연습문제 1장

1. 기술통계학이란 무엇인지 설명하라.
sol) 수집된 자료를 기술 • 요약하는 기초적인 통계

2. 추론통계학이란 무엇인지 설명하라.
sol) 표본으로부터 가설검정과 신뢰구간 등 모집단에 대한 추론을 다루는 통계

3. 다음 용어를 각각 정의하라.
sol) 1) **변수**: 데이터를 수집할 때 관찰을 하는 장소나 사람, 그 외에도 관련된 조건에 따라 다양한 값을 가질 수 있는 것
2) **양적변수**: 양(quantity)와 관련된 값을 갖는 변수
3) **질적변수**: 양적인 값이 아닌, 어떠한 범주에 속하는지를 나타내는 값을 갖는 변수
4) **확률변수**: 변수의 관측값이 미리 결정된 것이 아니라, 가질 수 있는 값 중에서 (우연히) 선택된 값을 갖는 변수
5) **모집단**: 개체의 측정값들의 전체집합을 의미하며, 유한 모집단과 무한 모집단으로 나눠진다.
6) **표본**: 모집단의 부분집합으로서 모집단으로부터 측정한 측정값의 집합을 의미
7) **이산형 변수**: 관측값이 이산형인 변수로, 관측값이 가질 수 있는 값의 종류가 유한하거나 자연수만큼 존재하는 경우를 이산형이라 한다. 이 때 크기순으로 나열한 모든 관측값 사이에 무수히 많은 실수가 존재한다.
8) **연속형 변수**: 관측값이 연속형인 변수로, 관측 가능한 값을 크기순으로 나열했을 때 연속한 두 값 사이에 실수가 존재하지 않는 경우를 연속형이라 한다.
9) **단순랜덤표본**: 크기가 n인 모든 가능한 표본이 뽑힐 확률이 동일하게 표본을 뽑는 방법에 의해서 얻어진 표본
10) **복원추출**: 한 번 뽑은 개체를 다시 뽑을 수 있는 추출 방법
11) **비복원추출**: 한번 뽑은 개체는 다시 뽑지 않는 추출 방법

4. 측정이란 무엇인지 정의하라.
sol) 적절한 과정에 따라 변수의 값을 얻는 과정을 의미하며 이렇게 얻어진 값을 측정값이라 한다.

5. 네 가지 척도를 나열하고, 각각을 설명하라.
sol)

척도	설명
명목척도	이름과 같이 분류가 목적인 경우의 측정 척도
순서척도	측정 대상을 분류하는 동시에 측정값을 가질 수 있는 값들 사이에 순서 개념이 포함되어 있어서 대소 비교가 가능한 척도
구간척도	측정 대상의 속성에 숫자를 부여하되, 숫자 사이의 간격을 비교할 수 있는 척도. 두 측정값 사이의 차이를 계산하여 비교할 수 있다는 점에서 순서척도에 비해 더 많은 정보를 가지는 척도
비척도	숫자 사이의 간격을 비교할 수 있고, 영점이 절대적인 의미를 갖는 척도로 가감승제의 모든 연산이 가능.

6. 다음의 각 변수가 양적변수인지 질적변수인지, 그리고 관측값은 어떠한 척도에 해당하는지 설명하라.

sol) 1) 한 반에서의 시험 성적 순위: 양적변수, 비척도
2) 입원한 사람들의 입원진단: 질적변수, 명목척도
3) 고등학생의 체중: 양적변수, 비척도
4) 성별: 질적변수, 명목척도
5) 관절 부상자들의 운동 범위 : 양적변수, 비척도
6) 체온: 양적변수, 구간척도

연습문제 2장

1. 다음을 정의하라.

sol) 1) **줄기-잎 그림**: 개개의 관측값을 줄기와 잎으로 분리한 그래프로, 관측값의 십진수의 각 단위 숫자 중 상위 단위 숫자가 줄기이고 나머지 숫자는 잎이다.

2) **상자그림**: 데이터의 사분위수를 이용하여 그린 그래프로 데이터의 산포, 중위수의 위치, 대칭성에 대한 정보를 제공한다.

3) **백분위수**: 관측값을 크기 순서로 배열했을 때 백분율로 나타낸 특정 위치의 값을 의미

4) **사분위수**: 관측값을 크기 순서로 배열했을 때 전체 관측값을 4등분한 점에 해당하는 값을 의미하며, 1사분위수, 2사분위수, 3사분위수로 구성된다.

5) **위치모수**: 도수 분포의 x-축의 위치와 관련된 모수를 의미하며, 평균과 중위수 등이 포함된다.

6) **도수분포**: 적절히 나눠진 계급구간에 포함되는 관측값들의 빈도를 도수라 하며, 이 도수의 분포 형태를 도수분포라 한다.

7) **상대도수분포**: 각 계급구간의 도수를 전체 도수의 합으로 나눈 비율을 상대도수라 하며, 이 상대도수의 분포형태를 상대도수분포라 한다.

8) **통계량**: 표본으로부터 계산된 기술통계량

9) **모수**: 모집단으로부터 얻어진 기술통계량

10) **도수다각형**: 선 그래프의 특별한 형태로, 히스토그램에서 각 계급 구간의 막대 윗부분의 중간점들을 직선으로 연결한 그래프

11) **히스토그램**: 도수분포 혹은 상대분포를 막대그래프의 형태로 표현한 그래프로 y-축은 도수 혹은 상대도수를 의미하며, x-축은 각 계급구간을 사각형의 막대로 표시한다.

2. 평균, 중위수, 최빈값을 정의하고 각각의 특징을 비교하라.

sol)

통계량	정의	특징
평균	모집단 또는 표본에 존재하는 모든 값을 더한 뒤에 그 측정값의 개수로 나눈 값	-평균과 중위수는 모든 데이터에서 오직 하나만 존재하는 반면 최빈값은 존재하지 않을 수도 있고, 2개 이상 존재할 수도 있다. -평균은 극단적인 값에 민감하지만 중위수는 평균에 비하여 이상치에 덜 민감하다.
중위수	데이터를 크기 순서대로 배열하였을 때 중앙에 위치한 값	
최빈값	관측값 중 가장 빈도가 높은 값	

3. 산포를 측정하는 데 사용하는 범위의 장점과 단점에 대하여 논하라.

sol)

장점	산포(관측값들의 흩어진 정도)를 측정하는 통계량 중 가장 계산이 간단하다.
단점	최솟값과 최댓값만 이용하여 계산하기 때문에 산포에 대하여 제한된 정보만을 제공한다.

4. 표본 분산 계산 시 $n-1$을 사용하는 이유에 대하여 설명하라.

sol) 관측값의 수가 아닌 자유도로 값을 나눠야 하기 때문이다. (불편성 만족)
이때, $n-1$개의 편차를 알고 있으면 마지막 편차는 자동으로 결정되므로 자유도는 $n-1$이다.

5. 변동계수는 어떠한 경우에 사용하는지 설명하라.

sol) 측정 단위가 다른 데이터의 산포를 비교할 때, 상대적 변동인 변동계수를 사용한다.

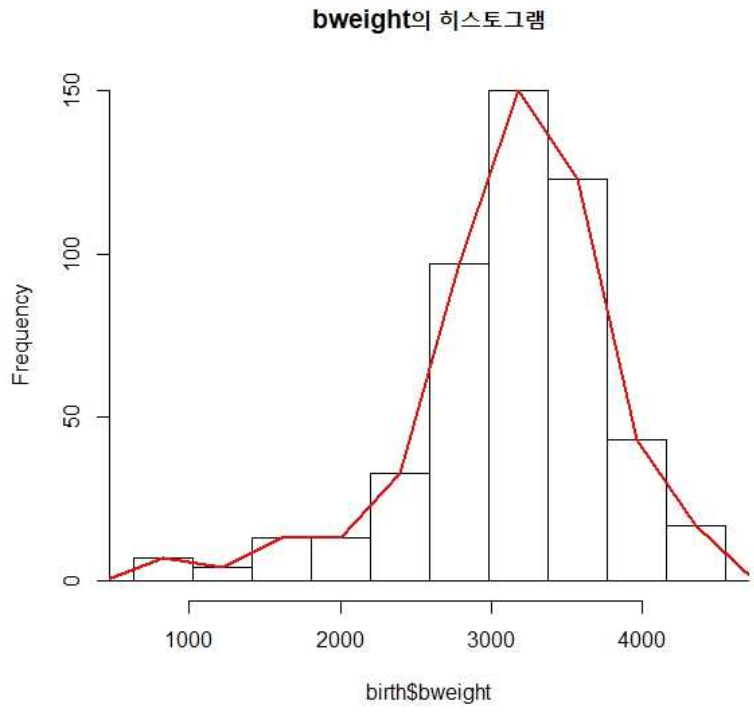
6. 50번째 백분위수(제2사분위수)를 다른 용어로 표현하라.

sol) 중위수(median)

7. [표 2.2.1]의 birth data를 이용하여 bweight의 데이터로 도수분포, 누적도수분포, 상대도수분포, 누적상대분포, 히스토그램, 도수다각형을 그려라.

sol)

```
> k<-1+3.322*log10(length(birth$bweight))
> k<-round(k)
> r<-max(birth$bweight)-min(birth$bweight)
> w<-r/k
> Freq<-table(cut(birth$bweight, breaks = seq(min(birth$bweight),max(birth$bweight),w),
                  dig.lab = 5, include.lowest = T))
> cbind(Freq, Cum_freq=cumsum(Freq), Rel_freq=prop.table(Freq),
  Cum_rel_freq=cumsum(prop.table(Freq)))
                Freq Cum_freq Rel_freq Cum_rel_freq
[628,1020.5]       7        7    0.014        0.014
(1020.5,1413]      4       11    0.008        0.022
(1413,1805.5]     13       24    0.026        0.048
(1805.5,2198]     13       37    0.026        0.074
(2198,2590.5]     33       70    0.066        0.140
(2590.5,2983]     97      167    0.194        0.334
(2983,3375.5]    150      317    0.300        0.634
(3375.5,3768]    123      440    0.246        0.880
(3768,4160.5]     43      483    0.086        0.966
(4160.5,4553]     17      500    0.034        1.000
> h<-hist(birth$bweight, breaks = seq(min(birth$bweight),max(birth$bweight),w),
           main = "bweight의 히스토그램")
> library(agricolae)
> polygon.freq(h, frequency=1, col="red", lwd=2)
```



8. [표 2.2.1]의 birth data를 이용하여 다음 물음에 답하라.

sol) 1) 남성(sex=1)의 bweight 변수의 평균, 중위수, 분산, 표준편차를 답하라.

```
> male_birth<-subset(birth, subset = sex==1)
> male_mean<-mean(male_birth$bweight)
> male_median<-median(male_birth$bweight)
> male_deviation<-sd(male_birth$bweight)
> male_variance<-var(male_birth$bweight)
> cbind(male_mean, male_median, male_variance, male_deviation)
     male_mean male_median male_variance male_deviation
[1,]  3229.902        3296      401503.2       633.6428
```

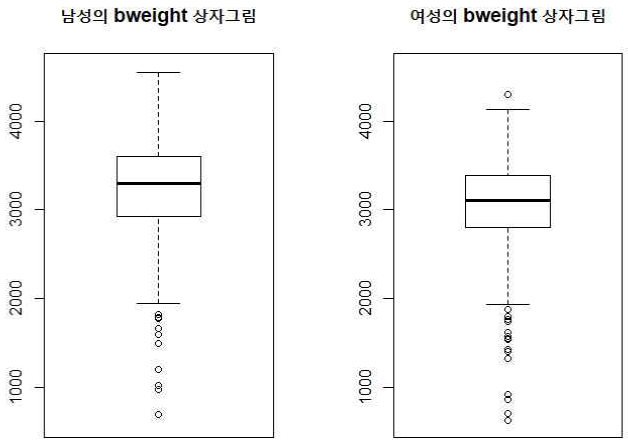
2) 여성(sex=2)의 bweight 변수의 평균, 중위수, 분산, 표준편차를 답하라.

```
> female_birth<-subset(birth, subset = sex==2)
> female_mean<-mean(female_birth$bweight)
> female_median<-median(female_birth$bweight)
> female_variance<-var(female_birth$bweight)
> female_deviation<-sd(female_birth$bweight)
> cbind(female_mean, female_median, female_variance, female_deviation)
     female_mean female_median female_variance female_deviation
[1,]    3032.831          3107        392898.3          626.816
```

3) 남성(sex=1)의 bweight 변수의 상자그림을 그려라.

4) 여성(sex=2)의 bweight 변수의 상자그림을 그려라.

```
> par(mfrow=c(1,2))
> boxplot(male_birth$bweight, main="남성의 bweight 상자그림", ylim=c(600,4600))
> boxplot(female_birth$bweight, main="여성의 bweight 상자그림",ylim=c(600,4600))
```



5) 여성과 남성 간의 bweight에 차이가 있다고 볼 수 있는지 설명하라.

차이가 있다고 할 수 있다. 남성의 bweight 변수의 중위수와 평균이 여성에 비해 모두 높으며 분산과 표준편차 또한 더 크다. 이는 상자그림을 통해서도 확인할 수 있는데 남성의 bweight의 분포가 여성에 비하여 위쪽에 분포하고 있으며, 분포의 길이도 여성에 비하여 더 긴 것을 확인할 수 있다.

9. [표 2.2.1]의 birth data를 이용하여 bweight 변수를 이용하여 다음 질문에 답하라.

sol) 1) 줄기-잎 그림을 그려라

```
> stem(birth$bweight, scale=2)

  The decimal point is 2 digit(s) to the right of the |

   6 | 39
   7 | 1
   8 | 6
   9 | 28
  10 | 2
  11 |
  12 | 0
  13 | 23
  14 | 03
  15 | 0457
  16 | 026
  17 | 4689
  18 | 0278
  19 | 45
  20 | 0099
  21 | 0359
  22 | 12567
  23 | 0345669
  24 | 0012233589
  25 | 01145556889
  26 | 00111122224567789
  27 | 0000000222233444566778
  28 | 000000334444556667889999999
  29 | 11112222334444455556667777778889999
  30 | 000001111223444444555566666888888899999
  31 | 0000001122222333444455556666788888888899999
  32 | 000012233344555566666777888999999
  33 | 0001111222223334444455555677788888999999
  34 | 000111223333455555566666777899
  35 | 00000112223445555555566667777778888899999
  36 | 00111223345556678889
  37 | 000022233333444455567777777788
  38 | 00011235889
  39 | 12344456889
  40 | 0223446779
  41 | 233488
  42 | 01239
  43 | 0024
  44 | 24
  45 | 0125
```

2) 줄기-잎 그림을 토대로 데이터 분포의 특성을 설명하라.

줄기-잎 그림을 보면 이상치를 제외하면 가운데를 기준으로 좌우 대칭되어 있는 형태를 띄고 있다.

3) 평균, 중위수, 분산, 표준편차를 구하라.

```
> bweight_mean<-mean(birth$bweight)
> bweight_median<-median(birth$bweight)
> bweight_variance<-var(birth$bweight)
> bweight_deviation<-sd(birth$bweight)
> cbind(bweight_mean, bweight_median, bweight_variance, bweight_deviation)
     bweight_mean bweight_median bweight_variance bweight_deviation
[1,]     3136.884          3188.5         406344.4           637.4515
```

10. 다음의 성질을 증명하라.

sol) 1) $\min_a \sum(x-a)^2 = \sum(x-\bar{x})^2$

$$\sum(x-a)^2 = \sum(x^2-2ax+a^2) = \sum x^2 - 2a\sum x + na^2$$

$$= n\left(a^2 - \frac{2a}{n}\sum x\right) + \sum x^2$$

$$= n\left\{a^2 - \frac{2a}{n}\sum x + \left(\frac{1}{n}\sum x\right)^2 - \left(\frac{1}{n}\sum x\right)^2\right\} + \sum x^2$$

$$= n\left(a - \frac{1}{n}\sum x\right)^2 - \frac{1}{n}(\sum x)^2 + \sum x^2$$

이므로 a가 $\bar{x}\left(=\frac{1}{n}\sum x\right)$일 때 최소가 됨.

