 165 ...
 $ M.I   : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
 $ Age   : num  18.2 17.6 16.9 20.3 23.7 ...

> head(survey[c("Sex","Exer")])
     Sex Exer
1 Female Some
2   Male None
3   Male None
4   Male None
5   Male Some
6 Female Some
```

survey 데이터에서 성별은 Sex, 운동을 얼마나 하는지는 Exer 열에 저장되어 있다. Sex는 Female, Male 두 가지 레벨을 가지는 팩터이며, Exer은 Freq, Some, None 3가지 레벨로 구성된 팩터이다. Freq는 운동을 자주함, Some은 운동을 약간 함, None은 운동을 하지 않음을 의미한다.

성별과 운동이 독립인지를 확인해보기 위해 분할표를 만들어보자.

```
> xtabs(~Sex+Exer, data = survey)
        Exer
Sex      Freq None Some
  Female   49   11   58
  Male     65   13   40
```

분할표를 작성하고 나면 chisq.test()를 통해 카이 제곱 검정을 수행할 수 있다.

다음은 성별과 운동 정도의 독립성 검정을 수행한 예다.

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

    Pearson's Chi-squared test

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

Note_ 귀무가설, 대립가설, p-value, 유의수준

가설 검정에서 사용하는 귀무가설, 대립가설, p-value에 대해 알아보자.

통계에서의 가설 검정은 측정된 두 현상 간에 관련이 없다는 귀무가설(Null Hypothesis : 흔히 H_0 으로 표시함)과 두 현상간에 '관련이 있다'고 보는 연구자가 알아보고자 하는 가설인 대립가설(Alternative Hypothesis : 흔히 H_0 로 표시) 을 사용한다. 귀무가설과 대립가설은 서로 모순 관계이다. 따라서 귀무가설이 참이면 대립가설이 거짓이고, 귀무가설이 거짓이면 대립가설이 참이다.

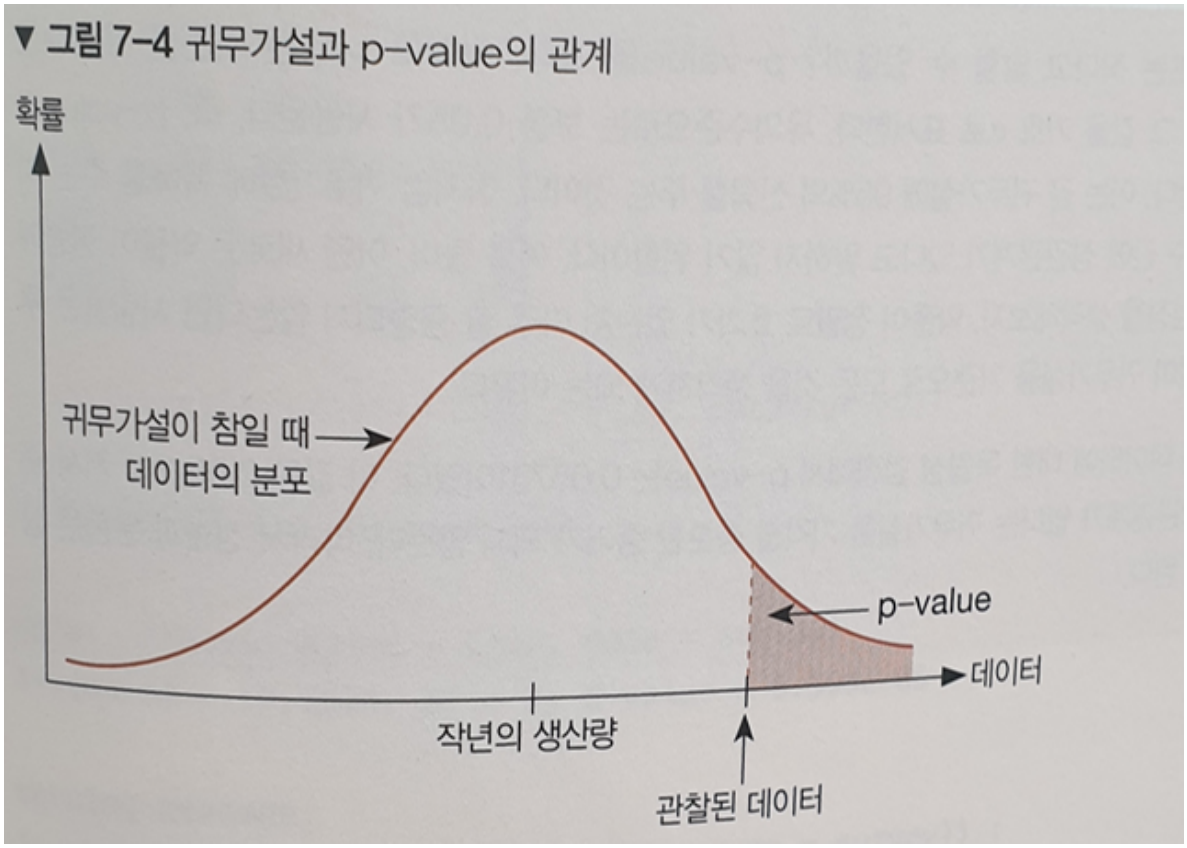
귀무가설은 '관련이 없다'는 형태의 가설이다. 귀무가설의 예에는 '두 변수가 독립이다', '두 변수의 평균에 차이가 없다' 등을 들 수 있다.

대립가설은 '관련이 있다'는 형태의 가설이다. 그 예로는 '두 변수가 독립이 아니다', '올 해 제품의 생산량과 작년의 생산량이 다르다' 등을 들 수 있다.

대립가설은 값이 '같지 않다', '작다' , '크다' 세 가지 형태로 나타낼 수 있다. 그 예로 '올해 생산량은 작년의 생산량과 다르다(즉, 올해 생산량이 크거나 작다)' , '올해 생산량은 작년의 생산량보다 작다' ,'올해 생산량은 작년의 생산량보다 크다' 를 들 수 있다. 이들 중 '같지 않다' 를 양측 검정(two sided test), '크다'와 '작다' 를 단측 검정(one sided test) 이라 한다.

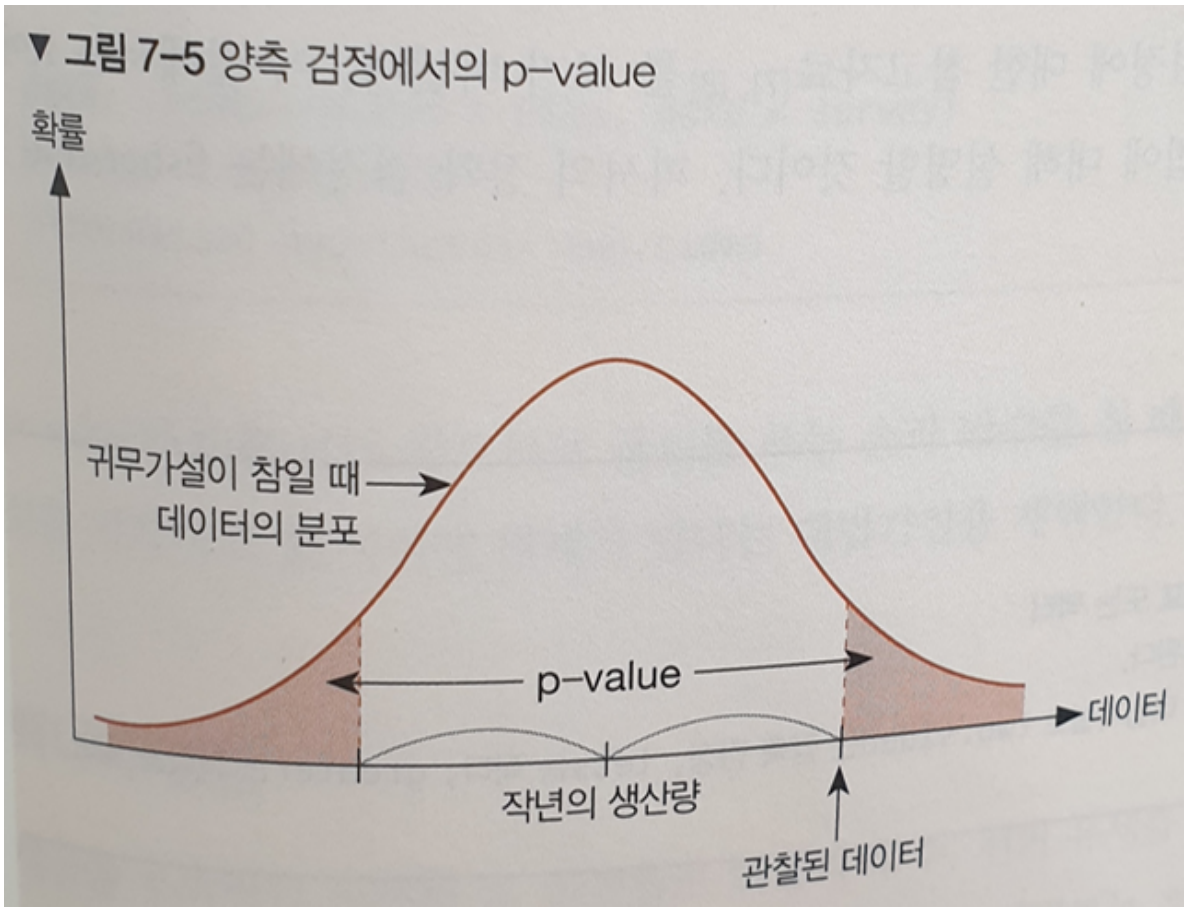
가설 검정은 귀무가설을 일단 참이라고 가정하고 시작한다. 그 뒤 귀무가설을 참이라고 생각했을 때 주어진 데이터 또는 그보다 극단적인 데이터가 관측될 확률을 구한다. 이를 p-value라고 한다. '더 극단적'이라는 개념은 대립가설의 형태마다 다르다. '크다' 형태의 대립가설이라면 관측값 또는 그 값보다 큰 값을 볼 확률이 될 것이고, '작다' 형태의 대립가설이면 관측값 또는 그보다 작은 값을 볼 확률이 된다. 반면 양측 검정('같지 않다' 형태)의 경우에는 작은 경우와 큰 경우를 모두 포함한다.

예를 들어, 공장에서 올해의 생산량이 작년의 생산량보다 큰지를 알아보기 위해 전체 공장 100곳 중 10곳의 생산량을 조사해봤다고 하자. 이때 귀무가설(H_0) 은 '올해 생산량은 작년의 생산량과 같다'이고 대립가설(H_1) 은 '올해 생산량은 작년의 생산량보다 크다' 이다. '크다' 형태의 대립가설에서 p-value는 다음 그림과 같다.



그림에서 곡선은 귀무가설이 참일 때(올해 생산량과 작년의 생산량이 같을 때) 10곳의 생산량이 어떻게 관찰되어야 하는지의 확률 분포를 나타낸다. 이 확률 분포는 작년의 생산량을 중심으로 서서히 작년의 생산량과 다른 값을 볼 확률이 낮아지는 형태다. p-value는 조사 결과 알게 된 10곳의 생산량 또는 그보다 큰 생산량이 관찰될 확률로, 색칠된 영역에 해당한다.

반면 대립가설을 'H_1: 올해의 생산량은 작년의 생산량과 다르다'로 놓으면 다음 그림과 같이 반대쪽 영역을 p-value 계산에 포함한다. 그 이유는 생산량이 같다는 귀무가설이 참이라고 할 때 '더 극단적'인 영역이란 작년의 생산량에 해당하는 가운데 부분에서 더 멀리 떨어지는 것을 의미하기 때문이다.



위 두 그림에 보인 것처럼 궁극적으로, p-value는 귀무가설이 참일 때 주어진 데이터가 관찰될 확률이다. 따라서 p-value가 작다면 귀무가설이 참이라고 믿었는데 관찰된 데이터는 그 가정 하에서는 좀처럼 볼 수 없는 값이었다는 뜻이다. 따라서 p -value가 작다면 귀무가설이 사실이 아니라고 볼 수 밖에 없으므로 대립가설을 참이라고 판단한다. 이를 통계 용어로 '귀무가설을 기각(reject)하고 대립가설을 채택(accept)한다'고 표현한다. 반대로 p-value가 크다면 귀무가설을 기각할 수 없으므로 대립가설을 기각하게 된다. 즉, p-value는 귀무가설을 지지하는 정도다.

7.4.1 피셔의 정확 검정

분할표를 그린 뒤 카이 제곱을 적용할 때 표본 수가 적거나 표본이 분할표의 셀에 매우 치우치게 분포되어 있다면 카이 제곱 검정의 결과가 부정확할 수 있다. 표본 수가 적다는 기준은 특정 짓기 어렵지만 기대빈도가 5 이하인 셀이 전체의 20% 이상인 경우 등이 이에 해당한다. chisq.test()는 이런 경우 경고 메시지를 출력하여 카이 제곱 검정이 부정확할 수 있음을 알린다.

카이 제곱 검정이 부정확한 경우에는 피셔의 정확 검정을 사용한다. 통계적인 계산식에 대한 설명은 참고자료를 봐라. 피셔의 정확 검정에는 fisher.test() 사용한다.

MASS::survey 데이터에서 손 글씨를 어느 손으로 쓰는지와 박수를 칠 때 어느 손이 위로 가는지 사이의 경우에 대해 피셔의 정확 검정을 수행해보자. 분할표를 xtab()으로 구한 뒤 카이 제곱 검정을 수행하면 카이 제곱 검정이 정확하지 않다는 경고 메시지가 나온다.

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

    Pearson's Chi-squared test

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

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

이 경우 fisher.test() 함수를 사용해야 한다.

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

    Fisher's Exact Test for Count Data

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

p-value 가 0.05보다 작으므로 글씨를 쓰는 손과 박수를 칠 때 위에 오는 손이 독립이라는 귀무가설을 기각하고 둘 사이에 관계가 있다는 대립가설을 채택한다.

7.4.2 맥니마 검정

벌금을 부과하기 시작한 후 안전벨트 착용자의 수, 선거 유세를 하고 난 뒤 지지율의 변화와 같이 응답자의 성향이 사건 전후에 어떻게 달라지는지를 알아보는 경우 맥니마 검정을 수행한다. 사건 전후에 응답자에게 설문을 하여 사건 발생 전 설문 결과를 Test1, 사건 발생 후 설문 결과를 Test2라고 명시한 다음 표를 보자.

	Test 2 양성(positive)	Test 2 음성(negative)	총계
Test 1 양성(positive)	a	b	a+b
Test 1 양성(positive)	c	d	c+d
총계	a+c	b+d	a+b+c+d

사건 전후에 설문 결과에 응답자 수 변화가 없다면 Test1의 positive와 Test2의 a+b = a+c 가 성립해야 한다. 또한, Test1의 negative와 Test2의 negative가 동일해야 하므로 c+d = b+d가 성립해야 한다. 이 둘을 정리해 결과적으로 b = c 여부를 검토하면 사건 전후에 성향 변화가 생겼는지 알 수 있다. b = c 가 성립하려면 b,c의 값이 b+c의 절반씩이 되어야 하므로 b는 이항 분포를 따른다.

$$b \sim B(b+c, \frac{1}{2})$$

Note 이항분포

이항 분포는 B(n,p)로 표현하며 성공 가능성이 p로 일정하고 성공과 실패의 두 가지 결과만을 가진 실험(이를 베르누이 실험이라 한다)을 나타낸다. 예를 들어, 동전을 던졌을 때 앞면이 나올 확률이 50%일 때 이 동전을 10번 던지는 시행은 B(10,0.5)이다.

이항 분포 B(n,p)에서 n(앞서 설문 조사 표에서는 b+c에 해당)이 크다면 이항 분포를 정규 분포로 근사할 수 있다.

$$b \sim N(\frac{b+c}{2}, \frac{b+c}{4})$$

Note 이항 분포의 정규 분포 근사

정규 분포는 평균 μ, 분산 σ^2 일 때, $N(\mu, \sigma^2)$ 으로 표현한다. 이항 분포 B(n,p)의 평균은 np, 분산은 np(1-p) 이므로 n이 클때, B(n,p)는 N(p,np(1-p))로 근사할 수 있다.

b를 표준화하여 N(0,1)을 따르게 하고 연속성 수정을 하면 다음이 성립한다.

$$\frac{(|b-c|-1)^2}{b+c} \sim \chi^2(1)$$

이제 ~의 좌측에 있는 통계량을 계산한 다음 이 값이 $\chi^2(1)$의 어디에 있는지를 보면 b = c 인지 여부를 알 수 있다. 이를 수행해주는 R 함수는 mcnemar.test()이다. 또는 b가 b+c의 절반에 해당하는지를 이항 분포를 사용해 검정할 수도 있다

이때 사용하는 함수가 mcnemar.test()이다. 또는 b가 b+c의 절반에 해당하는지를 이항분포를 통해 검정할 수도 있다. 이때 사용하는 함수는 binom.test()이다.

다음은 help(mcnemar.test)에서 가져온 것으로, 2x2 분할표에서 맥니마 검정을 사용하는 예다.

사용된 데이터는 투표권이 있는 나이의 미국인 1,600명에 대해 대통령 지지율을 조사한 것으로, 1차 조사와 2차 조사는 한 달 간격으로 수행되었다.

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