2) $n\bar{x} = \sum x$

$$n\bar{x} = n\left(\frac{1}{n}\sum x\right) = \sum x$$

3) $\sum(x-\bar{x}) = 0$

$$\sum(x-\bar{x}) = \sum x - \sum\bar{x} = \sum x - n\bar{x}$$
$$= \sum x - \sum x \quad (\because by\ (2))$$
$$= 0$$

11. 보건통계학 수업의 학점은 5번의 시험 결과로 결정된다. 5개 성적의 평균과 중위수 중 어떠한 통계량을 이용하여 최종 학점을 결정하는 것이 더 적절한가? 그리고 그 이유를 설명하라.

sol) 개별 성적의 값에 영향을 받는 평균으로 최종 학점을 결정하는 것이 더 적절하다.
5번의 시험 결과가 평균에 모두 반영되기 때문이다.

12. 다음은 집단선별검사에 참가한 대상들의 혈청 콜레스테롤 지수의 도수분포를 설정하는 데 가능한 계급 폭을 보여준다.

(a)	(b)	(c)
50~74	50~74	50~75
75~99	75~99	75~100
100~149	100~124	100~125
150~174	125~149	125~150
175~199	150~174	150~175
200~249	175~199	175~200
250~274	200~224	200~225
	225~249	225~250

세 가지 계급 구간 중 분석 목적에 가장 적합한 것은 무엇인가? 선택에 대한 적절한 근거를 마련하라.

sol) (b)가 가장 적합하다. (a)의 경우 계급 구간의 폭이 일정하지 않으며, (c)의 경우 계급구간이 중복되기 때문에 적합하지 않다.

13. 모집단의 중심성향을 측정할 때, 중위수보다 평균을 더 많이 이용하는 이유는 무엇인지 설명하라.

sol) 평균은 모든 관측값을 이용하여 계산하지만 중위수는 대소 비교만 할 뿐 모든 관측값들을 계산에 이용하지는 않는다. 따라서 평균이 중위수보다 데이터를 많이 반영하기 때문에 더 많이 이용한다.

14. 중심성향에 대한 통계량으로서, 평균보다 중위수가 더 적절한 경우는 언제인지 설명하라.

sol) 극단적인 값인 이상치가 존재할 때, 이상치에 민감한 평균보다 중위수를 사용하는 것이 더 적절하다.

15. 다음 변수에 대하여 중심 성향을 측정하려고 할 때, 각 변수에 대해서 평균, 중위수, 최빈값 중 어떠한 통계량이 가장 적절한지 선택하고 그 이유를 설명하라.

sol) 1) 의사들의 연봉 ⇒ 중위수 (∵ 의사들의 연봉은 개인병원 의사부터 종합병원 의사까지 다양한 크기의 값을 갖기 때문에 극단적인 값에 영향을 받지 않는 중위수가 적절하다.)

2) 대도시 병원 응급실 환자의 진단명: 최빈값 (∵ 환자의 진단명과 같은 질적변수의 경우 평균과 중위수가 존재하지 않기 때문에 최빈값이 적절하다.)

3) 고등학교 남자 축구선수의 몸무게 ⇒ 평균 (∵ 남고생 축구선수의 몸무게는 비교적 범위가 작고, 이상치가 거의 없기 때문에 평균이 적절하다.)

16. [표 2.2.1]의 birth data의 matage변수에 대하여 평균, 중위수, 최빈값을 구하라. 또한 만약 세 값이 동일하지 않으면, 그렇지 않은 이유를 설명하라.

sol)

```
> mode <- function(x) {
+   ux <- unique(x)
+   ux[which.max(tabulate(match(x, ux)))]
+ }
> matage_mean<-mean(birth$matage)
> matage_median<-median(birth$matage)
> matage_mode<-mode(birth$matage)
> cbind(matage_mean, matage_median, matage_mode)
     matage_mean matage_median matage_mode
[1,]      34.028            34           34
```

⇒ 평균이 중위수와 최빈값에 비해 0.028정도 크다. birth data의 matage변수의 분포가 약간 우측으로 치우쳐졌기 때문이다.

사용된 R Code

```
install.packages("readxl")
library(readxl)
birth<-read_excel("C:/Users/Desktop/example_data.xlsx", sheet = 1, col_names = T)
head(birth)
str(birth)
View(birth)

###7#####
k<-1+3.322*log10(length(birth$bweight))
k<-round(k)
r<-max(birth$bweight)-min(birth$bweight)
w<-r/k
Freq<-table(cut(birth$bweight, breaks = seq(min(birth$bweight),max(birth$bweight),w),
                dig.lab = 5, include.lowest = T))
cbind(Freq, Cum_freq=cumsum(Freq), Rel_freq=prop.table(Freq), Cum_rel_freq=cumsum(prop.table(Freq)))
h<-hist(birth$bweight, breaks = seq(min(birth$bweight),max(birth$bweight),w), main = "bweight의 히스토그램")
library(agricolae)
polygon.freq(h, frequency=1, col="red", lwd=2)


#####8#####
male_birth<-subset(birth, subset = sex==1)
female_birth<-subset(birth, subset = sex==2)

male_mean<-mean(male_birth$bweight)
male_median<-median(male_birth$bweight)
male_variance<-var(male_birth$bweight)
male_deviation<-sd(male_birth$bweight)
cbind(male_mean, male_median, male_variance, male_deviation)

female_mean<-mean(female_birth$bweight)
female_median<-median(female_birth$bweight)
female_variance<-var(female_birth$bweight)
female_deviation<-sd(female_birth$bweight)
cbind(female_mean, female_median, female_variance, female_deviation)

par(mfrow=c(1,1))
boxplot(male_birth$bweight, main="남성의 bweight 상자그림", ylim=c(600,4600))
boxplot(female_birth$bweight, main="여성의 bweight 상자그림",ylim=c(600,4600))

#####9#####
stem(birth$bweight, scale=2)
bweight_mean<-mean(birth$bweight)
bweight_median<-median(birth$bweight)
bweight_variance<-var(birth$bweight)
bweight_deviation<-sd(birth$bweight)
cbind(bweight_mean, bweight_median, bweight_variance, bweight_deviation)

####10####
mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
matage_mean<-mean(birth$matage)
matage_median<-median(birth$matage)
matage_mode<-mode(birth$matage)
cbind(matage_mean, matage_median, matage_mode)
hist(birth$matage)
```

연습문제 3장

1. 고전적 확률의 개념을 정의하라.

sol) N개의 결과가 가능한 시행을 한다고 가정 할 때, N개의 결과는 동시에 일어날 수 없으며, 일어날 확률이 동일하다고 가정한다. 총 N개의 결과 중 m개가 사건 E에 포함되면 E가 일어날 확률은 m/N이다.

2. 통계적 확률의 개념을 정의하라.

sol) 반복 가능한 시행을 고려할 때, 시행을 n번 반복하고 이 중 m번의 시행 결과가 사건 E를 만족하면, E의 상대도수는 m/n이다. 이 때, n이 충분히 크면 상대도수는 확률과 비슷해지므로 이를 통계적 확률이라고 한다.

3. 조건부확률의 정의는 무엇인지 설명하라.

sol) 사건 A가 일어났다는 가정 하에 사건 B의 조건부확률이란 사건 A에 포함되는 결과 중에서 사건 B에도 포함되는 결과의 비율을 의미한다. (기호: $P(B|A)$)

4. 독립사건을 정의하라.

sol) 두 사건 A, B가 있을 때, 사건 A가 일어날 확률이 사건 B가 일어났는지 여부에 관련이 없을 때 두 사건을 독립사건이라 한다. 즉, $P(A|B)=P(A)$이면, 사건 A와 B는 독립이다.

5. 확률의 세 가지 공리를 설명하라.

sol) ① 확률은 사건을 대응시키는 함수로서, 임의의 사건 E_k가 일어날 확률은 음이 아닌 실수이다.
② 상호배반인 사건 $E_1, E_2, \cdots, E_n$에 대하여 $P(E_1)+P(E_2)+\cdots+P(E_n)=P(E_1 \cup E_2 \cup \cdots E_n)$이다.
③ 표본 공간의 확률은 1이다.

6. 베이즈 정리를 설명하라.

sol) 표본공간 S가 k개의 사건 $B_1, \cdots, B_k$로 분할된다고 하면, 사건 A가 일어났을 때 사건 B_j가 일어날 확률은 $P(B_j|A)=\dfrac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$로 표현할 수 있으며 이를 베이즈 정리라고 한다.

7. 민감도, 특이도를 정의하라.

sol)

민감도	환자들의 선별검사 결과가 양성일 조건부확률
특이도	정상인의 선별검사 결과가 음성일 조건부확률

8. 다음의 표를 이용하여 각각의 확률을 구하라.

	폐암(A)	비폐암(B)
흡연 (S)	10	40
비흡연 (N)	50	300

sol)

1) $P(S)=\frac{10+40}{10+40+50+300}=\frac{50}{400}=\frac{1}{8}=0.125$

2) $P(S^c)=1-P(S)=1-\frac{1}{8}=\frac{7}{8}=0.875$

3) $P(A)=\frac{10+50}{10+40+50+300}=\frac{60}{400}=\frac{3}{20}=0.15$

4) $P(A^c)=1-P(A)=1-\frac{3}{20}=\frac{17}{20}=0.85$

5) $P(S\cap A)=\frac{10}{10+40+50+300}=\frac{10}{400}=\frac{1}{40}=0.025$

6) $P(S|A)=\frac{P(S\cap A)}{P(A)}=\frac{0.025}{0.15}=0.167,\ P(S|B)=\frac{P(S\cap B)}{P(B)}=\frac{P(S\cap A^c)}{P(A^c)}=\frac{0.1}{0.85}=0.118$

7) $P(N\cap B^c)=P(N\cap A)=\frac{50}{10+40+50+300}=\frac{50}{400}=\frac{1}{8}=0.125$

9. 사건 A와 사건 B는 독립이며, $P(A)=0.3, P(B)=0.2$이다. $P(A\cap B)$를 구하라.

sol) $P(A\cap B)=P(A)P(B)=0.3\times 0.2=0.06$

10. 배반사건과 독립사건의 차이를 기술하라.

sol) 두 사건 A, B가 동시에 일어날 수 없는 경우를 배반사건이라고 하고,
두 사건 A, B가 동시에 일어날 수는 있지만 각각의 사건이 발생할 확률이 서로에게 영향을 미치지 않는 경우를 독립사건이라고 한다.

11. $P(A)\neq 0, P(B)\neq 1$인 경우, $A\subset B$라면 A와 B는 독립일 수 있는지 설명하라.

sol) $A\subset B$이므로 $0<P(A)\leq P(B)<1$이다. 만약 A, B가 독립이라면 $P(A\cap B)=P(A)P(B)$이고,
$0<P(A)P(B)<1$이므로 A와 B는 독립일 수 있다.

12. 다음 식의 참-거짓을 판단하라.

sol)

1) $P(A\cap C)=P(C\cap A)$ ……(참)
2) $P(A\cup B)=P(A)+P(B)$ ……(거짓)
3) $P(D|A)=P(D)$ ……(거짓)
4) $P(D|A)=P(D\cap A)$ ……(거짓)
5) $P(A\cap B)=P(A|B)P(A)$ ……(거짓)
6) $P(A\cap B)=P(A)P(B)$ ……(거짓)
7) $P(A)=P(A\cap B)+P(A\cap C)$, 단 $P(A\cup B\cup C)=1$ ……(거짓)

13. WIN이 적혀 있는 공이 1개, LOSE가 적혀 있는 공이 4개 있다고 하자. A, B 두 사람이 공을 선택하여 먼저 WIN을 뽑는 사람이 이긴다고 했을 때, A가 이길 확률을 구하라(단, A가 먼저 선택한다).

sol)

1) 복원추출인 경우: $\frac{1}{5}+\left(\frac{4}{5}\right)^2\frac{1}{5}+\left(\frac{4}{5}\right)^4\frac{1}{5}+\cdots=\sum_{i=0}^{\infty}\left(\frac{4}{5}\right)^{2i}\frac{1}{5}=\frac{\frac{1}{5}}{1-\left(\frac{4}{5}\right)^2}=\frac{\frac{1}{5}}{\frac{9}{25}}=\frac{5}{9}$

2) 비복원추출인 경우: $\frac{1}{5}+\left(\frac{4}{5}\right)\left(\frac{3}{4}\right)\left(\frac{1}{3}\right)+\left(\frac{4}{5}\right)\left(\frac{3}{4}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)\left(\frac{1}{1}\right)=\frac{1}{5}+\frac{1}{5}+\frac{1}{5}=\frac{3}{5}$

14. 주머니에 초록색 공 3개, 파란색 공 5개가 있다고 하자. 비복원추출로 공을 하나씩 뽑을 때, 첫 번째는 초록색이고, 두 번째는 파란색일 확률을 구하라.

sol) $\left(\frac{3}{8}\right)\times\left(\frac{5}{7}\right)=\frac{15}{56}$

15. 그릇 A에는 3개의 빨간 칩과 7개의 파란 칩이 있다. 그릇 B에는 빨간 칩 8개, 파란 칩 2개가 있다. 주사위를 던졌을 때 앞면이 5나 6이 나오면 A에서 하나의 칩을 선택하고 그렇지 않으면 B에서 하나의 칩을 선택한다. 빨간 칩이 뽑혔을 때, 그 칩이 A에서 뽑혔을 확률을 구하라.

sol)

- $\begin{cases}A: \text{그릇 } A\text{에서 하나의 칩을 선택하는 사건} = \text{주사위를 던졌을 때, 앞면이 5나 6이 나올 사건}\\ B: \text{그릇 } B\text{에서 하나의 칩을 선택하는 사건} = \text{주사위를 던졌을 때, 앞면이 1부터 4가 나올 사건}\\ R: \text{빨간 칩이 나올 사건}\end{cases}$

$$\Rightarrow \begin{cases}P(A)=\frac{2}{6}\\ P(B)=\frac{4}{6}\end{cases}$$

- $P(A|R)=\frac{P(A\cap R)}{P(R)}=\frac{P(R|A)P(A)}{P(R|A)P(A)+P(R|B)P(B)}=\frac{\frac{3}{10}\times\frac{2}{6}}{\frac{3}{10}\times\frac{2}{6}+\frac{8}{10}\times\frac{4}{6}}=\frac{\frac{6}{60}}{\frac{38}{60}}=\frac{3}{19}$

16. 조건부 확률이 확률의 세 가지 공리를 만족하는지 확인하라.

sol) ① 임의의 사건 A, B에 대하여 $P(A)>0, P(B)>0$일 때, $P(A\cap B)\geq 0$이다.

따라서 $P(A|B)=\frac{P(A\cap B)}{P(B)}\geq 0$이므로 확률의 첫 번째 공리를 만족한다.

② 상호배반인 $A_1, A_2, \cdots$에 대하여 $(A_i\cap B)\cap(A_j\cap B)=A_i\cap A_j\cap B=\varnothing\ \ (i\neq j)$이다.

따라서 $P(A_1\cup A_2\cup\cdots|B)=\frac{P((A_1\cup A_2\cup\cdots)\cap B)}{P(B)}$

$$=\frac{P((A_1\cap B)\cup(A_2\cap B)\cup\cdots)}{P(B)}$$

$$=\frac{P(A_1\cap B)+P(A_2\cap B)+\cdots}{P(B)}$$

$$=\frac{P(A_1\cap B)}{P(B)}+\frac{P(A_2\cap B)}{P(B)}+\cdots=P(A_1|B)+P(A_2|B)+\cdots$$

이므로 확률의 두 번째 공리를 만족한다.

17. 민감도와 특이도를 수식을 이용하여 정의하고, 그 차이를 설명하라.

sol) ①민감도 $= P(\text{양성}\,|\,\text{환자}) = \dfrac{P(\text{양성} \cap \text{환자})}{P(\text{환자})}$

②특이도 $= P(\text{음성}\,|\,\text{정상}) = \dfrac{P(\text{음성} \cap \text{정상})}{P(\text{정상})}$

⇒ 민감도는 환자들 중에서 양성 판정이 나올 조건부 확률이므로 환자를 환자라고 선별하는 검사력이고, 특이도는 정상인 중에서 음성 판정이 나올 조건부 확률이므로 정상인을 정상인으로 선별하는 검사력이다.

18. 다음 표를 이용하여 다음을 계산하라. (단, 결핵의 유병률은 0.0008% 라고 가정한다.)

검사/질병 유무	결핵 있음	결핵 없음
양성	120	7
음성	20	250

sol) 1) 민감도: $\dfrac{120}{120+20} = \dfrac{120}{140} = \dfrac{6}{7} = 0.86$

2) 특이도: $\dfrac{250}{7+250} = \dfrac{250}{257} = 0.97$

3) 양성예측도: $\dfrac{0.86 \times 0.0008}{0.86 \times 0.0008 + (1-0.97) \times (1-0.0008)} = \dfrac{0.000688}{0.000688 + 0.029976} = 0.022$

4) 음성예측도: $\dfrac{0.97 \times (1-0.0008)}{0.97 \times (1-0.0008) + (1-0.86) \times 0.0008} = \dfrac{0.969224}{0.969224 + 0.000112} = 0.9999$

19. 양성예측도와 음성예측도를 수식을 이용하여 정의하고, 그 차이를 설명하라.

sol) 질병 여부에 따라 환자를 D, 정상인을 $\overline{D}$라 하고, 선별검사에 따라 양성을 T, 음성을 $\overline{T}$라 하자.

① 양성예측도: $P(D\,|\,T) = \dfrac{P(T\,|\,D)P(D)}{P(T\,|\,D)P(D) + P(T\,|\,\overline{D})P(\overline{D})}$

② 음성예측도: $P(\overline{D}\,|\,\overline{T}) = \dfrac{P(\overline{T}\,|\,\overline{D})P(\overline{D})}{P(\overline{T}\,|\,\overline{D})P(\overline{D}) + P(\overline{T}\,|\,D)P(D)}$

⇒ 양성예측도는 선별검사 결과가 양성일 때, 실제로 환자일 조건부 확률이고,
음성예측도는 선별검사 결과가 음성일 때, 실제로 정상일 조건부 확률이다.

연습문제 4장

1. 이산확률변수를 정의하고, 이산확률변수에 해당하는 예를 세 가지 이상 들어보라.

sol) 변수가 가질 수 있는 값의 종류가 유한하거나 자연수만큼 존재하는 경우를 이산형이라 하며, 관측값으로 이산형을 갖는 확률변수를 이산확률변수라고 한다.
ex_일별 종합병원을 방문한 환자의 수, 환자의 나이, 투약 횟수

2. 연속확률변수를 정의하고, 연속확률변수에 해당하는 예를 세 가지 이상 들어보라.

sol) 변수가 가질 수 있는 값들의 집합이 실수의 구간이나 실수집합인 확률변수를 연속확률변수라 한다.
ex_환자의 키, 투여량, 혈압

3. 이산확률변수의 확률분포함수를 정의하라.

sol) 이산확률변수가 가질 수 있는 값들의 확률을 함수를 이용하여 나타낸 것으로 확률질량함수라고 하며, $p(x)= P(X=x)$로 표현한다.

4. 연속확률변수의 확률분포함수를 정의하라.

sol) 음의 값을 갖지 않는 함수 $f(x)$와 x-축으로 둘러싸인 면적의 값이 1이고 x-축의 임의의 두 점인 a와 b에서 수직선과 $f(x)$, x-축으로 둘러싸인 영역의 면적이 $P(a \le X \le b)$와 같으면 $f(x)$를 연속확률변수 X의 확률분포함수 또는 확률밀도함수라고 한다.

5. 누적확률분포를 수식을 이용하여 정의하라.

sol) 누적확률분포는 개별 확률 $P(X=x)$을 누적하여 더하면 얻을 수 있다. 즉, $P(X \le x)$을 의미한다.

6. 베르누이 시행을 정의하라.

sol) 어떤 실험에서 배타적인 두 가지 결과가 가능한 시행을 의미한다.

7. 확률분포함수를 이용하여 이항분포를 설명하라.

sol) 이항분포의 확률분포함수는 $p(x)=\begin{cases} P(X=x) = \binom{n}{x} p^x q^{n-x} & , x = 0,1,2 \cdots, n \\ 0 & , \text{그 밖의 경우} \end{cases}$ 이며, 성공의 확률이 p, 실패의 확률이 q인 베르누이 시행을 독립적으로 n번 반복할 때, 성공의 횟수가 x번일 확률을 의미한다.

8. 이항분포를 따르는 확률변수의 예를 들어보라.

sol) 감기에 걸린 환자 수

9. 확률분포함수를 이용하여 포아송 분포를 설명하라.

sol) 포아송의 확률분포함수는 $p(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \cdots$이며, λ는 가정한 시간이나 공간에서의 평균 사건 발생 횟수를 의미한다.

10. 포아송 분포를 따르는 확률변수의 예를 들어보라.

sol) 숙주에서 발견되는 기생충 수

11. 확률분포함수를 이용하여 정규분포를 설명하라.

sol) 정규분포의 확률분포함수는 $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}, -\infty<x<\infty$ 이며, μ와 σ는 각각 중심 경향과 산포를 나타내는 모수이다.

12. 표준정규분포는 왜 중요한지 설명하라.

sol) 자연에서 관측되는 데이터들은 많은 경우에 근사적으로 정규분포를 따르기 때문에 정규분포는 데이터를 근사적으로 모형화하는 데 많이 사용된다. 또한 정규분포를 가정하면 다양한 종류의 확률을 쉽게 계산할 수 있으므로 유용하다.

13. 정규분포를 따르는 확률변수의 예를 들어보라.

sol) 성인 남성의 키

14. 다음의 약물의 수와 약물 중독자의 빈도를 나타낸 표이다. 다음 물음에 답하라.

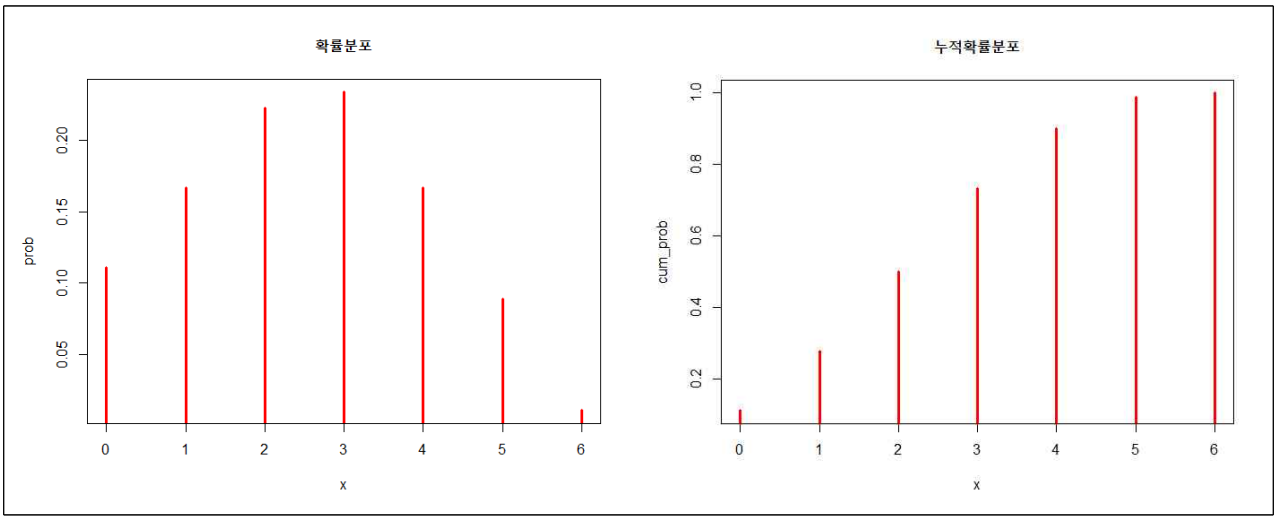
중독물질의 수	빈도
0	100
1	150
2	200
3	210
4	150
5	80
6	10
합계	900

sol)

1) 상대도수와 누적도수표를 만들어라.

중독물질의 수	빈도	상대도수	누적도수
0	100	$\frac{100}{900}$	100
1	150	$\frac{150}{900}$	$100+150=250$
2	200	$\frac{200}{900}$	$100+150+200=450$
3	210	$\frac{210}{900}$	$100+150+200+210=660$
4	150	$\frac{150}{900}$	$100+150+200+210+150=810$
5	80	$\frac{80}{900}$	$100+150+200+210+150+80=890$
6	10	$\frac{10}{900}$	$100+150+200+210+150+80+10=900$
합계	900	1	

2) 확률분포와 누적확률분포의 그래프를 그려라.



3) 랜덤으로 선택한 사람이 5개의 중독 물질을 사용할 확률을 구하라.

$P(X=5)=\frac{80}{900}=0.089$ ($X=$ 중독 물질의 수)

4) 랜덤으로 선택한 사람이 3개 미만의 중독 물질을 사용할 확률을 구하라.

$P(X<3)=\frac{100+150+200}{900}=\frac{450}{900}=0.5$ ($X=$ 중독 물질의 수)

5) 랜덤으로 선택한 사람이 6개 초과의 중독 물질을 사용할 확률을 구하라.

$P(X>6)=0$ ($X=$ 중독 물질의 수)

6) 랜덤으로 선택한 사람이 2개 이상 5개의 이하의 중독 물질을 사용할 확률을 구하라.

$P(2\le X\le 5)=\frac{200+210+150+80}{900}=\frac{640}{900}=0.711$

7) 도수분포를 이용하여 평균, 분산, 표준편차를 구하라.

①평균: $\left(0\times\frac{100}{900}\right)+\left(1\times\frac{150}{900}\right)+\left(2\times\frac{200}{900}\right)+\cdots+\left(6\times\frac{10}{900}\right)=\frac{2240}{900}=2.489$

②분산: $(0-2.489)^2\left(\frac{100}{900}\right)+(1-2.489)^2\left(\frac{150}{900}\right)+(2-2.489)^2\left(\frac{200}{900}\right)+\cdots+(6-2.489)^2\left(\frac{10}{900}\right)=2.25$

③표준편차: $\sqrt{\text{분산}}=\sqrt{2.25}=1.5$

※ 다음 연습문제에서 모집단의 크기 N은 표본의 크기 n에 비해 충분히 크다고 가정하자.

15. 고혈압을 앓고 있는 성인 비율은 20%라고 한다. 랜덤으로 20명의 성인을 추출했을 때

sol) 고혈압을 앓고 있는 환자의 수를 X라 하자.

1) 고혈압 환자가 3명일 확률: $P(X=3)=\binom{20}{3}(0.2)^3(0.8)^{17}=0.2054$

2) 고혈압 환자가 3명 이상일 확률:

$$P(X \geq 3) = 1 - P(X < 3)$$
$$= 1 - \left\{ \binom{20}{0}(0.2)^0(0.8)^{20} + \binom{20}{1}(0.2)^1(0.8)^{19} + \binom{20}{2}(0.2)^2(0.8)^{18} \right\}$$
$$= 1 - 0.2061 = 0.7939$$

3) 고혈압 환자가 3명 미만일 확률:

$$P(X < 3) = \left\{ \binom{20}{0}(0.8)^{20} + \binom{20}{1}(0.2)^1(0.8)^{19} + \binom{20}{2}(0.2)^2(0.8)^{18} \right\} = 0.2061$$

4) 고혈압 환자가 3명 이상, 7명 이하일 확률:

$$P(3 \leq X \leq 7) = \left\{ \binom{20}{3}(0.2)^3(0.8)^{17} + \binom{20}{4}(0.2)^4(0.8)^{16} + \cdots + \binom{20}{4}(0.2)^7(0.8)^{13} \right\}$$
$$= \sum_{x=3}^{7} \binom{20}{x}(0.2)^x(0.8)^{20-x} = 0.7618$$

16. 연습문제 15의 가정이 만족된다고 가정하자. 20명의 표본을 뽑았을 때 기대되는 고혈압 환자의 수를 구하라.

sol) $E(X) = n \times p = 20 \times 0.2 = 4$

17. 연습문제 15의 가정이 만족된다고 가정하자. 랜덤으로 5명의 성인을 선택했다고 가정했을 때, 다음의 확률을 계산하라.

sol)

1) 고혈압 환자가 0명일 확률:

$$P(X = 0) = \binom{5}{0}(0.2)^0(0.8)^5 = 0.3277$$

2) 고혈압 환자가 1명 초과일 확률:

$$P(X > 1) = 1 - P(X \leq 1)$$
$$= 1 - \left\{ \binom{5}{0}(0.2)^0(0.8)^5 + \binom{5}{1}(0.2)^1(0.8)^4 \right\} = 0.2627$$

3) 고혈압 환자가 1명 이상, 3명 이하일 확률:

$$P(1 \leq X \leq 3) = \left\{ \binom{5}{1}(0.2)^1(0.8)^4 + \binom{5}{2}(0.2)^2(0.8)^3 + \binom{5}{3}(0.2)^3(0.8)^2 \right\} = 0.6656$$

4) 고혈압 환자가 2명 이하일 확률:

$$P(X \leq 2) = \left\{ \binom{5}{0}(0.2)^0(0.8)^5 + \binom{5}{1}(0.2)^1(0.8)^4 + \binom{5}{2}(0.2)^2(0.8)^3 \right\} = 0.9421$$

5) 고혈압 환자가 5명일 확률:

$$P(X = 5) = \binom{5}{5}(0.2)^5(0.8)^0 = 0.0003$$

18. $p=0.8, n=3$인 이항분포를 고려하자. $(p+q)^n=1$를 이항전개(binomial expansion)하였을 때, 이항전개된 식의 $i+1$번째 항은 성공이 i번 일어날 확률과 같음을 확인하고, 이를 이용하여 $\sum_{i=0}^{n} p(i)=1$ 임을 증명하라.

sol) $(p+q)^n=\binom{n}{0}p^0q^n+\binom{n}{1}p^1q^{n-1}+\cdots+\binom{n}{i}p^iq^{n-i}+\cdots+\binom{n}{n}p^nq^0=\sum_{i=0}^{n}P(i)=1$

19. 울산의 공장에서 1년 평균 5건의 사고가 발생한다고 가정하자. 다음의 확률을 계산하라. 수를 구하라. ($X=$ 1년 동안 사건 발생 수)

sol)

1) 1년 동안 7건의 사고가 발생할 확률:

$$P(X=7)=\frac{e^{-5}5^7}{7!}=0.1044$$

2) 1년 동안 10건 이상의 사고가 발생할 확률:

$$P(X\geq 10)=1-P(X<10)$$
$$=1-\left\{\frac{e^{-5}5^0}{0!}+\frac{e^{-5}5^1}{1!}+\cdots+\frac{e^{-5}5^9}{9!}\right\}=1-\sum_{x=0}^{9}\left(\frac{e^{-5}5^x}{x!}\right)=0.0318$$

3) 1년 동안 사고가 발생하지 않을 확률:

$$P(X=0)=\frac{e^{-5}5^0}{0!}=0.0067$$

4) 1년 동안 5건 미만의 사고가 발생할 확률:

$$P(X<5)=\left\{\frac{e^{-5}5^0}{0!}+\frac{e^{-5}5^1}{1!}+\cdots+\frac{e^{-5}5^4}{4!}\right\}=\sum_{x=0}^{4}\left(\frac{e^{-5}5^x}{x!}\right)=0.4405$$

20. 확률변수 Z는 표준정규분포를 따른다는 가정하에서 다음을 계산하라.

sol)

1) 확률변수 Z의 확률밀도함수에 대하여 $x=0$과 $x=1.43$, x축, 그리고 표준정규분포의 확률밀도함수로 둘러싸인 영역의 면적을 구하라.
   $P(0\leq Z\leq 1.43)=0.4236$
2) Z가 -2.87보다 크고 2.64보다 작을 확률을 구하라.
   $P(-2.87\leq Z\leq 2.64)=0.9938$
3) $P(Z\geq 0.50)=0.3085$
4) $P(Z\geq -0.45)=0.6736$
5) $P(Z<-2.35)=0.0094$
6) $P(Z<2.35)=0.9906$
7) $P(-1.96\leq Z\leq 1.96)=0.9500$
8) $P(-2.58\leq Z\leq 2.58)=0.9901$
9) $P(-1.65\leq Z\leq 1.65)=0.9011$

21. 고등학교 2학년 여학생의 키는 평균 163, 표준편차 5인 정규분포를 따른다고 가정하자. 다음의 확률을 구하라. ($X=$ 고등학교 2학년 여학생의 키)

sol)

1) 키가 168 이상일 확률:

$$P(X \geq 168) = P\left(\frac{X-163}{5} \geq \frac{168-163}{5}\right) = P(Z \geq 1) = 0.1587$$

2) 키가 158 미만일 확률:

$$P(X < 158) = P\left(\frac{X-163}{5} < \frac{158-163}{5}\right) = P(Z < -1) = 0.1587$$