```
> mcnemar.test(Performance)

    McNemar's Chi-squared test with continuity correction

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

결과에서 p-value < 0.05가 나타나 사건 전후에 Approve, Disapprove에 차이가 없다는 귀무가설이 기각된다. 즉, 사건 전후에 Approve, Disapprove 비율에 차이가 발생했다.

앞서 mcnemar.test()는 이항 분포로부터 나왔으며, b가 b+c의 절반에 해당하는지를 보는 것이라고 설명했다. 따라서 binom.test()를 사용해 1차 조사에서의 Disapprove와 2차 조사에서의 Disapprove가 같은 값인지 확인할 수 있다. 다음 코드에서는 86이 86+150의 절반에 해당하는지를 검정하고 있다.

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

    Exact binomial test

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

여기에서도 p-value < 0.05 로 나타나 86이 86+150의 절반이라는 귀무가설이 기각되었다. 즉, 사건 전후에 Approve, Disapprove 성향 차이가 발생했다.

7.5 적합도 검정

통계 분석에서는 종종 데이터가 특정 분포를 따름을 가정한다. 특히 데이터의 크기가 일정 수 이상이라면 데이터가 정규분포를 따름을 별 의심 없이 가정하기도 한다. 하지만 실제로 그 분포를 따르는지 확인해볼 수도 있다.

7.5.1 카이 제곱 검정

데이터가 특정 분포를 따르는지 살펴보기 위해 분할표를 만들고, 카이 제곱 검정을 사용할 수 있다. 다만 독립성 검정과 달리 E_{ij} 를 비교하고자 하는 분포로부터 계산한다.