3) 키가 158보다 크고 168보다 작을 확률:

$$P(158 < X < 168) = P\left(\frac{158-163}{5} < \frac{X-163}{5} < \frac{168-163}{5}\right) = P(-1 < Z < 1) = 0.6827$$

4) 키가 153보다 크고 173보다 작을 확률

$$P(153 < X < 173) = P\left(\frac{153-163}{5} < \frac{X-163}{5} < \frac{173-163}{5}\right) = P(-2 < Z < 2) = 0.9545$$

22. $X \sim N(75, 625)$라고 한다. 이때 다음의 확률을 구하라.

sol)

1) $P(50 \leq X \leq 100) = P\left(\frac{50-75}{\sqrt{625}} \leq \frac{X-75}{\sqrt{625}} \leq \frac{100-75}{\sqrt{625}}\right) = P(-1 \leq Z \leq 1) = 0.6872$

2) $P(X > 90) = P\left(\frac{X-75}{\sqrt{625}} > \frac{90-75}{\sqrt{625}}\right) = P(Z > 0.6) = 0.2743$

3) $P(X < 60) = P\left(\frac{X-75}{\sqrt{625}} < \frac{60-75}{\sqrt{625}}\right) = P(Z < -0.6) = 0.2743$

4) $P(X \geq 85) = P\left(\frac{X-75}{\sqrt{625}} \geq \frac{85-75}{\sqrt{625}}\right) = P(Z \geq 0.4) = 0.3446$

5) $P(30 \leq X \leq 110) = P\left(\frac{30-75}{\sqrt{625}} \leq \frac{X-75}{\sqrt{625}} \leq \frac{110-75}{\sqrt{625}}\right) = P(-1.8 \leq Z \leq 1.4) = 0.8833$

23. 어떠한 이항분포의 평균과 분산은 각각 20, 16이라고 한다. 이때 이항분포의 시행횟수와 성공확률을 구하라.

sol) $np = 20,\ np(1-p) = 16$이므로 $n = 100, p = 0.2$

24. 정규분포를 따르는 확률변수 X의 평균은 100, 표준편차는 15일 때, 다음을 만족하는 k를 구하라.

sol)

1) $P(X \leq k) = 0.0094 \Rightarrow k = 64.7579$

2) $P(X \geq k) = 0.1093 \Rightarrow k = 118.4539$

3) $P(100 \leq X \leq k) = 0.4778 \Rightarrow k = 130.1544$

4) $P(\mu - k \leq X \leq \mu + k) = 0.9660 \Rightarrow k = 31.8011$

25. 다음의 분포표는 확률분포라고 할 수 있는가? 또한 답변에 대한 근거를 마련하라.

sol)

1)

X	$P(X=x)$
0	0.15
1	0.25
2	0.10
3	0.25
4	0.30

확률 값이 모두 음이 아닌 실수이지만,
모든 확률 값의 합이 1이 아니기 때문에
확률분포라고 할 수 없다.

2)

x	$P(X=x)$
0	0.15
1	0.20
2	0.30
3	0.10

확률 값이 모두 음이 아닌 실수이지만,
모든 확률 값의 합이 1이 아니기 때문에
확률분포라고 할 수 없다.

3)

x	$P(X=x)$
0	0.15
1	-0.20
2	0.30
3	0.20
4	0.15

확률 값으로 음수를 갖을 수 없기 때문에
확률분포라고 할 수 없다.

4)

x	$P(X=x)$
-1	0.15
0	0.30
1	0.20
2	0.15
3	0.10
4	0.10

확률 값이 모두 음이 아닌 실수이고,
모든 확률 값의 합이 1이기 때문에
확률분포라고 할 수 있다.

연습문제 5장

1. 표집분포가 무엇인지 설명하라.

sol) 어떤 모집단에서 무작위로 뽑은 표본에서 계산한 통계량이 가질 수 있는 값들의 분포를 통계량의 표집분포라고 한다.

2. 중심극한정리에 대하여 설명하라.

sol) 모집단이 정규분포가 아닌 다른 분포를 따르며 평균이 μ이고, 분산이 σ^2일 때, 크기가 n인 표본으로부터 계산한 $\bar{x}$의 표집분포는 n이 충분히 크면 평균과 분산이 각각 $\mu, \sigma^2/n$인 정규분포이다.

3. 두 표본평균 차이의 분포에 대하여 기술하라.

sol) 두 모집단이 정규분포를 따르며 평균은 각각 μ_1, μ_2이고 분산은 σ_1^2, σ_2^2이라고 하면, 두 모집단으로부터 크기가 n_1, n_2인 두 독립 표본의 평균 차이 $\overline{x_1} - \overline{x_2}$의 표집분포는 평균 $\mu_1 - \mu_2$, 분산 $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$인 정규분포 이다.

4. 20~47세의 혈청콜레스테롤 평균은 $204 mg/dl$이며, 표준편차는 대략 $44 mg/dl$이라고 한다. 이 모집단으로부터 50명의 표본을 추출하였을 때, 평균과 표준오차는 각각 무엇인가?

sol) ① 표본평균의 평균: $204 mg/dl$

② 표준오차: $\frac{44}{\sqrt{50}} = 6.22 mg/dl$

5. 20~29세 여성의 혈청콜레스테롤 평균은 $183 mg/dl$, 표준편차는 $37 mg/dl$이라고 한다. 이 모집단으로부터 60명의 여성을 표본으로 추출하였을 때, 다음의 확률을 구하라.

sol) 표본의 크기가 30이상이므로 중심극한정리에 의해 표본평균 $\overline{X}$의 평균은 183이고, 표준오차는 $37/\sqrt{50} = 4.78$인 정규분포를 따른다.

1) 표본평균이 170보다 크고 195보다 작을 확률:

$$P(170 \le \bar{x} \le 195) = P\left(\frac{170-183}{4.78} \le z \le \frac{195-183}{4.78}\right)$$
$$= P(-2.72 \le z \le 2.51) = 0.994 - 0.0033 = 0.9907$$

2) 표본평균이 175 미만일 확률:

$$P(\bar{x} < 175) = P\left(z < \frac{175-183}{4.78}\right)$$
$$= P(z < -1.67) = 0.0475$$

3) 표본평균이 190 이상일 확률:

$$P(\bar{x} \ge 190) = P\left(z \ge \frac{190-183}{4.78}\right)$$
$$= P(z \ge 1.46) = 1 - 0.9279 = 0.0721$$

6. 한 도시에서 진행된 연구에 따르면 60세 이상의 남성과 여성의 칼슘 수치의 평균과 표준편차는 다음과 같다고 한다.

	평균	표준 편차
남자	797	482
여자	660	414

sol) 표본의 크기가 모두 30보다 크므로 각 표본평균은 중심극한정리에 의하여 정규분포를 따른다고 가정할 수 있다. 따라서 표본평균의 차이의 표집분포는 평균 $\mu_{\bar{x}_1-\bar{x}_2}=797-660=137$이고, 분산 $\sigma^2_{\bar{x}_1-\bar{x}_2}=\frac{482^2}{40}+\frac{414^2}{35}=10705.13$인 정규분포이다.

1) 이 모집단으로부터 40명의 남성과 35명의 여성을 추출하였다. 이때 남성과 여성의 표본평균 차이가 100 이상일 확률은 얼마인가?

$$P(\overline{x_1}-\overline{x_2}\geq 100)=P\left(z\geq\frac{100-137}{\sqrt{10705.13}}\right)$$
$$=P(z\geq -0.36)=0.6406$$

2) 1)의 예에서 표본평균의 차이가 50 이상일 확률은 얼마인가?

$$P(\overline{x_1}-\overline{x_2}\geq 50)=P\left(z\geq\frac{50-137}{\sqrt{10705.13}}\right)$$
$$=P(z\geq -0.84)=0.7995$$

7. 20~74세의 한국 성인의 45%는 비만이라고 가정하자. 이 모집단으로부터 성인 125명을 임의 추출하였다. 이때 다음의 확률은 구하라.

sol) 125명의 표본에서 $np=125\times 0.45=56.25$, $n(1-p)=150\times 0.55=68.75$로 모두 5보다 크다. 따라서 표본비율 $\hat{p}$은 평균이 0.45이고, 분산은 $(0.45)(0.55)/125=0.00198$인 정규분포를 따른다.

1) 비만인 사람이 70% 이상일 확률:

$$P(\hat{p}\geq 0.7)=P\left(z\geq\frac{0.7-0.45}{\sqrt{0.00198}}\right)$$
$$=P(z\geq 5.62)=0.000000001$$

2) 비만인 사람이 70% 초과일 확률:

$$P(\hat{p}> 0.7)=P\left(z>\frac{0.7-0.45}{\sqrt{0.00198}}\right)$$
$$=P(z>5.62)=0.000000001$$

3) 비만인 사람이 90% 이상일 확률:

$$P(\hat{p}\geq 0.9)=P\left(z\geq\frac{0.9-0.45}{\sqrt{0.00198}}\right)$$
$$=P(z\geq 10.11)=0.000$$

4) 비만인 사람이 10% 미만일 확률:

$$P(\hat{p}< 0.1)=P\left(z<\frac{0.1-0.45}{\sqrt{0.00198}}\right)$$
$$=P(z< -7.87)=0.000$$

8. 성인의 35%는 한 가지 이상의 만성 질환을 겪고 있다고 한다. 이 모집단으로부터 200명의 성인을 추출하였을 때, 만성 질환이 있는 사람들이 80% 이상일 확률을 구하라.

sol) 200명의 표본에서 $np = 200 \times 0.35 = 70,\ n(1-p) = 200 \times 0.65 = 130$로 모두 5보다 크다. 따라서 표본비율 $\hat{p}$은 평균이 0.35이고, 분산은 $(0.35)(0.65)/200 = 0.00114$인 정규분포를 따른다.

$$\Rightarrow\ P(\hat{p} \geq 0.8) = P\left(z \geq \frac{0.8 - 0.35}{\sqrt{0.00114}}\right) = P(z \geq 13.33) = 0.000$$

9. A회사 노동자의 21%, B회사 노동자의 13%가 매년 1번 이상 병원을 찾는다고 한다. A회사 노동자 100명, B회사 노동자 120명의 표본을 추출하였을 때, 다음의 확률을 구하라.

sol) 표본이 충분히 크므로 표본비율 $\hat{p}_A - \hat{p}_B$은 근사적으로 정규분포를 따른다고 볼 수 있다. 두 회사 노동자의 병원 방문 비율의 차이 평균은 $\mu_{\hat{p}_A - \hat{p}_B} = 0.21 - 0.13 = 0.08$이고, 분산은

$\sigma^2_{\hat{p}_A - \hat{p}_B} = \dfrac{0.21(1-0.21)}{100} + \dfrac{0.13(1-0.13)}{120} = 0.0026$이다.

1) 표본 비율의 차이인 $\widehat{P_A} - \widehat{P_B}$가 0.04와 0.2 사이가 될 확률:

$$P(0.04 < \hat{p}_A - \hat{p}_B < 0.2) = P\left(\frac{0.04 - 0.08}{\sqrt{0.0026}} < z < \frac{0.2 - 0.08}{\sqrt{0.0026}}\right)$$
$$= P(-0.78 < z < 2.35) = 0.77$$

2) $\widehat{P_A} - \widehat{P_B}$가 0.3 이상일 확률:

$$P(\hat{p}_A - \hat{p}_B \geq 0.3) = P\left(z \geq \frac{0.3 - 0.08}{\sqrt{0.0026}}\right)$$
$$= P(z \geq 4.31) = 0.000$$

3) $\widehat{P_A} - \widehat{P_B}$가 0.1 이하일 확률:

$$P(\hat{p}_A - \hat{p}_B \leq 0.1) = P\left(z \leq \frac{0.1 - 0.08}{\sqrt{0.0026}}\right)$$
$$= P(z \leq 0.39) = 0.65$$

4) $\widehat{P_A} - \widehat{P_B}$가 0.5일 확률:

$$P(\hat{p}_A - \hat{p}_B = 0.5) = P\left(z = \frac{0.5 - 0.08}{\sqrt{0.0026}}\right)$$
$$= P(z = 8.24) = 0$$

5) $\widehat{P_A} - \widehat{P_B}$가 0 이상일 확률:

$$P(\hat{p}_A - \hat{p}_B \geq 0) = P\left(z \geq \frac{0.3 - 0.08}{\sqrt{0.0026}}\right)$$
$$= P(z \geq -1.57) = 0.94$$

10. 20~39세 남성의 체내 철분 보유량의 평균은 $18mg$이고, 표준 편차는 $10mg$라고 한다. 이 모집단으로부터 120명의 남성을 뽑았을 때 체내 철분 보유량의 표본평균이 $19mg$ 이상일 확률을 구하라.

sol) 표본의 크기가 30이상이므로 중심극한정리에 의해 표본평균 $\overline{X}$의 평균은 18이고, 표준오차는 $10/\sqrt{120} = 0.913$인 정규분포를 따른다.

$$\Rightarrow P(\overline{x} \geq 19) = P\left(z \geq \frac{19-18}{0.913}\right) = P(z \geq 1.095) = 0.1367$$

11. 65세 이상 남성의 암 발생 비율은 23%라고 한다. 250명의 남성 중에서 암이 발생한 남성의 비율이 20% 미만일 확률을 구하라.

sol) 250명의 표본에서 $np = 250 \times 0.23 = 57.5,\ n(1-p) = 250 \times 0.77 = 192.5$로 모두 5보다 크다. 따라서 표본비율 $\hat{p}$은 평균이 0.23이고, 분산은 $(0.23)(0.77)/250 = 0.00071$인 정규분포를 따른다.

$$\Rightarrow P(\hat{p} < 0.2) = P\left(z < \frac{0.2-0.23}{\sqrt{0.00071}}\right) = P(z < -1.126) = 0.1301$$

12. 두 모집단의 평균은 같고 모분산은 각각 $\sigma_1^2 = 100,\ \sigma_2^2 = 80$이라고 한다. 두 모집단에서 각각 $40, 60$개의 표본을 뽑았을 때, $\overline{x_1} - \overline{x_2}$가 8보다 크거나 같을 확률을 구하라.

sol) 표본의 크기가 모두 30보다 크므로 각 표본평균은 중심극한정리에 의하여 정규분포를 따른다고 가정할 수 있고, 표본평균의 차이의 표집분포는 평균 0, 분산 $\sigma^2_{\overline{x}_1 - \overline{x}_2} = 100/40 + 80/60 = 3.83$인 정규분포이다.

$$\Rightarrow P(\overline{x_1} - \overline{x_2} \geq 8) = P\left(z \geq \frac{0-8}{\sqrt{3.83}}\right) = P(z \geq -4.088) = 0.9999782$$

연습문제 6장

1. 통계적 추론 과정을 설명하라.
sol) 표본을 활용하여 모집단의 모수를 유추하는 과정으로 표본으로부터 모집단에 관한 정보를 얻고 도출하는 과정이다.

2. 통계적 추론에서 추정이 왜 중요한지 그 이유를 설명하라.
sol) 추정량은 모수를 추정하기 위해 사용되므로, 모수를 제대로 추정하기 위해서는 추정이 중요하다.

3. 점추정에 대하여 설명하라.
sol) 모집단의 모수를 하나의 값으로 추정하는 것이다.

4. 불편성에 대하여 설명하라.
sol) 모수 θ의 추정량 T의 기댓값 $E(T)$가 θ와 일치하는 것이다.

5. 다음을 정의하라.
sol) 1) 신뢰도: 다양한 표본에서 추정한 구간에 모수가 실제로 존재할 확률
2) 신뢰계수: 어떤 분포에서 특장 값이 나타날 확률이 특정한 신뢰도보다 작아지게 되는 값
3) 정밀도: 신뢰성계수에 평균의 표준오차를 곱한 값
4) 표준오차: 표본 분포의 표준 편차
5) 추정량: 모수를 추정하기 위해 표본에서 얻은 통계량
6) 오차범위: 정밀도와 동일

6. 신뢰구간을 구하는 일반적인 공식을 설명하라.
sol) 추정량 $\pm$(신뢰성계수)$\times$(표준오차)

7. 추정에서 중심극한정리는 어떻게 활용되는지 설명하라.
sol) 정규분포를 따른다고 가정할 수 있는 경우, 추정 및 검정 과정에 확률분포를 적용할 수 있다.

8. $t-$분포에 대하여 설명하라.
sol) 평균 0을 중심으로 좌우대칭 형태를 가지는 분포로 분산은 1보다 큰 값을 가지며, 모수는 자유도이며 자유도가 커질수록 분산이 1에 가까워진다. 정규분포에 비하여 중심이 덜 뾰족하고 꼬리는 더 두터운 형태를 보이며 자우도가 클수록 정규분포와 비슷해진다.

9. 모평균의 신뢰구간을 추정하기 위하여 $t-$분포를 이용하려고 한다. 이때 필요한 가정을 설명하라.
sol) 모집단이 모분산이 알려지지 않은 정규분포를 따라야 한다.

10. 유한 모집단 보정에 대하여 설명하라. 유한 모집단을 보정을 할 필요가 없는 경우는 언제인지 설명하라.
sol) 유한모집단으로부터 얻은 통계량에 일정한 계수를 곱해주어 편의를 보정하는 과정을 유한 모집단 보정이라 하며, 보정이 불필요한 경우는 모집단이 무한한 경우, 표본의 크기가 작은 경우이다.

11. 두 모평균 차이의 신뢰구간을 추정하기 위하여, $t-$분포를 이용하려고 한다. 이때 필요한 가정을 설명하라.

sol) 모집단이 모분산이 알려지지 않은 정규분포를 따라야 한다.

12. 15명의 남성의 체중은 다음과 같다고 한다.

75, 65, 60, 74, 75, 70, 68, 71, 83, 68, 65, 67, 73, 72, 65

남성의 체중은 정규분포를 따른다는 가정하에서 모평균의 95% 신뢰구간을 구하라.

sol) $\overline{X}=70.07$, $Var(\overline{X})=2.1$, $t_{0.975,14}=2.1448$

$$\Rightarrow \left(\overline{X}-t_{0.975,\,14}\sqrt{Var(\overline{X})},\ \overline{X}+t_{0.975,\,14}\sqrt{Var(\overline{X})}\right)$$
$$=(70.07-2.1448\times\sqrt{2.1}\,,\ 70.07+2.1448\times\sqrt{2.1}\,)$$
$$=(66.96,\ 73.17)$$

13. 15명의 천식환자 중에서 30%가 진드기 알레르기 검사에 양성 반응을 보였다. 모비율의 95% 신뢰구간을 구하라.

sol) $\hat{p}=0.3$, $n=150$, $z_{0.975}=1.96$

$$\Rightarrow \left(\hat{p}-z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\ \hat{p}+z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$
$$=\left(0.3-1.96\times\sqrt{\frac{0.3\times0.7}{150}}\,,\ 0.3+1.96\times\sqrt{\frac{0.3\times0.7}{150}}\right)$$
$$=(0.23,\ 0.37)$$

14. 공장의 안전성을 조사하기 위하여 80개의 공장을 조사한 결과, 25개의 공장이 '위험' 등급을 받았다고 한다. '위험' 등급에 해당하는 공장의 모비율의 95% 신뢰구간을 구하라.

sol) $\hat{p}=0.3125$, $n=80$, $z_{0.975}=1.96$

$$\Rightarrow \left(\hat{p}-z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\ \hat{p}+z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$=\left(0.3125-1.96\times\sqrt{\frac{0.3125\times0.6875}{80}}\,,\ 0.3125+1.96\times\sqrt{\frac{0.3125\times0.6875}{80}}\right)$$
$$=(0.21,\ 0.41)$$

15. 14번 문제에서 95% 신뢰구간의 길이가 0.05 이하가 되기 위해 필요한 표본의 최소 크기를 추정하라.

sol) $\hat{p}=0.3125$, $n=80$, $z_{0.975}=1.96$

$$\Rightarrow\ 0.2\times1.96\times\sqrt{\frac{0.3125\times0.6875}{80}}<0.05$$
$$\Rightarrow\ n>1320.55 \qquad \therefore 1321\text{명}$$

16. 500명의 성인을 대상으로 가장 최근에 병원을 방문한 이유를 조사하였다. 210명의 남성 중에서 42명이 건강검진을 목적으로 방문했다고 답하였다. 또한 290명의 여성 중에서 140명이 건강검진을 목적으로 방문했다고 답하였다. 두 모비율의 차에 대한 95% 신뢰구간을 구하라.

sol) $\widehat{p_{남}} = \frac{42}{210}$, $\widehat{p_{여}} = \frac{140}{290}$, $z_{0.975} = 1.96$

$\Rightarrow p_{여} - p_{남}$의 신뢰구간

$$= \left(\left(\frac{140}{290} - \frac{42}{210}\right) - 1.96 \times \sqrt{\frac{\frac{42}{210}\left(1-\frac{42}{210}\right)}{210} + \frac{\frac{140}{290}\left(1-\frac{140}{290}\right)}{290}} \right.$$
$$\left. , \left(\frac{140}{290} - \frac{42}{210}\right) + 1.96 \times \sqrt{\frac{\frac{42}{210}\left(1-\frac{42}{210}\right)}{210} + \frac{\frac{140}{290}\left(1-\frac{140}{290}\right)}{290}} \right)$$
$$= (0.20, 0.36)$$

17. 다음은 종양의 크기를 조사한 결과이다.

종양의 종류	n	$\overline{x}$	s
A	20	3.75cm	1.90cm
B	15	2.70cm	1.75cm

두 모평균의 차이에 대해 95% 신뢰구간을 구하라.

sol) ① 종양 A, B의 모분산이 같은 경우,

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{19 \times 1.9^2 + 14 \times 1.75^2}{20+15-2} = 3.78,\ t_{0.975,\,33} = 2.0345$$

$\Rightarrow \mu_A - \mu_B$의 신뢰구간

$$= \left((3.75-2.70) - 2.0345 \times \sqrt{\frac{3.78}{20} + \frac{3.78}{15}}, (3.75-2.70) + 2.0345 \times \sqrt{\frac{3.78}{20} + \frac{3.78}{15}} \right)$$
$$= (-0.3, 2.4)$$

① 종양 A, B의 모분산이 다른 경우,

$$df^* = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{1.90^2}{20} + \frac{1.75^2}{15}\right)^2}{\frac{1}{20-1} \times \left(\frac{1.90^2}{20}\right)^2 + \frac{1}{15-1} \times \left(\frac{1.75^2}{15}\right)^2} = 31.53511,$$

$t_{0.975,\,31.53511} = 2.038112$

$\Rightarrow \mu_A - \mu_B$의 신뢰구간

$$= \left((3.75-2.07) - 2.038112 \times \sqrt{\frac{1.90^2}{20} + \frac{1.75^2}{15}}, (3.75-2.07) + 2.038112 \times \sqrt{\frac{1.90^2}{20} + \frac{1.75^2}{15}} \right)$$

18. 알레르기 환자 200명 중 180명은 신약을 먹은 후 알레르기 증상이 호전되었다고 한다. 모비율의 90% 신뢰구간을 구하라.

sol) $\hat{p} = 0.9$, $n = 200$, $z_{0.95} = 1.645$

$$\Rightarrow \left(0.9 - 1.645 \times \sqrt{\frac{0.9(1-0.9)}{200}}, 0.9 + 1.645 \times \sqrt{\frac{0.9(1-0.9)}{200}} \right)$$
$$= (0.87, 0.93)$$

19. 다리에 울형성궤양이 있는 환자를 60명씩 랜덤으로 뽑아 두 집단에 서로 다른 치료약을 처방하였다. 두 집단의 표본평균과 표본표준편차는 다음과 같다.

치료약	$\overline{x}$	s
A	85	25
B	115	30

두 모평균의 차이에 대한 95% 신뢰구간을 구하라.

sol) $\mu_B - \mu_A$의 신뢰구간

$$= \left((115-85) - 1.96 \times \sqrt{\frac{25^2}{60} + \frac{30^2}{60}}, (115-85) + 1.96 \times \sqrt{\frac{25^2}{60} + \frac{30^2}{60}}\right)$$

$$= (20.12, 39.88)$$

20. 초등학교 1학년 학생 10명의 몸무게는 다음과 같다고 한다.

20.5, 24.8, 21.3, 22.7, 18.2, 31.6, 25.4, 21.9, 19.7, 22.2

모집단은 정규분포를 따른다고 알려져 있을 때, 모평균의 95% 신뢰구간을 구하라.

sol) $\overline{X} = 22.83$, $Var(\overline{X}) = 14.19$, $z_{0.975} = 1.96$

$$\Rightarrow \left(\overline{X} - z_{0.975}\sqrt{Var(\overline{X})},\ \overline{X} + z_{0.975}\sqrt{Var(\overline{X})}\right)$$

$$= \left(22.83 - 1.96 \times \sqrt{\frac{14.19}{10}},\ 22.83 + 1.96 \times \sqrt{\frac{14.19}{10}}\right)$$

$$= (20.5,\ 20.16)$$

21. 중학교 1학년 학생들을 대상으로 2개의 독립 표본을 뽑아 구강 내 산도(pH)를 측정하였다. A 집단은 충치가 없는 학생들의 집합이고, B 집단은 충치가 많은 학생들의 집합이다. 측정 결과는 다음과 같다.

A: 7.04, 7.12, 7.51, 7.88, 7.26, 7.85, 7.14, 7.96, 7.57, 7.92, 7.37, 7.66, 7.62, 7.65
B: 7.16, 7.05, 7.09, 7.31, 7.14, 7.25, 7.32, 7.47, 7.10, 7.35, 7.19

등분산을 가정했을 때, 두 모평균의 차에 대한 90% 신뢰구간을 구하라.

sol) $\overline{X_A} = 7.52$, $\overline{X_B} = 7.26$, $s_A^2 = 0.096$, $s_B^2 = 0.038$, $t_{0.95,\,25} = 1.71$

$$\Rightarrow S_p^2 = \frac{14 \times 0.096 + 11 \times 0.038}{15 + 12 - 2} = 0.07$$

$\Rightarrow \mu_A - \mu_B$의 신뢰구간

$$= \left((7.52-7.26) - 1.71 \times \sqrt{\frac{0.07}{15} + \frac{0.07}{12}}, (7.52-7.26) + 1.71 \times \sqrt{\frac{0.07}{15} + \frac{0.07}{12}}\right)$$

$$= (0.085, 0.435)$$

22. 불면증을 겪는 12명의 환자에게 약물 A, 그리고 또 다른 불면증 환자 16명에게는 약물 B를 처방하였다. 약물을 처방한 다음 날 수면시간은 다음과 같다.

A: 4.0, 5.2, 4.2, 5.9, 16.8, 3.5, 3.0, 6.4, 6.8, 3.6, 6.9, 5.7

B: 5.0, 11.2, 11.6, 3.5, 5.3, 3.5, 6.2, 6.6, 7.1, 6.4, 4.5, 5.1, 3.2, 4.7, 4.5, 3.0

등분산을 가정했을 때, 두 모평균의 차에 대한 95 신뢰구간을 구하라.

sol) $\overline{X_A}=6,\ \overline{X_B}=5.7125,\ s_A^2=13.39, s_B^2=6.45, t_{0.975,\,26}=2.0555$

$\Rightarrow\ S_p^2=\dfrac{11\times13.39+17\times6.45}{12+16-2}=9.38$

$\Rightarrow\ \mu_A-\mu_B$의 신뢰구간

$=\left((6-5.7125)-2.0555\times\sqrt{\dfrac{9.38}{12}+\dfrac{9.38}{16}}\,,(6-5.7125)+2.0555\times\sqrt{\dfrac{9.38}{12}+\dfrac{9.38}{16}}\right)$

$=(-2.12, 2.69)$

23. 신뢰구간을 구할 때, 일반적으로 높은 신뢰계수를 선호한다. 만약 신뢰계수를 100%로 설정하면 신뢰구간의 폭은 어떻게 될지 설명하라.

sol) 신뢰계수가 100%로 설정되면 모든 실수 범위가 신뢰구간에 포함된다.

※ 통계소프트웨어 R에 내장되어 있는 chickwts 데이터를 이용하여 다음 물음에 답하라. R 함수 data(chickwts)를 실행하면 chickwts 데이터를 이용할 수 있다.

24. weight 변수의 95% 신뢰구간을 구하라.

sol)

```
> t.weight = t.test(chickwts$weight)
> t.weight$conf.int
[1] 242.8301 279.7896
attr(,"conf.level")
[1] 0.95
```

weight의 95% 신뢰구간은 $(242.8301, 279.7896)$이다.

25. 모집단은 정규분포를 따른다고 가정했을 때, feed변수가 horsebean, linseed인 두 그룹의 몸무게 평균의 차이에 대한 95% 신뢰 구간을 구하라.

sol)

```
> hb.weight = subset(chickwts, feed == 'horsebean')$weight
> ls.weight = subset(chickwts, feed == 'linseed')$weight
> ttest = t.test(hb.weight, ls.weight, var.equal = T)
> ttest$conf.int
[1] -100.17618  -16.92382
attr(,"conf.level")
[1] 0.95
```

두 그룹의 몸무게 평균의 차이에 대한 95% 신뢰구간은 $(-100.17618, -16.92382)$이다.

연습문제 7장

1. 가설검정은 무엇인가? 그리고 가설검정은 왜 필요한지 설명하라.

sol) 가설검정이란 수진한 데이터를 활용하여 관심가설이 적절한지 의사결정을 하는 과정이며, 의사결정을 돕기 위하여 수행한다.

2. 가설검정에 사용되는 두 가지 가설에 대하여 설명하라.

sol)

연구가설	연구의 동기가 되는 추측 혹은 가정
통계가설	수집한 데이터를 통계적 방법을 이용하여 검증할 수 있는 가설

(또는)

영가설	보편적으로 모집단에서 만족될 것으로 생각되는 성질에 대한 진술
대립가설	영가설가 정반대의 의미를 갖는 가설, 많은 경우 과학적으로 의미가 있는 가설

3. 가설 검정의 단계에 대하여 서술하라.

sol) (1) 데이터 (수집 방법, 변수 의미 등)
(2) 가정 (정규성, 등분산성, 독립성 등)
(3) 가설 (영가설, 대립가설 설정)
(4) 검정통계량 설정
(5) 검정통계량 분포
(6) 결정 규칙 (기각역/비기각역, 유의수준)
(7) 검정통계량 계산
(8) 통계적 결정 (영가설 기각 여부)
(9) 결론 (영가설/대립가설 선택)
(10) p-value

4. 1종 오류와 2종 오류를 정의하라.

sol)

1종 오류	영가설이 참일 때 영가설을 기각하여 발생하는 오류
2종 오류	대립가설이 참임에도 영가설을 기각하지 못하여 발생하는 오류

5. $p-$값의 의미를 설명하라.

sol) 영가설이 사실이라는 가정 하에서 계산되는 값으로, 검정통계량이 관측된 검정통계량보다 더 강한 영가설을 기각하라 증거를 보여줄 확률
(또는) 영가설이 사실일 때 표본이 영가설에 얼마나 부합하는지 보여주는 값

6. 정규분포를 따르는 40대 남녀 집단에서 각각 18, 15명을 무작위로 추출하여 혈액수치의 차이가 있는지를 보고자 한다. 다음의 자료를 분석하라.

남: 0.9, 2.2, 1.6, 2.8, 4.2, 3.7, 2.6, 2.9, 3.3, 1.2, 3.2, 2.7, 3.8, 4.5, 4.0, 2.2, 0.8, 1.2
여: 1.4, 2.7, 2.1, 1.8, 3.0, 3.1, 1.6, 1.7, 2.6, 2.5, 2.3, 1.4, 2.6, 3.5, 2.7

sol) ⅰ) 등분산성 검정: $H_0 : \sigma_1^2 = \sigma_2^2$, $H_A : \sigma_1^2 \neq \sigma_2^2$

-검정통계량: $F = \frac{s_1^2}{s_2^2} = \frac{1.358}{0.418} = 3.248$ ⇒영가설 하에서 $F_{17,14}$를 따름

-기각역: $F \geq F_{0.975,17,14} = 2.900$ 또는 $F \leq F_{0.025,17,14} = 0.363$

-통계적 결정: 검정통계량이 기각역에 포함되므로 영가설을 기각한다.
즉, 유의수준 $\alpha = 0.05$에서 분산이 같지 않다는 유의한 근거가 있다.