MASS::survey 데이터를 사용해 글씨를 왼손으로 쓰는 사람과 오른손으로 쓰는 사람의 비율이 30% : 70%인지 여부를 분석해보자. 아래에서 수행한 chisq.test()에서 귀무가설은 분할표에 주어진 관측 데이터가 30% : 70%를 따른다는 것이다.

```
> table(survey$W.Hnd)

 Left Right
   18   218
> chisq.test(table(survey$W.Hnd), p=c(.3,.7))

    Chi-squared test for given probabilities

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

p-value < 0.05이므로 글씨를 왼손으로 쓰는 사람과 오른손으로 쓰는 사람의 비가 30% : 70% 라는 귀무가설을 기각한다.

7.5.2 샤피로 윌크 검정

샤피로 윌크 검정은 표본이 정규 분포로부터 추출된 것인지 테스트하기 위한 방법이다. 검정은 shapiro.test() 함수를 사용하며, 이때 귀무가설은 주어진 데이터가 정규분포로부터의 표본이라는 것이다.

shapiro.test : 데이터가 정규 분포를 따르는지 샤피로 윌크 검정을 수행한다. 귀무가설은 정규 분포를 따른 다는 것이다.

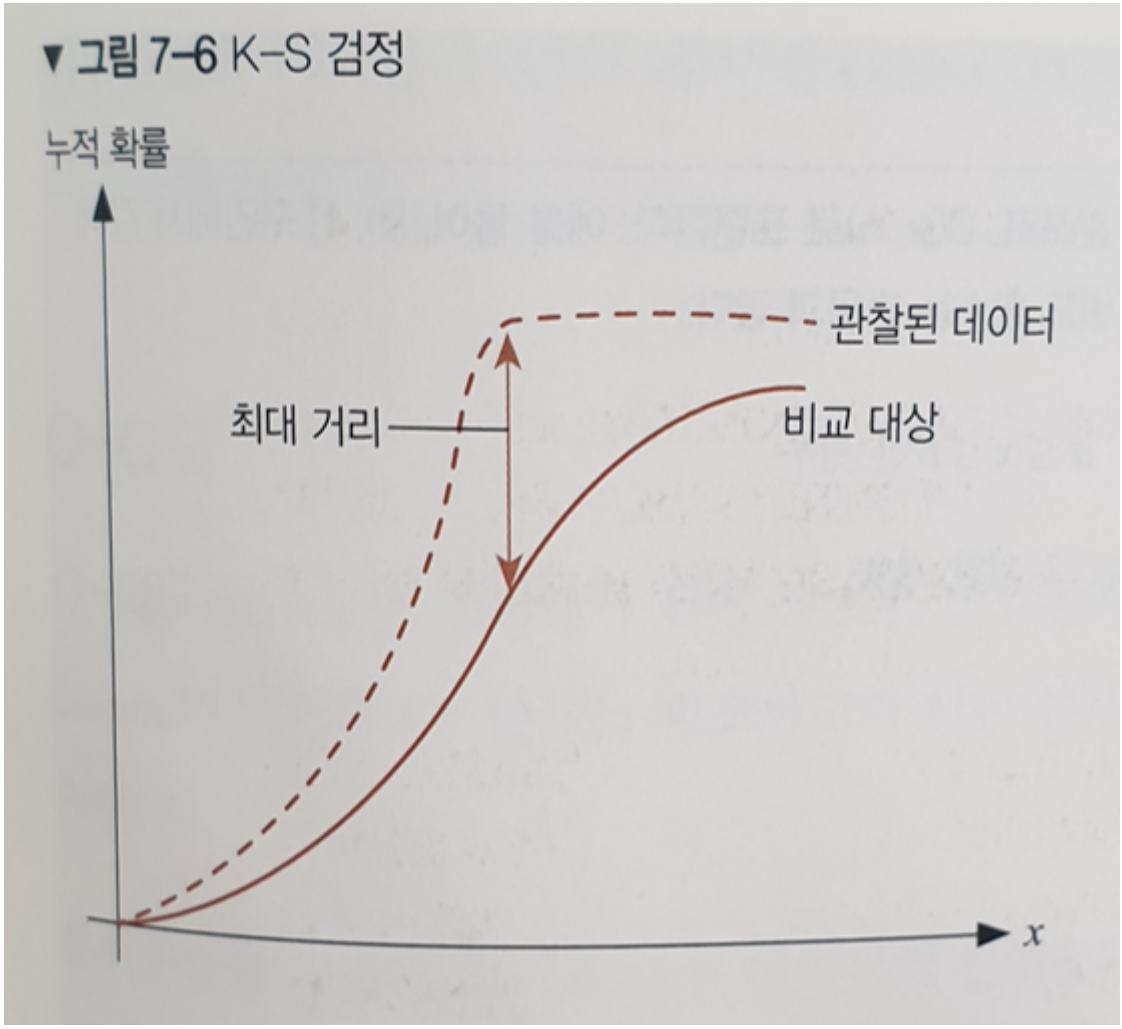
다음은 정규 분포를 따르는 1,000개의 난수를 발생시킨 뒤 이 숫자들이 정규 분포를 따르는지 샤피로 윌크 검정을 수행한 예다.

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

    Shapiro-Wilk normality test

data:  rnorm(1000)
W = 0.99885, p-value = 0.784
```

p-value > 0.05 이므로 데이터가 정규 분포를 따른다는 귀무가설을 기각할 수 없다. shapiro.test()외에도 nortest 패키지에는 앤더스 달링 검정, 피어슨 카이 제곱 검정 등을 사용해 정규성을 검정하는 다양한 함수가 있으니 참고해라.

7.5.3 콜모고로프 스미르노프 검정

콜모고로프 스미르노프 검정 (K-S Test 라고도 부름)은 데이터의 누적 분포 함수와 비교하고자 하는 분포의 누적 분포 함수 간의 최대 거리를 통계량으로 사용하는 가설 검정 방법이다. 다음 그림은 K-S 검정의 기본 아이디어를 보여준다.



K-S 검정은 ks.test() 함수를 사용해 수행한다.

ks.test : K-S 검정을 수행한다. 귀무가설은 x,y에 차이가 없다고 보는 것이다.

다음은 정규분포를 따르는 두 난수 데이터 간에 분포가 동일한지 살펴본 예다.

```
> ks.test(rnorm(100), rnorm(100))

    Two-sample Kolmogorov-Smirnov test

data:  rnorm(100) and rnorm(100)
D = 0.14, p-value = 0.281
alternative hypothesis: two-sided
```

p-value > 0.05 이므로 두 난수가 같은 분포라는 귀무가설을 기각할 수 없다. 다음은 정규 분포와 균등 분포 간의 비교를 해보자.

Note_ 균등분포

균등 분포는 주어진 구간의 값이 관찰될 확률이 일정한 확률 분포로, U(a,b)로 표현한다. 예를 들어, [0,1] 구간에서 값이 관찰될 확률이 동일한 확률 분포는 U(0,1)로 표현한다. 확률 밀도 함수는 다음과 같다.

$$f(x) = \frac{1}{b-a} \quad (a \le x \le b \text{ 인 경우})$$
$$0 \qquad (\text{그 외인 경우})$$

```
> ks.test(rnorm(100), runif(100))

    Two-sample Kolmogorov-Smirnov test

data:  rnorm(100) and runif(100)
D = 0.51, p-value = 1.012e-11
alternative hypothesis: two-sided
```

p-value가 0.05보다 작아 서로 다른 분포로 판단되었다. 다음은 K-5 검정을 사용해 데이터가 평균 0, 분산 1 인 정규 분포로부터 뽑은 표본인지 확인하는 예이다.

```
> ks.test(rnorm(100), "pnorm", 0,1)

    One-sample Kolmogorov-Smirnov test

data:  rnorm(100)
D = 0.081704, p-value = 0.5167
alternative hypothesis: two-sided
```

귀무가설을 기각할 수 없어, 주어진 rnorm(100)은 평균 0, 분산 1인 정규분포으로부터의 표본이라고 결론을 내린다.

7.5.4 Q-Q도

Q-Q도 (Q-Q plot)는 데이터가 특정 분포를 따르는지를 시각적으로 검토하는 방법이다. Q는 분위수 (Quantile) 의 약어로 Q-Q도는 비교하고자 하는 분포의 분위수끼리 좌표 평면에 표시하여 그린 그림이다.

분위수들을 차트에 그리고 나면 데이터의 분위수와 비교하고자 하는 분포의 분위수 간에 직선 관계가 보이는지 확인한다. 예를 들어, X가 정규 분포를 따르는지 살펴보고 싶다고 가정하자.

X~N(μ, α^2) 이라면 다음 관계가 성립한다.

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

즉, 정규 분포를 따르는 확률 변수는 정규화된 뒤에 평균이 0, 분산이 1인 정규 분포를 따른다. X를 Z로 표현하면 다음과 같다.

$$X = \mu + \sigma Z$$

따라서 X가 정규 분포를 따를 때 (X,Z) 를 좌표 평면에 표시한다면 위 식에 보인 직선이 나타나야 한다. Q-Q도는 이와 같은 직선 관계가 실제로 성립하는지 시각적으로 보여주는 도구다.

(X,Z)에서 X는 주어진 데이터므로 이미 알고 있는 값이다. 따라서 X에 대응하는 Z만 찾으면 된다. 이때 분위수가 사용된다. X가 몇 % 분위수인지를 안다면 N(0,1)에서 해당 % 분위수를 찾아 Z로 하면 된다.

X값들이 몇 % 분위수인지 찾아보자. X를 크기 순서로 정렬했을 때 관측값 $x_1 < x_2 < \cdots < x_n$ 이 된다고 하고, 이 데이터들의 분포를 표현하는 분포 함수가 다음과 같다고 하자.

$$G(x_i) = \frac{i - 3/8}{n + 1/4}$$

$G(x_i)$는 표본 데이털르 정렬했을 때 i번째 데이터가 몇 %분위수인지 알려주는 역활을 한다. 예를 들어, 표본의 크기 n이 20이라면 $G(x_1) = \frac{1-3/8}{20+1/4} = 0.03, G(x_2) = \frac{2-3/8}{20+2/4}$ 이 된다. 따라서 x_1은 X가 따르는 분포의 3% 분위수, x_2은 X가 따르는 분포의 7% 분위수 등이 된다.