ⅱ) 평균 검정: $H_0 : \mu_1 = \mu_2$, $H_A : \mu_1 \neq \mu_2$

① 등분산을 가정한 경우

-공통분산: $S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(18-1)\times 1.358 + (15-1)\times 0.418}{18+15-2} = 0.933$

-검정통계량: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1-\mu_2)_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(2.656-2.333)-0}{\sqrt{0.933\left(\frac{1}{18}+\frac{1}{15}\right)}} = 0.954$ ⇒영가설 하에서 t_{31} 따름

-기각역: $t \geq t_{0.975,31} = 2.040$ 또는 $t \leq t_{0.025,31} = -2.040$

-통계적 결정: 검정통계량이 비기각역에 포함되므로 영가설을 기각할 수 없다.
즉, 유의수준 $\alpha = 0.05$에서 각 성별의 혈액 수치에 차이가 있다는 유의한 증거가 없다.

② 이분산을 가정한 경우

-검정통계량: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1-\mu_2)_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{(2.656-2.333)-0}{\sqrt{\frac{1.358}{18}+\frac{0.418}{15}}} = 1.002$

⇒영가설 하에서 근사적으로 t분포를 따름

-기각역: $t \geq t'_{0.975} = \frac{\left(\frac{s_1^2}{n_1}\right)t_{0.975,17} + \left(\frac{s_2^2}{n_2}\right)t_{0.975,14}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \frac{\left(\frac{1.358}{18}\right)\times 2.110 + \left(\frac{0.418}{15}\right)\times 2.145}{\frac{1.358}{18}+\frac{0.418}{15}} = 2.119$ 또는

$t \leq t'_{0.025} = \frac{\left(\frac{s_1^2}{n_1}\right)t_{0.025,17} + \left(\frac{s_2^2}{n_2}\right)t_{0.025,14}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \frac{\left(\frac{1.358}{18}\right)\times(-2.110) + \left(\frac{0.418}{15}\right)\times(-2.145)}{\frac{1.358}{18}+\frac{0.418}{15}} = -2.119$

-통계적 결정: 검정통계량이 비기각역에 포함되므로 영가설을 기각할 수 없다.
즉, 유의수준 $\alpha = 0.05$에서 각 성별의 혈액 수치에 차이가 있다는 유의한 증거가 없다.

7. 흡연자와 비흡연자에 대하여 폐기종 발전 이전의 폐파괴지수를 측정하였다(높을수록 손상 정도가 큰 것을 의미함). 두 집단은 정규분포를 따르며, 모분산은 같다고 한다. 두 모집단의 파괴지수 모평균이 같은지 검정하라.

흡 연 자: 16.6, 13.9, 11.3, 26.5, 17.4, 15.3, 15.8, 12.3, 18.6, 11.0, 26.5, 17.3, 22.9, 15.0, 21.0, 18.5

비흡연자: 18.1, 6.0, 4.2, 11.4, 6.7, 19.5, 7.3, 13.0, 18.7

sol) $H_0 : \mu_1 = \mu_2$, $H_A : \mu_1 \neq \mu_2$

-공통분산: $S_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \dfrac{(16-1)\times 22.74 + (9-1)\times 35.71}{16+9-2} = 27.25$

-검정통계량: $t = \dfrac{(\overline{x}_1 - \overline{x}_2) - (\mu_1-\mu_2)_0}{\sqrt{s_p^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \dfrac{(17.49-11.66)-0}{\sqrt{27.25\left(\dfrac{1}{16}+\dfrac{1}{9}\right)}} = 2.684$ ⇒영가설 하에서 t_{23} 따름

-기각역: $t \geq t_{0.975,\,23} = 2.069$ 또는 $t \leq t_{0.025,\,23} = -2.069$

-통계적 결정: 검정통계량이 기각역에 포함되므로 영가설을 기각할 수 있다.
즉, 유의수준 $\alpha = 0.05$에서 두 모집단의 파괴지수 모평균이 다르다는 유의한 증거가 있다.

8. 동전을 1,000번 던졌을 때 앞면이 560번, 뒷면이 440번 나왔다고 한다. 이 결과를 이용하여, 앞면이 나올 확률이 1/2이라고 할 수 있는지 가설검정 하라.

sol) $H_0 : p = 1/2$, $H_A : p \neq 1/2$

-$n = 1000$, $\hat{p} = \dfrac{560}{1000}$ ⇒ $n\hat{p} > 5, n(1-\hat{p}) > 5$이므로 중심극한정리 적용

-검정통계량: $z = \dfrac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \dfrac{0.56-0.5}{\sqrt{\dfrac{0.5\times 0.5}{1000}}} = 3.795$ ⇒영가설 하에서 표준정규분포 따름

-기각역: $z \geq z_{0.975} = 1.960$ 또는 $z \leq z_{0.025} = -1.960$

-통계적 결정: 검정통계량이 기각역에 포함되므로 영가설을 기각할 수 있다.
즉, 유의수준 $\alpha = 0.05$에서 앞면이 나올 확률이 1/2가 아니라는 유의한 증거가 있다.

9. 8번의 문제에서 뒷면이 500번 나왔다면, 앞면이 나올 확률이 1/2이라고 할 수 있는지 가설검정 하라.

sol) $\hat{p} = 0.5$이므로 검정통계량 $z = 0$, 유의수준에 관계없이 영가설을 기각할 수 없다.
즉, 앞면이 나올 확률이 1/2이 아니라는 유의한 근거가 없다.

10. 평균이 μ이고 표준편차가 5인 정규모집단에서 크기 $n=25$인 표본을 추출하여 다음의 가설을 검정하려고 한다.

$$H_0 : \mu \le 50,\ H_A : \mu > 50$$

오른쪽 검정이므로 $\overline{X}$이 어느 정도보다 클 때 영가설 H_0를 기각하며, 따라서 기각역은 $\overline{X} > c$의 형태로 표현할 수 있다. 유의수준 0.05에서 c를 구하라.

sol) $H_0 : \mu \le 50,\ H_A : \mu > 50$

-검정통계량: $z = \dfrac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \dfrac{\overline{X}-50}{\sqrt{\frac{5^2}{25}}} = \overline{X}-50$ ⇒영가설 하에서 표준정규분포 따름

-기각역: $z > z_{0.95} = 1.645 \Leftrightarrow \overline{X} > 50 + 1.645 = 51.645 \quad \therefore c = 51.645$

11. 10번 문제에서 실제 평균은 $\mu = 55$라고 한다. 유의수준 0.01에서 10번과 동일한 형태의 가설을 검정하려고 한다. 검정력이 0.9가 되기 위한 최소 표본의 크기를 구하라.

sol) $P\left(z = \dfrac{\overline{X}-50}{\sqrt{\frac{\sigma^2}{n}}} > z_{1-0.01} = 2.326 \middle| \mu = 55\right) \ge 0.9$

$$\Rightarrow P\left(\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{X}-55}{\sqrt{\frac{5^2}{n}}} > 2.326 - \frac{5}{\sqrt{\frac{5^2}{n}}} = 2.326 - \sqrt{n} \middle| \mu = 55\right) \ge 0.9$$

$$\Rightarrow 2.326 - \sqrt{n} \le z_{1-0.9} = -1.282$$

$$\Rightarrow n \ge (2.326 + 1.282)^2 = 13.02 \qquad \therefore \text{최소 } 14\text{개가 필요하다.}$$

12. 11명의 환자에게 새로운 혈압 강하제를 처방한 이후, 처방 전후의 혈압을 비교하였다. 혈압 강하제가 효과가 있다고 볼 수 있는지 검정하라.

환자번호	복용 전 혈압	복용 후 혈압
1	90	77
2	42	53
3	64	57
4	88	70
5	93	74
6	86	92
7	71	58
8	55	54
9	64	71
10	65	62
11	91	70

sol) $H_0 : d = \mu_1 - \mu_2 \le 0,\ H_A : d = \mu_1 - \mu_2 > 0$

-검정통계량: $t = \dfrac{\overline{d}-d_0}{s_d/\sqrt{n}} = \dfrac{6.455-0}{11.27/\sqrt{11}} = 1.899$ ⇒영가설 하에서 t_{10} 따름

-기각역: $t > t_{0.95,\,10} = 1.812$

-통계적 결정: 검정통계량이 기각역에 포함되므로 영가설을 기각할 수 있다.
즉, 유의수준 $\alpha = 0.05$에서 복용 후 혈압이 복용 전 혈압보다 낮다는 유의한 증거가 있다.

13. 비타민 C가 감기 치료에 효과가 있는지 검정하기 위하여 179명의 감기 환자에게 비타민 C를 처방한 뒤에 다음의 결과를 얻었다. 비타민이 감기 치료에 효과가 있다고 할 수 있는지 검정하라.

	변화 없음	감기가 완화됨
위약	6	84
비타민 C	17	72

sol) $H_0 : p_1 \geq p_2,\ H_A : p_1 < p_2$

-모비율의 합병추정값: $\bar{p} = \frac{84+72}{179} = \frac{156}{179}$ 이므로 중심극한정리 적용

-검정통계량: $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{\left(\frac{84}{90} - \frac{72}{89}\right) - 0}{\sqrt{\frac{156}{179}\left(1 - \frac{156}{179}\right)\left(\frac{1}{90} + \frac{1}{89}\right)}} = 2.486$

⇒영가설 하에서 근사적으로 표준정규분포 따름

-기각역: $z \leq z_{0.05} = -1.645$

-통계적 결정: 검정통계량이 비기각역에 포함되므로 영가설을 기각할 수 없다.
즉, 유의수준 $\alpha = 0.05$에서 비타민 C가 감기 치료에 효과가 있다는 유의한 증거가 없다.

14. 정규분포를 따르는 모집단으로부터 크기가 15인 확률표본을 뽑았다. 표본평균은 4.7이고, 모분산은 5.67이라고 알려져 있을 때, 모평균에 대한 90% 신뢰구간을 구하라.

sol) $\overline{X} = 4.7,\ \sigma^2 = 5.67,\ n = 15,\ z_{1-\frac{1-0.9}{2}=0.95} = 1.645$

$$\Rightarrow \left(\overline{X} - z_{0.95}\sqrt{\frac{\sigma^2}{n}},\ \overline{X} + z_{0.95}\sqrt{\frac{\sigma^2}{n}}\right)$$
$$= \left(4.7 - 1.645 \times \sqrt{\frac{5.67}{15}},\ 4.7 + 1.645 \times \sqrt{\frac{5.67}{15}}\right)$$
$$= (3.689, 5.711)$$

15. $\overline{X}$를 $N(\mu, 9)$에서 추출한 크기 n인 표본의 표본평균이라 하자. 이때 $P(\overline{X} - 1 < \mu < \overline{X} + 1) = 0.95$가 되게 하는 n을 구하라.

sol) $\sigma^2 = 9,\ z_{1-\frac{1-0.95}{2}=0.975} = 1.96$

$$\Rightarrow P\left(\overline{X} - z_{0.975}\sqrt{\frac{\sigma^2}{n}},\ \overline{X} + z_{0.975}\sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

$$\Rightarrow z_{0.975}\sqrt{\frac{\sigma^2}{n}} = 1.96 \times \sqrt{\frac{9}{n}} = 1$$

$$\Rightarrow n = 1.96^2 \times 9 = 34.57 \qquad \therefore n \approx 35$$

16. 두 정규분포 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$에서 각각 크기가 100인 표본을 뽑았다. 이때, 서로 독립인 두 표본의 표본평균과 표본분산은 다음과 같다고 한다.

$$\overline{X} = 4.8,\ \ s_1^2 = 8.64,\ \ \overline{Y} = 5.6,\ \ s_2^2 = 7.8$$

sol)

1) $\sigma_1^2 = \sigma_2^2$일 때, $\mu_1 - \mu_2$에 대한 95% 신뢰구간을 구하라.

$$\Rightarrow\ S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(100-1)\times 8.64 + (100-1)\times 7.8}{100+100-2} = 8.22,$$

$$t_{1-\frac{1-0.95}{2} = 0.975,\ 100+100-2=98} = 1.984$$

$\Rightarrow$ $\mu_1 - \mu_2$의 신뢰구간

$$= \left((\overline{X}-\overline{Y}) - t_{0.975,98}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},\ (\overline{X}-\overline{Y}) + t_{0.975,98}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \right)$$

$$= \left((4.8-5.6) - 1.984\times\sqrt{8.22\left(\frac{1}{100}+\frac{1}{100}\right)},\ (4.8-5.6) + 1.984\times\sqrt{8.22\left(\frac{1}{100}+\frac{1}{100}\right)} \right)$$

$$= (-1.600, 0.000)$$

2) $\sigma_1^2 \neq \sigma_2^2$일 때, $\mu_1 - \mu_2$에 대한 95% 신뢰구간을 구하라.

$$\Rightarrow\ t'_{0.975} = \frac{\left(\frac{s_1^2}{n_1}\right)t_{0.975,\,99} + \left(\frac{s_2^2}{n_2}\right)t_{0.975,\,99}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = t_{0.975,\,99} = 1.984,$$

$\Rightarrow$ $\mu_1 - \mu_2$의 신뢰구간

$$= \left((\overline{X}-\overline{Y}) - t'_{0.975}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\ (\overline{X}-\overline{Y}) + t'_{0.975}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \right)$$

$$= \left((4.8-5.6) - 1.984\times\sqrt{\frac{8.64}{100}+\frac{7.8}{100}},\ (4.8-5.6) + 1.984\times\sqrt{\frac{8.64}{100}+\frac{7.8}{100}} \right)$$

$$= (-1.604, 0.004)$$

연습문제 8장

1. 분산분석을 정의하라.
sol) 총 변동을 2개 이상의 변동으로 나누고, 총 변동에 대한 각 변동의 상대적인 크기를 계산하여 여러 모집단의 모평균에 차이가 있는지 검정하는 방법이다.

2. 완전확률화 블록 계획법을 설명하라.
sol) 실험에 참여하는 실험단위를 특성이 비슷한 그룹으로 나눈 뒤, 그룹별로 각 처리를 받은 실험단위의 수를 동일하게 유지하는 실험방법이다.

3. 요인 실험을 설명하라.
sol) 검정하고자 하는 관심 변수를 요인이라고 하며, 둘 이상의 요인을 동시에 고려하는 실험을 요인실험이라고 한다.

4. Tukey의 HSD 검정에 대하여 설명하라.
sol) 처리그룹별로 표본의 크기가 같을 때, 모든 가능한 쌍의 두 평균이 서로 같다는 영가설을 검정할 때 적절한 다중비교 분석방법이다.

5. 다중 검정에 대하여 설명하라.
sol) 분산분석을 수행한 결과, 모평균 사이에 차이가 없다는 영가설을 기각했을 때, 어떤 쌍의 모평균이 서로 다른지 확인하는 분석방법이다.

6. 완전확률화 블록 계획법을 수행하는 목적은 무엇인지 설명하라.
sol) 실험단위가 유사한 특성을 가질 수 있도록 블록을 잘 정의하면 대립가설이 사실일 때 분산분석표의 오차평균제곱이 작아져 검정통계량이 커지므로, 영가설을 쉽게 기각할 수 있다.

7. 교호작용은 무엇인지 설명하라.
sol) 한 요인의 수준에 따라 다른 요인이 반응변수에 미치는 정도에 차이가 있으면 교호작용이 존재한다고 한다.

8. 분산분석표는 무엇을 의미하는지 설명하라.
sol) 분산분석의 결과를 표의 형태로 정리한 것으로 요인, 제곱합, 자유도, 평균제곱합, 검정통계량 F-값으로 구성된다. 분산분석표의 집단 간 제곱합이 집단 내 제곱합보다 상대적으로 크면 처리그룹 간 평균의 차이가 존재함을 의미한다.

9. 다음 세 집단의 표본 평균 간에 유의한 차이가 있는지 검정하라.($\alpha = 0.05$)

A	18, 18, 20, 21, 23, 23, 24, 26, 26, 27, 28, 29, 29, 29, 30, 30, 30, 30, 32
B	10, 16, 22, 22, 23, 26, 28, 28, 28, 29, 29, 30, 31, 32, 32, 33, 33, 38, 39
C	17, 24, 26, 27, 29, 30, 30, 33, 34, 35, 35, 36, 39

sol)

```
> aov<-aov(value~group, data=data_9)
> summary(aov)
             Df  Sum Sq Mean Sq F value Pr(>F)
group         2   152.1   76.04   2.211  0.121
Residuals    48  1650.6   34.39
```

-가설: $H_0 : \mu_A = \mu_B = \mu_C$, $H_A : not\, H_0$

-결론: F-값이 2.211이고 p-값이 0.121이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 세 집단의 표본 평균 간에 유의한 차이가 없다.