X가 몇 % 분위수인지 알면 Z는 손쉽게 찾을 수 있다. Z가 표준 정규 분포를 따르므로 3% 분위수 인 z_1 , 7% 분위수 인 z_2 등은 Z의 누적 분포 함수가 Φ 라 할 때 $z_1 = \Phi^{-1}(0.03), z_2 = \Phi^{-1}(0.07)$이다.

지금까지 설명한 내용을 정리하면 X가 정규 분포를 가정 하에서 다음이 성립한다.

$$x_i = \mu + \sigma z_i = \mu + \sigma\phi^{-1}(G(x_i))$$

이제 X에 해당하는 Z를 찾았은 (X,Z)를 그려볼 차례다. 이 목적으로는 qqnorm() 함수를 사용한다.

qqline()은 Q-Q도 에서 데이터가 만족해야 하는 직선 관계를 그린다.

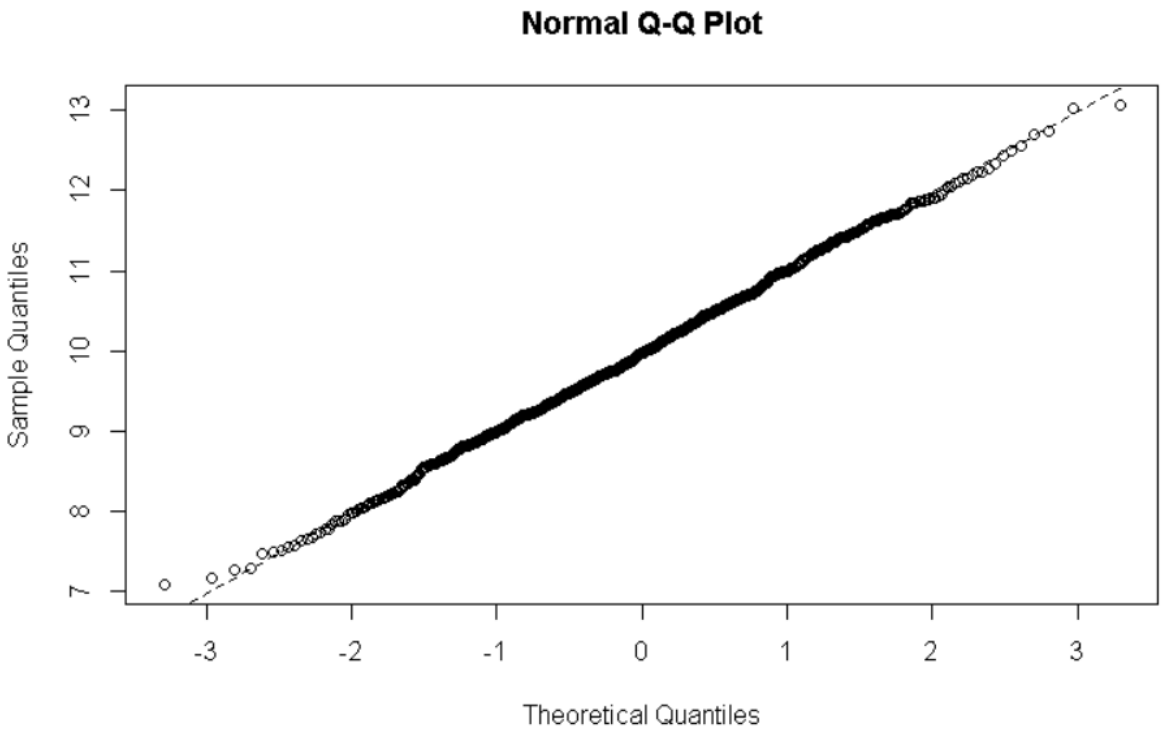
qqnorm : 주어진 데이터와 정규 확률 분포를 비교하는 Q-Q도를 그린다.

qqplot : 두 데이터 셋에 대한 Q-Q도를 그린다.

qqline : 데이터와 분포를 비교해 이론적으로 성립해야 하는 직선 관계를 그린다.

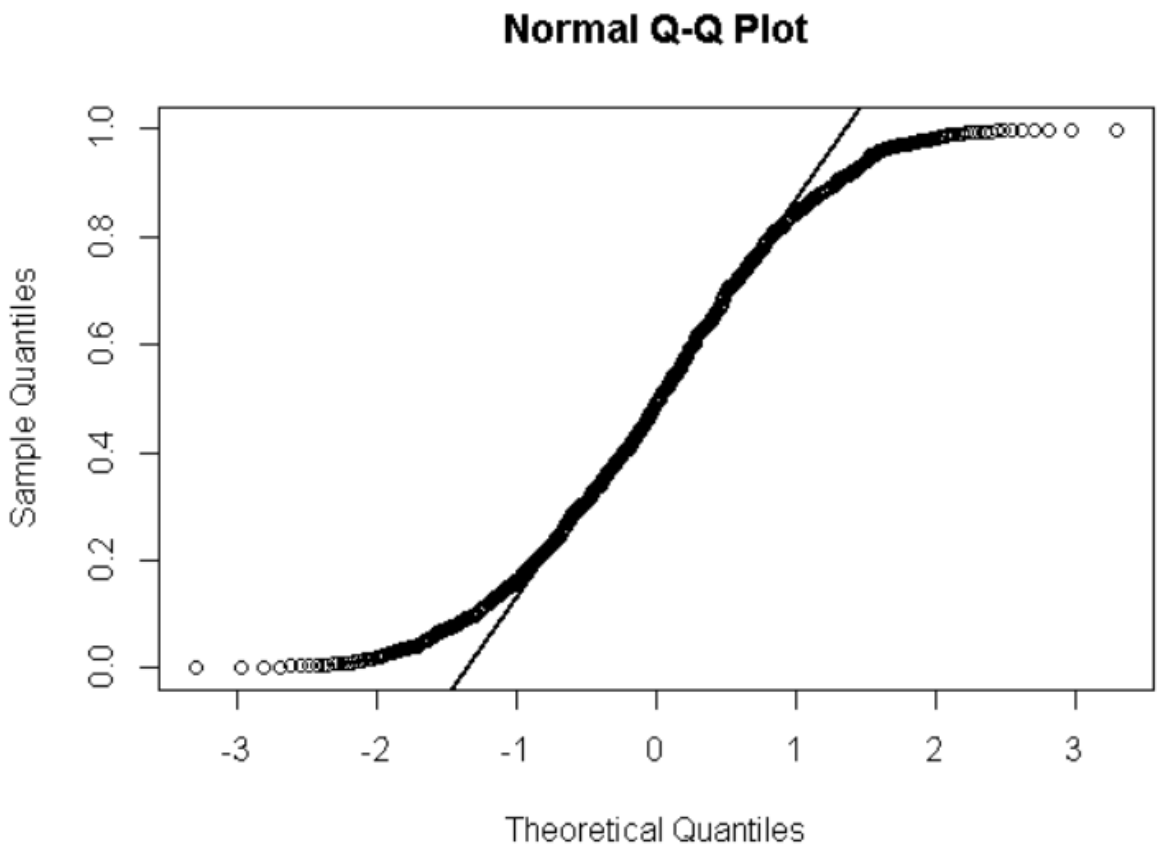
다음은 N(0,1)의 정규 분포를 따르는 난수 1,000개 구하고 Q-Q도를 그려 해당 숫자들이 정규 분포를 따르는지 확인해본 예다.

```
x<- rnorm(1000, mean = 10, sd = 1)
qqnorm(x)
qqline(x,lty =2 )
```



비교를 위해 균등 분포로부터 생성한 데이터로 정규 확률 그림을 그려보자. 결과를 보면 한눈에도 직선 관계가 성립하지 않음을 알 수 있다.

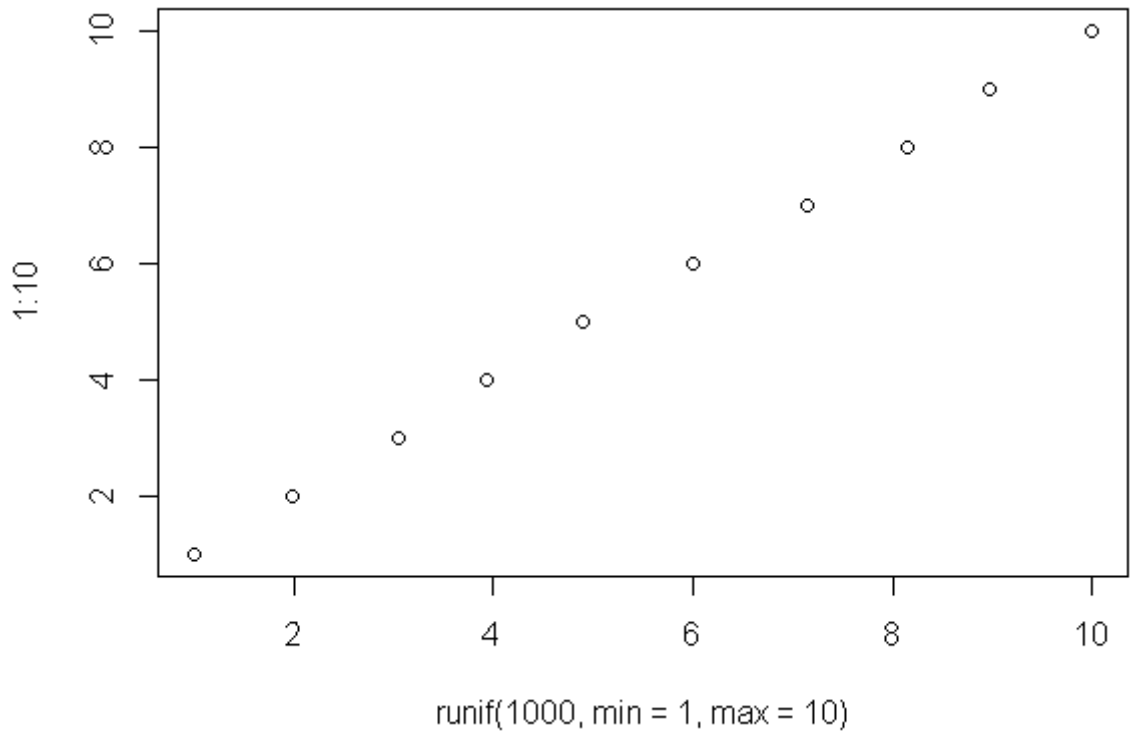
```
x<- runif(1000)
qqnorm(x)
qqline(x,lwd =2 )
```



데이터의 정규성이 Q-Q도를 통해 항상 명확히 판단되는 것은 아니다 . 직선 관계가 성립한다 할지라고 데이터의 출처 및 데이터가 정규성을 따를 이유에 대한 고민이 항상 필요하다.

정규 확률 그림이 아닌 분포에 대해서는 qqplot() 함수를 사용한다. 다음은 c(1,2,3,4,5,6,7,8,9,10)이 U(1,10)의 균등 분포를 따르는지 확인하기 위해 Q-Q도를 그린 예다. 비교할 대상인 U(1,10)은 runif()로 데이터를 만들었다.

```
qqplot(runif(1000,min = 1, max = 10), 1:10)
```



7.6 상관분석

상관 분석은 두 확률 변수 사이의 관련성을 파악하는 방법이다. 상관 계수는 두 변수 간 관련성의 정도를 의미하며, 이를 계산하는 방법에는 피어슨 상관계수, 스피어만 상관계수, 켄달의 순위 상관 계수 등이 있다. 그러나 흔히 상관 계수라고 하면 피어슨 상관 계수를 뜻한다.

7.6.1 피어슨 상관 계수

피어슨 상관 계수는 선형 관계를 판단하므로 $Y = aX + b$ 와 같은 형태의 관계를 잘 찾는다. 반면 $Y = aX^2 + b$ 와 같은 비선형 관계에서 Y는 X가 증가함에 따라 값이 커지는 것이 확실하지만 값의 증가 형태가 선형이 아니므로 피어슨 상관 계수가 작게 나타날 수 있다.

피어슨 상관 계수는 다음과 같이 정의된다.

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

식에서 $cov(X,Y)$는 X,Y의 공분산, σ_X, σ_Y X,Y의 표준 편차다.

Note_ 공분산

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

참고로 분산은 다음과 같이 정의 된다.

$$Var(X) = E[(X - E(X))^2]$$

따라서 공분산은 분산을 2개 변수로 확장한 형태로 생각할 수 있다.

다음은 R에서 공분산을 계산하는 예이다.

```
> cov(1:5, 2:6)
[1] 2.5
> cov(1:5,c(0,0,0,0,0))
[1] 0
> cov(1:5,5:1)
[1] -2.5
```

cor : 상관 계수를 구한다.

symnum : 숫자를 심볼로 표현한다.

corrgram::corgram : 상관 계수 행렬을 그림으로 보여준다.

다음은 아이리스 데이터에서 Sepal.Width, Sepal.Length의 피어슨 상관 계수를 구하는 예다. 상관 계수 값이 작아 두 값 사이에 큰 상관관계는 없지만, Sepal.Width가 커짐에 따라 sepal.Length가 작아지는 경향이 있음을 알 수 있다.

```
> cor(iris$Sepal.Width, iris$Sepal.Length)
[1] -0.1175698
```

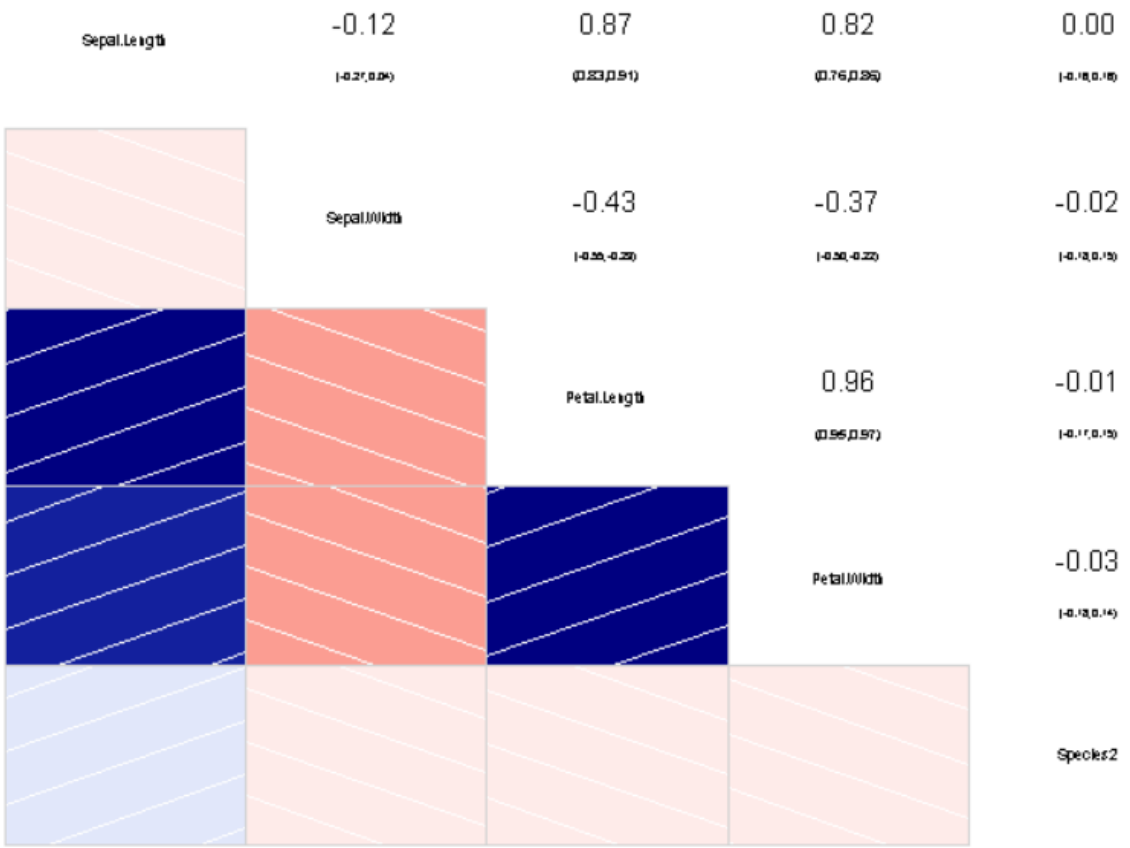
아이리스에서 Species를 제외한 모든 컬럼의 피어슨 상관 계수를 구해보자.

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

만약 살펴봐야 할 컬럼의 수가 많다면 숫자가 잘 안들어 올 수 있다. symnum()을 사용하면 숫자들을 간략한 기호로 볼 수 있다. 결과에 표시된 기호들을 보면 P.L(Petal.Length)과 Petal.Width 간의 상관 관계가 가장 크고 (B로 표시됨), 그 다음으로 Sepal.Length와 Petal.Length, Sepal.Length와 Petal.Width가 상관 관계가 큼(+로 표시됨)을 쉽게 알 수 있다.

```
> symnum(cor(iris[,1:4]))
             S.L S.W P.L P.W
Sepal.Length 1
Sepal.Width      1
Petal.Length +   .   1
Petal.Width  +   .   B   1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

corrgram 패키지는 상관관계를 시각화하는 데 유용하다. 다음 코드는 아이리스의 상관 계수를 그림의 우측 상단에 배치하고 (upper.panel = panel.conf), 대각선에는 컬럼의 이름을 적고, 좌측 하단에는 상관계수를 그림으로 표현하는 예다. 그림에서 오른쪽 위에서 왼쪽 아래로 빗금이 쳐진 부분은 양의 상관 계수를 뜻하고, 왼쪽 위에서 오른쪽 아래로 빗금이 쳐진 영역은 음의 상관계수를 뜻한다. 색의 짙기는 상관계수의 크기를 뜻해 절댓값이 큰 상관 계수일수록 더 짙은 색을 띤다.



피어슨 상관 계수는 데이터의 선형 관계 정도를 판단한다. 따라서 Y=X, Y = 2X는 모두 선형 관계가 성립하므로 피어슨 상관 계수가 1이다.

7.6.2 스피어만 상관 계수

스피어만 상관 계수는 상관 계수를 계산할 두 데이터의 실제값 대신 두 값의 순위를 사용해 상관 계수를 계산하는 방식이다. 피어슨 상관 계수와 마찬가지로 값의 범위는 [-1,1]이며 1은 한쪽의 순위가 증가함에 따라 다른 쪽의 순위도 증가함을 뜻하고, -1은 한쪽의 순위가 증가할 때 다른 쪽의 순위는 감소함을 뜻한다. 0은 한쪽의 순위 증각가 다른 쪽의 순위와 연관이 없음을 뜻한다.

스피어만 상관 계수의 계산식은 피어슨 상관계수와 유사해 이해하기 쉽고, 피어슨 상관 계수와 달리 비선형 관계의 연관성을 파악할 수 있다는 장점이 있다. 또한, 데이터에 순위만 매길 수 있다면 적용이 가능하므로 연속형 데이터에 적합한 피어슨 상관계수와 달리 이산형 데이터, 순서형 데이터에 적용이 가능하다. 예를 들어, 국어 점수와 영어 점수 간의 상관 계수는 피어슨 상관 계수로 계산할 수 있고, 국어 성적 석차와 영어 성적 석차의 상관 계수는 스피어만 상관 계수로 계산할 수 있다.

스피어만 상관 계수는 다음과 같이 정의한다.

$$p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \text{ ----식 } 7.5$$

위 식에서 x_i는 변수 X에서 i 번째 데이터의 순위, y는 Y에서 i번째 데이터의 순위, $\bar{x}, \bar{y}$는 x,y의 평균을 뜻한다. 이 식을 피어슨 상관 계수와 비교해보면 모양이 상당히 유사함을 알 수 있다.

식 7.5는 $d_i = x_i - y_i$ 라 놓으면 다음과 같이 간단히 할 수 있다.

$$p = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \text{ ------식 } 7.6$$

데이터가 주어졌을 때 순위 계산은 다음과 같이 한다. 예를 들어, 3,4,5,3,2,1,7,5를 가정해보자. 이 데이터를 정렬해보면 1,2,3,3,5,5,7이 된다. 또, 각 값의 순위는 1,2,3.5,3.5,5,6.5,6.5,8 이 된다. 데이터에서 3, 3은 순위 3,4에 해당하므로 이들의 평균은 3.5가 순위로 주어지고 5,5는 순위가 6,7 이므로 평균인 6.5 순위로 주어진다. R코드를 사용해 순위를 확인해보자.