10. 다음의 표는 세 병원에 4가지 치료 물질에 대한 효과를 연구한 자료이다. 수치가 낮을수록 치료의 효과가 높음을 의미한다. 분산 분석을 수행하고 다음의 물음에 답하라.

	A	B	C	D
병원1	4.31	4.68	4.16	5.67
	4.68	6.18	3.05	6.78
	4.17	4.37	5.10	5.89
	5.75	4.35	3.56	7.90
병원2	3.13	3.56	3.87	4.58
	4.67	4.21	3.67	5.21
	3.78	5.23	4.11	5.78
	6.89	3.78	5.18	6.15
병원3	3.79	4.55	3.67	8.78
	4.17	6.77	5.67	8.31
	4.35	2.33	7.80	6.79
	3.67	3.44	3.45	7.01

sol)

1) 치료 물질 간 치료효과의 차이가 있는가?

```
> anova(lm(effect~treatment, data=data_10))
Analysis of Variance Table

Response: effect
           Df  Sum Sq Mean Sq F value    Pr(>F)
treatment   3  40.589  13.530  9.2926 7.086e-05 ***
Residuals  44  64.062   1.456
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-가설: $H_0 : \alpha_A = \alpha_B = \alpha_C = \alpha_D$, $H_A : not\, H_0$

-결론: F-값이 9.2926이고 p-값이 7.086e-05이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, 치료 물질 간 치료효과의 차이가 있다.

2) 병원 간 치료효과의 차이가 있는가?

```
> anova(lm(effect~hospital, data=data_10))
Analysis of Variance Table

Response: effect
           Df  Sum Sq Mean Sq F value Pr(>F)
hospital    2   3.696  1.8480  0.8237 0.4453
Residuals  45 100.955  2.2434
```

-가설: $H_0 : \beta_1 = \beta_2 = \beta_3$, $H_A : not\, H_0$

-결론: F-값이 0.8237이고 p-값이 0.4453이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 병원 간 치료효과의 차이가 없다.

3) 병원과 치료 물질 간의 교호작용이 유의한가?

```
> anova(lm(effect~hospital*treatment, data=data_10))
Analysis of Variance Table

Response: effect
                  Df  Sum Sq Mean Sq F value    Pr(>F)
hospital           2   3.696  1.8480  1.3856    0.2632
treatment          3  40.589 13.5295 10.1445 5.532e-05 ***
hospital:treatment 6  12.354  2.0589  1.5438    0.1921
Residuals         36  48.012  1.3337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-가설: $H_0 : (\alpha\beta)_{ij} = 0,\ i = A, B, C, D,\ j = 1, 2, 3$, $H_A : not\, H_0$

-결론: F-값이 1.5438이고 p-값이 0.1921이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 교호작용이 유의하지 않다.

4) 3)의 교호작용이 유의하면 분산분석을 수행하라.
유의하지 않기 때문에 분산분석을 수행하지 않아도 된다.

11. 분산분석표를 완성하라.

Source	SS	d.f.	MS	V.R.	p-값
처리	154	4	③	⑤	⑥
오차	①	②	④		
합	200	39			

sol)

① $200 - 154 = 46$

② $39 - 4 = 35$

③ $154/4 = 38.5$

④ $46/35 = 1.314285$

⑤ $38.5/1.314285 = 29.2935$

⑥ `> pf(29.2935,4,35,lower.tail=F)` > $9.79155e-11$

12. 다음의 표를 완성하라.

Source	SS	d.f.	MS	V.R.	p-값
처리	①	3	③	⑥	⑦
블록	183.5	3	④		
오차	26.0	②	⑤		
합	709	15			

sol)

① $709 - 183.5 - 26 = 499.5$

② $15 - 3 - 3 = 9$

③ $499.5/3 = 166.5$

④ $183.5/3 = 61.167$

⑤ $26/9 = 2.889$

⑥ $166.5/2.889 = 57.6324$

⑦ > pf(57.6324,3,9,lower.tail=F) 4 > $3.378135e-06$

13. 다음 분산분석표를 보고 질문에 답하라.

Source	SS	d.f.	MS	V.R.	p-값
A	12.315	2	6.157	29.402	<0.005
B	19.784	3	6.594	31.489	<0.005
AB	8.941	6	1.490	7.115	<0.005
처리	41.041	11			
오차	10.052	48	0.209		
합	51.093	59			

sol)

1) 어떤 종류의 실험계획법을 사용하였는가?

⇒ 이요인 완전확률화 계획법

2) 처리의 효과는 유의한가?

⇒ 처리의 평균제곱합은 $41.041/11 = 3.731$이고, F-값은 $3.731/0.209 = 17.85167$이다.
R을 통해 p-값을 구하면 $2.140674e-13$로 유의수준 0.05하에서 처리의 효과는 유의하다.

3) 교호작용은 유의한가?

⇒ AB의 p-값이 0.005보다 작으므로, 유의수준 0.05하에서 교호작용은 유의하다.

4) 분산분석 결과를 적절히 기술하라.

⇒ A의 p-값이 0.005보다 작으므로, 유의수준 0.05하에서 영가설 $H_0 : \alpha_i = 0, i = 1,2$을 기각할 수 있다. 즉, 요인 A의 수준에 따라 평균에 차이가 있다.

⇒ B의 p-값이 0.005보다 작으므로, 유의수준 0.05하에서 영가설 $H_0 : \beta_j = 0, i = 1,2,3$을 기각할 수 있다. 즉, 요인 B의 수준에 따라 평균에 차이가 있다.

14. 다음 분산분석표를 보고 질문에 답하라.

Source	SS	d.f.	MS	V.R.	p-값
처리	231	2			
오차	98	7			
합					

sol)

1) 어떤 종류의 실험계획법을 사용하였는가?

⇒ 완전확률화 계획법

2) 처리변수의 수준은 몇 개 인가?

⇒ $2+1=3$개

3) 몇 개의 관측치가 사용되었는가?

⇒ $2+7+1=10$개

4) 처리변수의 효과가 유의한가?

⇒ 처리의 평균제곱합은 $231/2=115.5$이고, 오차의 평균제곱합은 $98/7=14$이다. 따라서 F-값은 $115.5/14=8.25$이다. R을 통해 p-값을 구하면 0.01442466로 유의수준 0.05하에서 처리변수의 효과는 유의하다.

사용된 R Code

```
##8_9
group<-c(rep("A",19),rep("B",19),rep("C",13))
value<-c(18, 18, 20, 21, 23, 23, 24, 26, 26, 27, 28, 29, 29, 29, 30, 30, 30, 30, 32,
         10, 16, 22, 22, 23, 26, 28, 28, 28, 29, 29, 30, 31, 32, 32, 33, 33, 38, 39,
         17, 24, 26, 27, 29, 30, 30, 33, 34, 35, 35, 36, 39)
data_9<-data.frame(group, value)
aov<-aov(value~group, data=data_9)
summary(aov)

##8_10
effect<-c(4.31, 4.68, 4.17, 5.75, 4.68, 6.18, 4.37, 4.35, 4.16, 3.05, 5.10, 3.56, 5.67, 6.78, 5.89, 7.90,
          3.13, 4.67, 3.78, 6.89, 3.56, 4.21, 5.23, 3.78, 3.87, 3.67, 4.11, 5.18, 4.58, 5.21, 5.78, 6.15,
          3.79, 4.17, 4.35, 3.67, 4.55, 6.77, 2.33, 3.44, 3.67, 5.67, 7.80, 3.45, 8.78, 8.31, 6.79, 7.01)
hospital<-factor(rep(1:3, each=16))
treatment<-factor(rep(rep(c("A","B","C","D"),each=4),3))
data_10<-data.frame(hospital, treatment, effect)

anova(lm(effect~treatment, data=data_10))
anova(lm(effect~hospital, data=data_10))
anova(lm(effect~hospital*treatment, data=data_10))

##8_11
pf(29.2935,4,35,lower.tail=F)

##8_12
pf(57.6324,3,9,lower.tail=F)

##8_13
pf(17.85167,11,48,lower.tail=F)

##8_14
pf(8.25,2,7,lower.tail=F)
```

연습문제 9장

1. 회귀분석의 가정에 대하여 기술하라.

sol) ① 독립변수 X는 연구자에 의해서 미리 결정된 값이거나, 고정된 값이라고 가정
② 독립변수 X는 측정오차 없이 관측된다고 가정
③ 독립변수 X의 개별 관측값별로 Y의 부분모집단이 존재하며, 부분모집단은 정규분포를 따른다
④ Y의 모든 부분모집단의 분산은 같고 이를 σ^2이라 가정
⑤ Y의 부분모집단의 모평균은 일직선상에 있다고 가정(선형성의 가정)
⑥ Y는 통계적으로 서로 독립이라 가정

2. 최소제곱법을 정의하라.

sol) 모든 데이터의 편차의 제곱합을 최소로 만들어주는 β_0, β_1을 $\hat{\beta_0}, \hat{\beta_1}$로 이용하는 방법으로 $\hat{\beta_0}, \hat{\beta_1}$는 다음의 부등식을 만족해야 한다.

$$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \geq \sum_{i=1}^{n}(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)^2, \quad \forall \beta_0, \forall \beta_1$$

3, 4. 단순회귀분석에서 기울기의 추정량을 유도하라.

sol) $SS=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$이라고 하자.

$$\begin{cases} \dfrac{\partial SS}{\partial \beta_0}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0 \\ \dfrac{\partial SS}{\partial \beta_1}=-2\sum_{i=1}^{n}x_i(y_i-\beta_0-\beta_1 x_i)=0 \end{cases}$$ 을 만족시키는 β_0, β_1가 최소제곱추정량 $\hat{\beta_0}, \hat{\beta_1}$이 된다.

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)=0 & \cdots(1) \\ \sum_{i=1}^{n}x_i(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)=0 & \cdots(2) \end{cases}$$

이 때, (1)의 양변을 $\frac{1}{n}$으로 나눈 식과 (2)$-$(1)$\times\bar{x}$식을 구하면

$$\Rightarrow \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{1}{n}\sum_{i=1}^{n}\hat{\beta_0}-\frac{1}{n}\sum_{i=1}^{n}\hat{\beta_1}x_i=\bar{y}-\hat{\beta_0}-\hat{\beta_1}\bar{x}=0 & \cdots(3) \\ \sum_{i=1}^{n}x_i(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)-\bar{x}\sum_{i=1}^{n}(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)=0 & \cdots(4) \end{cases}$$

따라서, (3)을 정리하면 $\hat{\beta_0}=\bar{y}-\hat{\beta_1}\bar{x}$를 유도해 낼 수 있으며 (4)식에 구한 $\hat{\beta_0}$을 대입하면

$$\Rightarrow \sum_{i=1}^{n}(x_i-\bar{x})\{y_i-(\bar{y}-\hat{\beta_1}\bar{x})-\hat{\beta_1}x_i\}=\sum_{i=1}^{n}(x_i-\bar{x})\{y_i-\bar{y}-\hat{\beta_1}(x_i-\bar{x})\}=0$$

$$\Rightarrow \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-\hat{\beta_1}\sum_{i=1}^{n}(x_i-\bar{x})^2=0 \qquad \therefore \hat{\beta_1}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

5. R^2에 대하여 설명하라.

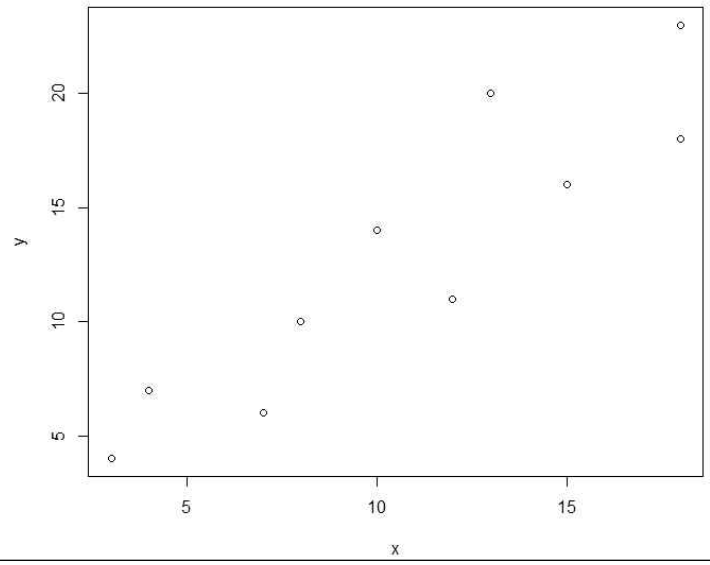
sol) 두 변수 사이의 선형성의 크기를 수학적으로 계산한 값으로 편차의 산포와 회귀방정식에서 계산한 예측 값의 산포를 비교하여 계산한다.

6. 다음은 간호사(X)와 의사(Y)에 대한 평가 점수이다. 다음 데이터를 이용하여 산점도를 그려라.

X	18	13	18	15	10	12	8	4	7	3
Y	23	20	18	16	14	11	10	7	6	4

sol)

```
> x<-c(18, 13, 18, 15, 10, 12, 8, 4, 7, 3)
> y<-c(23, 20, 18, 16, 14, 11, 10, 7, 6, 4)
> plot(x, y)
```



7. X(설명변수)와 Y(반응변수)로 회귀분석을 실시하고, 다음 물음에 답하라.

X	Y	X	Y
10	600	6.5	785
5.5	625	6	765
9	560	6.6	611
9	585	2.7	600
8	590	6	625
12.6	500	5.4	650
3	700	6	635
11	570	3.3	522
6.5	540		

sol)

1) 위 데이터를 활용하여 결정계수를 추정하라.

```
> model7 <- lm(Y~X, data=data9_7)
> summary(model7)$r.sq
[1] 0.1441538
```

추정된 결정계수의 값은 약 0.144이다.

2) $\beta_1 = 0$이라는 귀무가설에 대한 ANOVA 표를 그리고 유의수준 0.05에서 $F-$검정을 수행하라.

```
> model7_a <- lm(Y~1, data=data9_7)
> anova(model7_a, model7)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X
  Res.Df   RSS  Df  Sum of Sq    F    Pr(>F)
1     16 94986
2     15 81294   1      13693 2.5265 0.1328
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$

-결론: F-값이 2.5265이고 p-값이 0.1328이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, $\beta_1 = 0$이다.

3) $\beta_1 = 0$이라는 귀무가설에 대한 유의수준 0.05에서 $t-$검정을 수행하라.

```
> coef(summary(model7))
             Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 688.50572 49.295648 13.966866 5.293036e-10
X           -10.60288  6.670573 -1.589501 1.327986e-01
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$

-결론: T-값이 –1.589501이고 p-값이 0.1327986이므로 유의수준 0.05하에서 영가설을 기각 할 수 없다. 즉, $\beta_1 = 0$이다.

4) β_1에 대한 95% 신뢰 구간을 구하라.

```
> confint(model7, level = 0.95)
                2.5 %     97.5 %
(Intercept) 583.43454 793.576909
X           -24.82087   3.615109
```

β_1에 대한 95% 신뢰 구간은 $(-24.82087,\ 3.615109)$이다.

8. 중학교 2학년 학생 중에서 15명을 임의로 추출하여 각 학생의 악력(손의 힘, kg), 신장(cm), 체중(kg)과 원반던지기에서 던진 거리(m)를 측정하여 다음의 데이터를 얻었다.