```
> x<- c(3,4,5,3,2,1,7,5)
> rank(sort(x))
[1] 1.0 2.0 3.5 3.5 5.0 6.5 6.5 8.0
```

이처럼 비교할 수 없는 데이터로부터 순위를 구한 다음 식 7.6에 적용하면 된다. R 코드에서는 스피어만 상관 계수 계산 시와 마찬가지로 cor() 함수를 사용한다. 다음은 가상의 순위가 저장된 행렬로부터 스피어만 상관 계수를 계산한 예다. 숫자가 양쪽에서 모두 증가하므로 스피어만 상관 계수는 1로 계산된다.

```
> cor(m,method = 'spearman')
     [,1] [,2]
[1,]    1    1
[2,]    1    1

> cor(m,method = 'pearson')
          [,1]      [,2]
[1,] 1.0000000 0.9745586
[2,] 0.9745586 1.0000000
>
```

반면 같은 값을 피어슨 상관 계수에 넣으면 선형 관계를 따지게 되어 컬럼 1,2 간의 상관 계수가 0.9745586 으로 1보다 작게 나타난다.

7.6.3 상관 계수 검정

cor.test()를 사용해 상관 계수 검정을 수행하여 상관 계수의 통계적 유의성을 판단할 수 있다. 이때 귀무가설은 'H_0 : 상관 계수가 0이다'이며, 대립가설은 'H_1 : 상관 계수가 0이 아니다'이다.

cor.test : 상관 계수에 대한 가설 검정을 수행한다.

c(1,2,3,4,5)와 c(1,0,3,4,5) 간의 피어슨 상관 계수, 스피어만 상관계수에 대해 상관 계수 검정을 수행해보자.

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

    Pearson's product-moment correlation

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

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

    Spearman's rank correlation rho

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

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

    Kendall's rank correlation tau

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

코드 수행 결과 피어슨 상관 계수에서만 p-value가 0.05 보다 작아 상관관계가 유의한 것으로 나타났다. 이처럼 세 가지 상관 계수의 값은 서로 다른 값이 될 수 있다. 이런 경우 더 작은 숫자를 사용하는 것이 바람직하다. 위의 경우라면 켄달의 상관 계수 0.8이 가장 작은 값이므로 이 값을 사용한다. 또는 세 가지 값이 계산하는 것이 무엇인지 알고 데이터의 성격에 비추어 올바른 상관 계수를 찾아 사용해야 한다.

7.7 추정 및 검정

이 절에서는 평균, 분산, 비율에 대해 추정하고 검정하는 방법에 대해 살펴본다. 이 절에서 주로 설명할 추정은 구간 추정이다. 예를 들어, 평균의 구간 추정은 데이터로부터 표본을 추출하여 표본의 평균을 구한 뒤 전체 데이터의 평균이 어떤 구간 [a,b]에 있다고 말하는 것이다.

이 절의 내용에서 이론적 배경 부분은 참고자료를 봐라.

Note_ 점 추정, 구간 추정, 신뢰 수준, 신뢰 구간

이 절에서 살펴볼 추정은 전체 데이터(이후 모집단 이라고 부름)로부터 일부를 표본으로 취한 뒤 표본으로부터 신뢰 수준 95%의 신뢰 구간을 구하는 내용이다.

평균을 구하는 경우를 생각해보자. 모딥단에서 일부를 표본으로 추출한 뒤 표본의 평균을 구하면 이 값을 전체 데이터의 평균으로 볼 수 있다. 예를 들어, 우리나라 20대의 평균 키를 알고자 한다면 전국 20대 중 100명을 뽑아 이들의 키를 재고 평균을 구할 수 있다. 그러면 그 평균이 전국 20대 평균에 대한 추정값이다. 이처럼 **하나의 값으로 추정하는 것을 점 추정**이라고 한다.

그러나 실제 우리나라 20대의 평균값이 정확히 점 추정된 값은 아닐 수 있다. 그보다는 점 추정된 값 근처 어딘가에 있다고 보는 것이 더 타당할 것이다. 이때 사용되는 개념이 신뢰 구간(confidence interval) 이다. 신뢰 구간은 (a,b) 의 구간 형태 또는 x±y 형태로 표현한다. 예를 들어, **(160cm, 180cm) 이렇게 추정하는 것이 구간 추정**이다.

그렇다면 신뢰 구간은 어찌 구할까? 신뢰 구간은 신뢰 수준(confidence level)을 먼저 전제하여 계산한다. 신뢰 수준은 신뢰 구간을 구하는 작업을 여러 번 반복했을 때 참 값이 그 구간에 얼마나 자주 속하는지를 뜻한다. 예를 들어, 95% 신뢰 수준으로 20대 평균 키를 구간 추정했을 때 (160cm, 180cm)였다는 말은 20대 평균 키를 알아내기 위해 20대 중 100명을 뽑아 평균 키에 대한 신뢰 구간을 구하는 작업을 여러 번 반복했을 때 그 중 95%에서 신뢰 구간 안에 우리나라 20대 평균 키가 속해 있었다는 의미다.

신뢰 구간을 구하는 데는 약간의 통계적 가정이 필요하다. 예를 들면 우리나라 20대의 키는 정규 분포를 따른다거나 하는 것이다. 이렇게 분포를 가정하면 통계적 이론에 의해 주어진 신뢰 수준에 해당하는 신뢰 구간을 구할 수 있다.

7.7.1 일표본 평균

이 절에서는 하나의 모집단으로부터 표본을 추출하고 표본으로부터 모집단 평균의 신뢰 구간을 구하는 방법을 살펴본다.

이론적 배경

확률 변수 $X_1, X_2, \cdots, X_n$ 이 서로 독립이고 정규 분포 $N(\mu, \sigma^2)$ 을 따른다고 하자. 즉, X_i는 $N(\mu, \sigma^2)$ 으로부터 구한 표본이며, 표본의 크기는 n 이다. 표본의 평균이 X라 할 때, 다음과 같이 평균이 0, 분산이 1인 정규 분포를 따르는 식을 구할 수 있다.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$ ----------식 7.8

모집단의 분산 σ^2 은 보통 알려진 값이 아니므로 다음에 정의된 표본 분산을 사용한다.

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

식 7.8에 σ^2 대신 S^2 을 사용하면 자유도가 n-1인 t 분포를 따르게 된다.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$ ----------식 7.9

따라서 유의수준 α 가 0.05일 때 모평균에 대한 95% 신뢰 수준의 신뢰 구간은 다음과 같다.

$$(\bar{X} - t(n-1;\alpha/2)S/\sqrt{n}, \bar{X} + t(n-1;\alpha/2)S/\sqrt{n})$$

Note_ 모평균, 모분산, 표본 평균, 표본 분산

모평균은 모집단 (연구 대상이 되는 전체)에서 구한 평균이다. 마찬가지로 모분산은 전체 데이터에 대한 분산을 뜻한다. 표본 평균은 모집단으로부터 추출한 표본의 평균이며, 표본 분산은 표본의 분산을 뜻한다. 모평균은 μ,모분산은 σ^2 으로 표현한다. 표본 평균과 분산은 표본을 $X_1, X_2, \cdots, X_n$ 으로 표현할 때 각각 $\bar{X}, S_X^2$ 으로 적는다.

가설 검정 및 추정에서 하는 일은 표본의 평균과 분산으로부터 모집단의 평균과 분산을 알아내고자 하는 것이다.

추정 및 검정의 예

일표본 평균의 구간 추정 및 가설 검정에는 t.test() 함수를 사용한다.

t.test : t검정을 수행한다. 귀무가설은 '모평균이 mu와 같다'이다.

일표본 평균의 추정 및 가설 검증에서는 t.test()에서 보통 x, alternative, mu 인자만 사용한다. 다음은 평균 0, 분산 1인 정규 분포 N(0,1) 로부터 30개의 표본을 뽑아 모평균의 구간을 측정한 예다.

```
> x <- rnorm(30)
> x
 [1] -0.44842386 -0.56426144 -1.65950317
 [4] -2.08691128  0.42263301  0.34304892
 [7] -0.70535280  0.87046463  0.88957625
[10]  1.15749590 -0.77569322  0.12267276
[13]  0.74534774  0.47894066 -2.14755773
[16]  0.13479327 -0.43325201  0.02388258
[19]  0.29187521 -0.88409876 -1.04946767
[22]  0.32642121 -0.71705350 -1.39180631
[25] -0.17871056  0.62159094 -0.93195104
[28]  1.16389431  1.07986841 -0.61792838

> t.test(x)

    One Sample t-test

data:  x
t = -1.1622, df = 29, p-value = 0.2546
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.5445612  0.1499301
sample estimates:
 mean of x
-0.1973155
```

t.test()는 이 코드의 실행 결과에서 보인 것처럼 평균에 대한 신뢰 구간 및 가설 검정 결과를 한번에 보여준다. 실행 결과 모평균은 -0.1973155, 모평균의 95% 신뢰구간은 (-0.5445612 , 0.1499301)로 추정되었다. 기본 인자에 의한 귀무가설은 'H_0 모평균이 0이다'고, p-value가 0.2546로 0.05보다 귀무가설을 기각하지 못하므로 모집단의 평균은 0으로 본다. 이는 95% 신뢰 구간에 0이 포함되어 있다는 것으로도 알 수 있는 내용이다.

다음은 평균이 10, 분산이 1인 정규 분포 $N(10,1)$ 에서 30개의 표본을 뽑아 모평균의 구간을 추정한 예다.

```
> x<-rnorm(30,mean = 10)
> t.test(x,mu=10)

    One Sample t-test

data:  x
t = -0.081372, df = 29, p-value = 0.9357
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
  9.626154 10.345237
sample estimates:
mean of x
 9.985695
```

t.test() 수행 결과 표본 평균은 9.985695, 평균의 신뢰 구간은 (9.626154, 10.345237)이다. t.test()에 mu = 10을 지정했으므로 모평균이 10인지에 대한 가설 검증이 수행되었는데, p-value > 0.05며 신뢰 구간이 10을 포함하므로 귀무가설 '$H_0 : mu = 10$ ' 을 기각하지 못한다.

이 절의 방법은 데이터가 정규 분포로부터 나온 것임을 가정하고 식 7.9를 사용해 구간 추정을 수행했다. 앞서 보인 예제 코드들에서는 rnorm ()을 사용해 데이터를 생성했으므로 데이터가 정규분포로부터 나온 표본임이 보장 되었다. 그러나 데이터가 정규 분포를 따르는지 불명확한 경우에는 샤피로 윌크 검정, 콜모고로프 스미로느프 검정, Q-Q도를 사용해 데이터의 정규성을 검토할 수 있다.

7.7.2 독립 이표본 평균

이 절에서는 두 모집단에서 각각 표본을 추출한 뒤 표본으로부터 두 모집단의 평균이 같은지 알아보는 내용을 다룬다. 예를 들어, A학교 학생의 평균 키와 B학교 학생의 평균 키를 비교하기 위해 각 학교에서 100명씩 총 200명을 뽑아 키를 측정하여 두 학교 학생 키의 평균에 차이가 있는지 알아보는 경우를 생각해볼 수 있다.

이론적 배경

독립 이표본은 서로 독립인 두 개의 표본 집단이 있는 경우를 지칭한다. $X_1, X_2, \cdots, X_m$ 이 $N(\mu_1, \sigma_1^2)$ 으로부터의 표본이고, $Y_1, Y_2, \cdots, Y_m$ 이 $N(\mu_2, \sigma_2^2)$으로부터의 표본이라고 하자. X,Y는 독립이다. 그러면 표본의 평균 X,Y는 다음의 정규 분포를 따른다.

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{m})$$

$$\bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n})$$

X와 Y가 독립이므로 다음이 성립한다.

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim N(0,1) \quad \text{----------식 7.10}$$

일표본 평균의 경우와 마찬가지로 σ_1^2, σ_2^2은 모집단의 분산이고, 이 값은 보통 사전에 알려지지 않은 값이다. 따라서 표본 분산을 대신 사용하게 된다. 표본 분산을 구할 때는 $\sigma_1 = \sigma_2$ 인 경우와 $\sigma_1 \neq \sigma_2$ 를 나누어 풀이한다 . 여기서는 모분산이 같은 경우 ($\sigma_1 = \sigma_2$) 만 알아보자.

$\sigma_1 = \sigma_2$ 인 경우, 합동 표본 분산 S_p^2 을 구하여 σ_1, σ_2 대신 사용한다. 합동 표본 분산은 두 집단의 표본 분산이 각각 S_1, S_2 일 때 다음과 같다.

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

S_p^2를 식 7.10에 대압히면 다음과 같이 자유도가 m+n-2인 t분포를 따르게 된다.

$$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_p\sqrt{1/m+1/n}} \sim t(m+n-2)$$

따라서 두 표본으로부터 구한 모평균 사이의 95% 신뢰 구간은 다음과 같다.

$$(\bar{X}-\bar{Y}-t(m-n-2;\alpha/2)S_p\sqrt{1/m+1/n}, \bar{X}-\bar{Y}+t(m+n-2;\alpha/2)S_p\sqrt{1/m+1/n})$$

추정 및 검정의 예

help(t.test)에 있는 예제를 살펴보자. t.test()의 도움말에 나와 있는 예제는 sleep 데이터를 사용한다. sleep 데이터는 최면성 약물을 10명의 환자에게 투여했을 때 수면 시간의 증가를 기록한 데이터다. 각 컬럼의 의미는 다음과 같다.

- extra : 수면 시간의 증가량
- group : 사용한 약물의 종류 =
- ID : 환자 식별 번호

다음은 sleep 데이터 전체의 내용이다.

```
> sleep
   extra group ID
1    0.7     1  1
2   -1.6     1  2
3   -0.2     1  3
4   -1.2     1  4
5   -0.1     1  5
6    3.4     1  6
7    3.7     1  7
8    0.8     1  8
9    0.0     1  9
10   2.0     1 10
11   1.9     2  1
12   0.8     2  2
13   1.1     2  3
14   0.1     2  4
15  -0.1     2  5
16   4.4     2  6
17   5.5     2  7
18   1.6     2  8
19   4.6     2  9
20   3.4     2 10
```

예를 들어, 1번 수면제를 사용했을 때 환자 1의 수면 시간 증가량은 0.7이었고, 2번 수면제를 사용했을 때 환자 1의 수면시간 증가량은 1.9 였다. 이처럼 동일한 대상에 대해 서로 다른 처방을 했을 때의 비교는 뒤에서 설명할 짝지은 이표본 평균의 비교 문제다. 그러나 여기서는 독립 이표본을 살펴보기 위해 환자 식별 번호가 없다고 가정해보자.

```
> sleep2 <- sleep[,-3]
> sleep2
   extra group
1    0.7     1
2   -1.6     1
3   -0.2     1
4   -1.2     1
5   -0.1     1
6    3.4     1
7    3.7     1
```

```
8    0.8     1
9    0.0     1
10   2.0     1
11   1.9     2
12   0.8     2
13   1.1     2
14   0.1     2
15  -0.1     2
16   4.4     2
17   5.5     2
18   1.6     2
19   4.6     2
20   3.4     2
```

이제 데이터의 의미가 달라졌다. 1번 행은 어떤 환자인지는 모르지만 누군가가 수면제 1을 복용했더니 수면 시간이 0.7 증가했다는 의미다. 마찬가지로 11번 행 역시 누군가에게 수면제 2를 투여했더니 수면 시간이 1.9 증가했다는 의미다. 이처럼 두 그룹이 있고 각 그룹에서의 표본이 상대 그룹에서의 표본과 아무런 상관관계가 없는 관찰 결과가 독립 이표본 추정 및 검정 대상에 해당한다.

수면제별 수면 시간 증가량의 평균을 계산해보자. 다음은 tapply()를 사용한 예다.

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

doBy::summaryBy()를 사용할 수도 있다.

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

이 장에서는 모분산이 같은 경우만 살펴보기로 했으므로 모분산이 같은지 먼저 검정한다. 분산의 비교에 대한 상세 설명은 이미 했다. 여기서는 분산 비교를 수행하는 var.test() 함수의 결과만 활용하자.

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

    F test to compare two variances

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

var.test()는 두 집단의 분산비 σ_1/σ_2 에 대한 가설 검정을 수행한다. var.test() 수행 결과 p-value가 0.05보다 커서 귀무가설 'H_0 : 분산의 비가 1이다'를 기각할 수 없다. 또는 신뢰 구간의 개념을 활용해 분산비의 95% 신뢰 구간이 (0.198297, 3.214123) 으로 그 안에 1이 포함되어 있어 분산의 비가 1임을 반박할 증거가 없다고 읽어도 된다.

이제 $\sigma_1 = \sigma_2$ 라고 보고 t.test()를 적용해보자. t.test의 주요 인자에는 paired, var.equal이 있다. paired = FALSE는 독립 이표본 검정을 뜻하고, paired = TRUE는 짝지은 이표본 검정을 뜻한다. var.equal은 두 집단의 모분산이 같은지 여부를 뜻한다.

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

    Two Sample t-test

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

분석 결과 p-value > 0.05 이므로 'H_0 : 모평균에 차이가 없다'는 가설을 기각할 수 없다. 또는 신뢰 구간이 0을 포함하여 평균에 차이가 없다고 읽어도 된다.

7.7.4 짝지은 이표본 평균

짝지은 이표본 평균은 표본을 추출할 때 연관된 것끼리 짝을 지어 데이터를 추출하는 경우를 말한다. 예를 들어, 앞 절에서 살펴본 sleep 데이터는 두 가지 약물을 10명에게 투여하여 수면시간의 증가를 측정한 것이다. 이때 각 환자마다 약물1로 인한 수면 시간 증가, 약물 2로 인한 수면시간 증가가 측정되었다. 따라서 sleep 데이터로부터 서로 다른 약물의 수면 시간 증가폭의 차를 개인별로 알아볼 수 있다. 그리고 이 개인별로 측정한 값의 평균을 구하면 두 약물 중 어느 것이 더 많은 수면 시간 증가를 가져오는지 알 수 있다. 이처럼 개인별로 측정한 수면 시간 증가량의 차를 구한 뒤 다시 평균을 구하면, 수면 시간 증가폭 계산 시 개인의 신체적 특성으로 인한 차이가 상쇄되어 좀 더 정확한 비교가 가능하다.

또 다른 예로 다이어트 약의 효과를 보기 위해 50명의 표본을 조사하는데, i번째 사람에 대해 X_i

에 대해 X_i에는 약물 섭취 전의 체중,Y_i에는 약물 섭취 후의 체중을 측정해 (X_i, Y_i) 형태로 기록했다면 이 역시 짝지은 이표본에 해당한다.

이론적 배경

짝지은 이표본 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ 이 있을 때 $D_i = X_i - Y_i$ 가 정규 분포를 따른다고 가정하자. 그러면 D의 평균 $\bar{D}$ 는 다음의 정규 분포를 따른다.

$$\bar{D} \sim N(\mu_D, \sigma_D^2/n)$$

그러나 σ_D 는 모분산이므로 보통 알려지지 않은 값이다. 이를 표본 분산으로 치환하면 t분포를 따른다.

$$\frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} \sim t(n-1)$$

$\bar{D}$의 95% 신뢰 구간은 $\alpha = 0.05$ 일 때 다음과 같다.

$$(\bar{D} - t(n-1;\alpha/2)S_D/\sqrt{n}, \bar{D} + t(n-1;\alpha/2)S_D/\sqrt{n})$$

추정 및 검정의 예

sleep 데이터를 다시 써보자. 그룹별 평균을 구하는 방법도 해당 절을 참고해라.

sleep 데이터는 다음과 같이 수면제 1(group==1) 과 수면제 2(group ==2) 각각에 대해 환자 ID가 오름차순으로 정렬되어 있다. 따라서 수면제별로 데이터를 잘라냈을 때 수면제 1과 수면제 2의 환자가 동일한 순서로 오게 된다.

```
> sleep
   extra group ID
1    0.7     1  1
2   -1.6     1  2
3   -0.2     1  3
4   -1.2     1  4
5   -0.1     1  5
6    3.4     1  6
7    3.7     1  7
8    0.8     1  8
9    0.0     1  9
10   2.0     1 10
11   1.9     2  1
12   0.8     2  2
13   1.1     2  3
14   0.1     2  4
15  -0.1     2  5
16   4.4     2  6
17   5.5     2  7
18   1.6     2  8
19   4.6     2  9
20   3.4     2 10
```

t.test()에 pair=TRUE를 지정하고 짝지은 이표본 검정을 수행해보자. 앞서 설명한 것처럼 그룹 별로 데이터를 잘라냈을 때1,2,3, ..., 10 환자 순서로 t.test()의 인자로 넘겨지고 있다.

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

    Paired t-test

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

p-value < 0.05 이므로 귀무가설 'H_0' : 모평균의 차이가 0이다' 를 기각한다. 따라서 두 수면제의 수면 시간 증가 정도가 다르다고 결론 내린다.

이 결론은 sleep 데이터를 독립 이표본으로 본 경우와 다른 결과다. 독립 이표본의 경우에는 신뢰 구간이 (-3.363874 , 0.203874) 로 구간 안에 0을 가까스로 포함한 형태 였다. 짝지은 이표본 검정의 경우 독립 이표본 검정에 비해 추정의 정확도가 높아 신뢰 구간이 좁아지면서 신뢰 구간에서 0이 빠지게 됐다. 그 결과 수면제 간 수면 시간 연장 정도에 차이가 존재함을 보였다.

7.7.5 이표본 분산

이 절에서는 두 모집단으로부터의 표본으로부터 분산을 구해 두 모집단의 분산이 동일한지를 알아보는 방법에 대해 설명한다. 보통은 이표본 분산을 그 자체로 사용하기보다는 다른 통계 추정, 검정에서 이표본 분산의 결과를 사용한다. 예를 들어, 독립 이표본 검정에서 두 모집단의 분산이 같으면, t.test() 인자로 var.equal = TRUE를, 다르면 var.equal = FALSE를 지정해야 한다고 설명한 바 있다.

이론적 배경

확률 변수 X,Y가 독립이며 X~$N(\mu_1, \sigma_1^2), Y$~$N(\mu_2, \sigma_2^2)$으로 각각 정규 분포를 따른다고 가정하자. m은 X에서의 표본 수, n은 Y에서의 표본 수이다. 이때 표본 분산과 모분산의 비가 다음과 같이 F분포를 따른다.

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(m-1, n-1)$$

따라서 모분산 비에 대한 95% 신뢰 구간은 $\alpha = 0.05$ 라 할 때 다음과 같다.

$$(\frac{S_1^2/S_2^2}{F(n-1, m-1; \alpha/2)}, \frac{S_1^2/S_2^2}{F(n-1, m-1; 1-\alpha/2)})$$

추정 및 검정의 예

분산의 비교에는 var.test() 함수를 사용한다.

var.test() : 분산 비교를 하기 위한 F 검정을 수행한다. 귀무가설은 '모분산 비가 ratio와 같다'이다.

iris의 Sepal.Width와 Sepal.Length가 같은지 var.test()를 사용하여 검정해보자.

```
> with(iris, var.test(Sepal.Width, Sepal.Length))

    F test to compare two variances

data:  Sepal.Width and Sepal.Length
F = 0.27706, num df = 149, denom df = 149,
p-value = 3.595e-14
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2007129 0.3824528
sample estimates:
ratio of variances
         0.2770617
```

수행결과 p-value가 매우 작게 나타났다. 따라서 모분산에 차이가 없다(ratio = 1) 는 귀무가설을 기각한다.

7.7.6 일표본 비율

모집단에서 표본을 추출하여 그 표본에서 계산한 비로부터 모집단의 비를 추정 및 가설 검정하는 내용에 대해 살펴보자. 일표본 비율의 예로는 국민 투표로 찬반 투표를 실시할 때 출구 조사를 들 수 있다. 투표자 중 임의의 표본(예를 들면, 1,000명의 유권자)에게 찬성과 반대 중 어느 쪽에 투표했는지를 물으면 찬성과 반대 중 어느 쪽의 비율이 크게 나타날지 추정할 수 있다.

이론적 배경

베르누이 시행을 n회 수행하여 X번 성공했다고 하자.

Note 베르누이 시행

베르누이 시행이란 '성공', '실패' 두 가지 결과만 있는 실험을 말한다. 실험을 여러 번 반복할 때 성공의 확률은 매번 일정하다.

이때 X는 성공 확률이 p인 베르누이 시행을 n회 수행했을 때 성공 횟수를 뜻하는 이항 분포를 따른다. 이를 다음과 같이 표현한다.

$$X \sim B(n, p)$$

일표본 비율에서는 p를 구하는 것이 목적이다.

이항 분포 B(n,p)는 평균이 np, 분산이 np(1-p)이며, n이 크면 정규 분포로 근사할 수 있다.

$$X \sim N(np, np(1-p))$$

모비율에 대한 추정값 $\hat{p}$ 는 X/n으로 계산할 수 있으므로 위 식의 양변을 n으로 나누면 다음과 같다.

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

따라서 $\alpha = 0.05$ 라 할 때, 모비율의 95% 신뢰 구간은 다음과 같다.

$$(\hat{p} - Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$$

추정 및 검정의 예

비율에 대한 가절 검정 시 이항 분포의 정규 분포 근사를 사용할 경우 prop.test()를 사용한다. 반면 이항 분포를 정규 분포로 근사하지 않고 그대로 사용해 계산하고 싶다면 binom.test()를 사용한다.

prop.test : 비율에 대한 가설 검정을 수행한다. 귀무가설은 '두 그룹의 비율이 같다' 또는 '비율이 p와 같다'이다.

동전을 100번 던졌더니 앞면이 42번 나왔다고 하자. 이때 동전의 앞면이 나오는 비율이 50%가 아니라고 할 수 있을까? prop.test() 함수로 확인할 수 있다.

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

    1-sample proportions test with continuity
    correction

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

수행 결과 p-value > 0.05이므로 동전의 앞면이 나올 확률이 0.5라는 귀무가설을 기각할 수 없다. 이는 신뢰 구간 (0.32,0.52)에 0.5가 포함되어 있는 것으로도 확인할 수 있다.

비율 구간 추정 또는 검정 시 반드시 정규 분포 등으로 근사해야 하는 것은 아니다. binom.test()를 사용하면 이항분포로부터의 신뢰구간을 직접 계산한다.

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

    Exact binomial test

data:  42 and 100
number of successes = 42, number of trials
= 100, p-value = 0.1332
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3219855 0.5228808
sample estimates:
probability of success
                  0.42
```

이항분포를 통한 정확한 계산의 경우에도 동전의 앞면이 나올 확률이 0.5라는 귀무가설을 기각하지 못한다.

7.7.7 이표본 비율

이표본 비율은 두 집단에서 표본을 추출해 표본의 비율을 보고 모집단에서의 비율을 비교하는 경우다. 예를 들어, 남성의 흡연율과 여성의 흡연율에 차이가 있는지를 비교하기 위해 남성 100명, 여성 100명의 흡연율을 계산하고 이로부터 남성의 흡연율과 여성의 흡연율에 차이가 있는지 가설 검정하는 경우를 생각해볼 수 있다.

이론적 배경

독립인 두 집단 X, Y가 이항 분포를 따른다고 하자.

$$X \sim B(n_1, p_1)$$

$$Y \sim B(n_2, p_2)$$

n이 충분히 클 때 X,Y가 근사적으로 정규 분포를 따른다. 다음 식

$\hat{p} \sim N(p, p(1-p)/n)$를 X,Y에 적용하고 비율의 차를 구하면 그 역시 정규 분포를 따른다.

따라서 $p_1 - p_2$의 95% 신뢰 구간은 $\alpha = 0.05$ 라 할 때 다음과 같다.

$$(\hat{p}_1 - \hat{p}_2 - Z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2},$$

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

추정 및 검정의 예

두 개의 동전을 각각 100회, 90회 던졌을 때 각각 앞면이 45회, 55회 나왔다고 하자. 이때 두 동전의 앞면이 나올 확률이 같은지 검정해보자.

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

    2-sample test for equality of proportions
    with continuity correction

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

p < 0.05가 나와 두 동전의 앞면이 나올 확률이 같다는 가설을 기각한다. 즉, 두 동전의 앞면이 나올 확률은 서로 다르다.