식별번호	독립변수 악력(X_1)	독립변수 신장(X_2)	독립변수 체중(X_3)	종속변수 원반던지기 거리(Y)
1	28	146	34	22
2	46	169	57	36
3	39	160	48	24
4	25	156	38	22
5	34	161	47	27
6	29	168	50	29
7	38	154	54	26
8	23	153	40	23
9	42	160	62	31
10	27	152	39	24
11	35	155	46	23
12	39	154	54	27
13	38	157	57	31
14	32	162	53	25
15	25	142	32	23

단순회귀모형을 추정하고 그때의 결정계수의 값을 구하라.

sol)

```
> model8_a<-lm(y~x1, data=data9_8)
> model8_a

Call:
lm(formula = y ~ x1, data = data9_8)

Coefficients:
(Intercept)           x1
     11.768        0.433

> summary(model8_a)$r.sq
[1] 0.5592272
```

독립변수가 악력(X_1)일 때의 추정된 단순회귀모형은 $Y = 11.768 + 0.433X_1$이고, 이때의 결정계수는 약 0.56이다.

```
> model8_b<-lm(y~x2, data=data9_8)
> model8_b

Call:
lm(formula = y ~ x2, data = data9_8)

Coefficients:
(Intercept)           x2
    -33.777        0.383

> summary(model8_b)$r.sq
[1] 0.4698578
```

독립변수가 신장(X_2)일 때의 추정된 단순회귀모형은 $Y = -33.777 + 0.383X_2$이고, 이때의 결정계수는 약 0.47이다.

```
> model8_c<-lm(y~x3, data=data9_8)
> model8_c

Call:
lm(formula = y ~ x3, data = data9_8)

Coefficients:
(Intercept)          x3
     9.5961       0.3503

> summary(model8_c)$r.sq
[1] 0.6208367
```

독립변수가 체중(X_3)일 때의 추정된 단순회귀모형은 $Y = 9.5961 + 0.3503X_3$이고, 이때의 결정계수는 약 0.621이다.

9. 표준식단의 식사 가격이 제공되는 식사 수에 따라 달라지는지를 조사하였다.

식사 수	식사 가격
30	1.15
35	1.1
40	0.98
45	1.01
50	0.97
55	0.9
60	0.89
70	0.85
75	0.78
80	0.7
65	0.8

sol)

1) 결정계수를 계산하라.

```
> model9<-lm(price~meal, data=data9_9)
> summary(model9)$r.sq
[1] 0.9442436
```

추정된 결정계수의 값은 약 0.944이다.

2) 분산분석표를 만들고 $F-$통계량을 이용하여 $H_0 : \beta_1 = 0$을 검정하라.

```
> model9_a<-lm(price~1, data=data9_9)
> anova(model9_a, model9)
Analysis of Variance Table

Model 1: price ~ 1
Model 2: price ~ meal
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     10  0.18809
2      9 0.010487  1   0.1776 152.42 6.043e-07 ***
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$

-결론: F-값이 152.42이고 p-값이 6.043e-07이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, $\beta_1 \neq 0$이다.

3) 각 영가설에 대한 $p-$값을 구하라.

```
> coef(summary(model9))
                Estimate     Std. Error        t value      Pr(>|t|)
(Intercept)  1.362909091  0.0372519775   36.58622  4.21800e-11
meal        -0.008036364  0.0006509441  -12.34571  6.04307e-07
```

①$H_0 : \beta_0 = 0,\ H_A : \beta_0 \neq 0$
⇒p-값은 4.21800e-11이다.
②$H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$
⇒p-값은 6.04307e-07이다.

4) 위 결과를 활용하여, 결과를 해석하라.
⇒식사 가격에 대한 식사 수의 효과는 제공되는 식사 수가 1 증가할 때 식사 가격이 –0.008만큼 변화하는 것이며, 이는 유의수준 0.05 하에서 유의하다.

5) β_1에 대한 95% 신뢰 구간을 구하라.

```
> confint(model9, level = 0.95)
                  2.5 %        97.5 %
(Intercept)  1.278639263   1.447178919
X           -0.009508901  -0.006563826
```

β_1에 대한 95% 신뢰 구간은 $(-0.009608901, -0.006563826)$이다.

10. 메틸 수은에 오염된 물고기를 먹은 12명을 대상으로 메틸 수은의 섭취량과 혈중 수은 농도를 조사한 결과, 다음과 같은 표를 얻었다.

메틸수은 섭취량(x)	혈중 수은량(y)
180	90
200	120
230	125
410	290
600	310
550	290
275	170
580	375
105	70
250	105
460	205
650	480

sol)

1) 두 변수의 관계를 설명하는 회귀 방정식을 구하고 r^2을 계산하라.

```
> model10<-lm(y~x, data=data9_10)
> summary(model10)

Call:
lm(formula = y ~ x, data = data9_10)

Residuals:
    Min      1Q  Median      3Q     Max
-69.164 -36.413   5.319  23.462  84.094

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.57906   30.66189  -0.671    0.517
x             0.64075    0.07373   8.691 5.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.37 on 10 degrees of freedom
Multiple R-squared:  0.8831,    Adjusted R-squared:  0.8714
F-statistic: 75.53 on 1 and 10 DF,  p-value: 5.66e-06
```

추정된 회귀 방정식은 $Y=-20.57906+0.64075X$이고,
이때의 결정계수는 약 0.8831이다.

2) $H_0 : \beta_1 = 0$에 대해서 $F-$검정과 $t-$검정을 수행하라.

①$F-$검정

```
> model10_a<-lm(y~1, data=data9_10)
> anova(model10_a, model10)
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1     11 183892
2     10  21500   1    162392 75.531 5.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$
-결론: F-값이 75.531이고 p-값이 5.66e-06이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, $\beta_1 \neq 0$이다.

②$t-$검정

```
> coef(summary(model10))
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -20.5790583 30.66188745 -0.6711608 5.173203e-01
x             0.6407458  0.07372638  8.6908623 5.659702e-06
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$
-결론: T-값이 8.6908623이고 p-값이 5.659702e-06이므로 유의수준 0.05하에서 영가설을 기각 할 수 있다. 즉, $\beta_1 \neq 0$이다.

11. 16명의 건강한 성인 남자들의 체중(kg)과 혈당량$(mg/100ml)$을 조사하여 다음과 같은 결과를 얻었다.

체중	혈당량	체중	혈당량
64	108	82.1	101
75.3	109	78.9	85
73	104	76.7	99
82.1	102	82.1	100
76.2	105	83.9	108
95.7	121	73	104
59.4	79	64.4	102
93.4	107	77.6	87

sol)

1) 단순선형회귀 방정식을 구하고 ANOVA와 $t-$검정을 이용해서 $H_0 : \beta_1 = 0$을 검정하라.

①$F-$검정

```
> model11<-lm(blood~weight, data=data9_11)
> model11_a <-lm(blood~1, data=data9_11)
> anova(model11_a, model11)
Analysis of Variance Table

Model 1: blood ~ 1
Model 2: blood ~ weight
  Res.Df    RSS  Df  Sum of Sq      F   Pr(>F)
1     15 1573.4
2     14 1204.6   1     368.8 4.2861 0.0574 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$

-결론: F-값이 4.2861이고 p-값이 0.0574이므로 유의수준 0.05하에서 영가설을 기각 할 수 없다. 즉, $\beta_1 = 0$이다.

②$t-$검정

```
> coef(summary(model11))
             Estimate  Std. Error   t value   Pr(>|t|)
(Intercept) 61.8769330 19.1890348 3.224599 0.006113973
weight       0.5097504  0.2462225 2.070283 0.057397282
```

-가설: $H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0$

-결론: T-값이 2.070283이고 p-값이 0.057397282이므로 유의수준 0.05하에서 영가설을 기각 할 수 없다. 즉, $\beta_1 = 0$이다.

2) $H_0 : \rho = 0$을 검정하고 ρ에 대한 95% 신뢰구간을 구하라.

```
> cor.test(~blood+weight, data=data9_11)

        Pearson's product-moment correlation

data:  blood and weight
t = 2.0703, df = 14, p-value = 0.0574
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01521939  0.79020296
sample estimates:
      cor
0.4841384
```

-가설: $H_0 : \rho = 0,\ H_A : \rho \neq 0$
-결론: t-값이 2.0703이고 p-값이 0.0574이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, $\rho = 0$이다.
- ρ에 대한 95% 신뢰구간은 $(-0.01521939, 0.79020296)$이다.

3) 체중이 $95kg$인 사람의 혈당량 수준은 얼마나 될 것으로 예측되는가? 그에 대한 95% 예측 구간을 구하라. 이때 α는 0.05로 한다.

```
> new9_11<-data.frame(weight=95)
> predict(model11, new9_11, interval="prediction")
        fit      lwr      upr
1  110.3032 87.77959 132.8269
```

-예측: 체중이 95kg인 사람은 혈당량 수준이 110.3032일 것으로 예측된다.
-95% 신뢰구간은 $(87.77959, 132.8269)$이다.

12. 대학교 1학년 학생을 무작위로 20명 추추하여 이들의 대입입학성적(x)와 1학년 1학기 중간고사 성적(y)간의 관계를 분석하려고 한다. 데이터가 다음과 같을 때, 단순회귀모형을 추정하라.

X	170	174	128	147	166	152	157	125	182	174
Y	698	645	578	518	725	625	558	485	745	698
X	192	179	130	150	102	125	146	171	185	133
Y	778	798	471	538	458	625	485	611	745	538

sol)

```
> model12<-lm(y~x, data=data9_12)
> model12

Call:
lm(formula = y ~ x, data = data9_12)

Coefficients:
(Intercept)            x
     42.699        3.714
```

추정된 회귀 방정식은 $Y = 42.699 + 3.714X$이다.

13. 다음의 자료에서 다중회귀방정식을 구하라.

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
28	39	4	18	88	76	50	5
68	52	33	9	54	79	44	11
59	89	17	3	73	48	57	9
91	57	31	7	87	90	33	6
70	28	35	19	47	55	11	20
38	34	3	25	60	83	24	11
46	42	16	17	65	50	21	25
57	52	6	26	57	44	31	18
89	88	41	13	85	79	30	20
48	90	24	3	28	24	5	22
74	38	22	13	40	40	20	17
78	83	41	11	87	35	15	27
43	30	9	24	80	55	9	21
76	45	33	14	49	45	28	17
72	47	36	18	57	46	19	17
61	90	17	0	32	37	4	21
63	63	14	16	52	47	29	3
77	34	35	22	42	28	23	21
85	76	33	23	49	61	8	7
31	26	13	18	63	35	31	26
79	68	34	26	89	68	65	6
92	85	28	10	67	80	29	10

sol)

```
> model13<-lm(y~x1+x2+x3, data=data9_13)
> model13

Call:
lm(formula = y ~ x1 + x2 + x3, data = data9_13)

Coefficients:
(Intercept)          x1          x2          x3
     7.5375      0.4178      0.7946      0.7901
```

추정된 회귀 방정식은 $Y = 7.5375 + 0.4178X_1 + 0.7946X_2 + 0.7901X_3$이다.

14. 13번에서 구한 식을 이용하여 다음 각 조건의 y값을 계산하고 X_j값에 대한 ① 95% 신뢰구간과 ② 95% 예측구간을 설정하라.

sol)

1) $x_{1j} = 95, x_{2j} = 35, x_{3j} = 18$
2) $x_{1j} = 50, x_{2j} = 20, x_{3j} = 10$
3) $x_{1j} = 50, x_{2j} = 6, x_{3j} = 11$
4) $x_{1j} = 95, x_{2j} = 35, x_{3j} = 18$

```
>new9_14<-data.frame(x1=c(95,50,50,1),
                     x2=c(35,20,6,2), x3=c(18,10,11,15))
```

①

```
> predict(model13, new9_14, interval="confidence")
       fit       lwr       upr
1 89.26210 77.796310 100.72788
2 52.22099 45.740297  58.70168
3 41.88609 33.386008  50.38617
4 21.39586  6.577963  36.21375
```

②

```
> predict(model13, new9_14, interval="prediction")
       fit       lwr       upr
1 89.26210 60.610800 117.91339
2 52.22099 25.175991  79.26598
3 41.88609 14.287468  69.48471
4 21.39586 -8.753811  51.54552
```

사용된 R Code

```
##9_6
x<-c(18, 13, 18, 15, 10, 12, 8, 4, 7, 3)
y<-c(23, 20, 18, 16, 14, 11, 10, 7, 6, 4)
plot(x, y)

##9_7
X<-c(10, 5.5, 9, 9, 8, 12.6, 3, 11, 6.5, 6.5, 6, 6.6, 2.7, 6, 5.4, 6, 3.3)
Y<-c(600, 625, 560, 585, 590, 500, 700, 570, 540, 785, 765, 611, 600, 625, 650, 635, 522)
data9_7<-data.frame(X, Y)

model7 <- lm(Y~X, data=data9_7); summary(model7)$r.sq

model7_a <- lm(Y~1, data=data9_7); anova(model7_a, model7)

coef(summary(model7))

confint(model7, level = 0.95)

##9_8
x1<-c(28, 46, 39, 25, 34, 29, 38, 23, 42, 27, 35, 39, 38, 32, 25)
x2<-c(146, 169, 160, 156, 161, 168, 154, 153, 160, 152, 155, 154, 157, 162, 142)
x3<-c(34, 57, 48, 38, 47, 50, 54, 40, 62, 39, 46, 54, 57, 53, 32)
y<-c(22, 36, 24, 22, 27, 29, 26, 23, 31, 24, 23, 27, 31, 25, 23)
data9_8<-data.frame(y, x1, x2, x3)

model8_a<-lm(y~x1, data=data9_8)
model8_a
summary(model8_a)$r.sq

model8_b<-lm(y~x2, data=data9_8)
model8_b
summary(model8_b)$r.sq

model8_c<-lm(y~x3, data=data9_8)
model8_c
summary(model8_c)$r.sq

##9_9
meal<-c(30, 35, 40, 45, 50, 55, 60, 70, 75, 80, 65)
price<-c(1.15, 1.1, 0.98, 1.01, 0.97, 0.9, 0.89, 0.85, 0.78, 0.7, 0.8)
data9_9<-data.frame(meal, price)

model9<-lm(price~meal, data=data9_9)
summary(model9)$r.sq

model9_a<-lm(price~1, data=data9_9)
anova(model9_a, model9)

coef(summary(model9))

confint(model9)
```

```
##9_10
x<-c(180, 200, 230, 410, 600, 550, 275, 580, 105, 250, 460, 650)
y<-c(90, 120, 125, 290, 310, 290, 170, 375, 70, 105, 205, 480)

data9_10<-data.frame(y, x)
model10<-lm(y~x, data=data9_10)
summary(model10)

model10_a<-lm(y~1, data=data9_10)
anova(model10_a, model10)
coef(summary(model10))

##9_11
weight<-c(64, 75.3, 73, 82.1, 76.2, 95.7, 59.4, 93.4, 82.1, 78.9, 76.7, 82.1, 83.9, 73, 64.4, 77.6)
blood<-c(108, 109, 104, 102, 105, 121, 79, 107, 101, 85, 99, 100, 108, 104, 102, 87)
data9_11<-data.frame(weight, blood)

model11<-lm(blood~weight, data=data9_11)
model11_a <-lm(blood~1, data=data9_11)
anova(model11_a, model11)
coef(summary(model11))

cor.test(~blood+weight, data=data9_11)

new9_11<-data.frame(weight=95)
predict(model11, new9_11, interval="prediction")



##9_12
x<-c(170, 174, 128, 147, 166, 152, 157, 125, 182, 174, 192, 179, 130, 150, 102, 125, 146, 171, 185, 133)
y<-c(698, 645, 578, 518, 725, 625, 558, 485, 745, 698, 778, 798, 471, 538, 458, 625, 485, 611, 745, 538)
data9_12<-data.frame(y, x)

model12<-lm(y~x, data=data9_12); model12

##9_13
y<-c(28, 68, 59, 91, 70, 38, 46, 57, 89, 48, 74, 78,  43, 76, 72, 61, 63, 77, 85, 31, 79, 92,
     88, 54, 73, 87, 47, 60, 65, 57, 85, 28, 40, 87, 80, 49, 57, 32, 52, 42, 49, 63, 89, 67)
x1<-c(39, 52, 89, 57, 28, 34, 42, 52, 88, 90, 38, 83, 30, 45, 47, 90, 63, 34, 76, 26, 68, 85,
      76, 79, 48, 90, 55, 83, 50, 44, 79, 24, 40, 35, 55, 45, 46, 37, 47, 28, 61, 35, 68, 80)
x2<-c(4, 33, 17, 31, 35, 3, 16, 6, 41, 24, 22, 41, 9, 33, 36, 17, 14, 35, 33, 13, 34, 28,
      50, 44, 57, 33, 11, 24, 21, 31, 30, 5, 20, 15, 9, 28, 19, 4, 29, 23, 8, 31, 65, 29)
x3<-c(18, 9, 3, 7, 19, 25, 17, 26, 13, 3, 13, 11, 24, 14, 18, 0, 16, 22, 23, 18, 26, 10,
       5, 11, 9, 6, 20, 11, 25, 18, 20, 22, 17, 27, 21, 17, 17, 21, 3, 21, 7, 26, 6, 10)
data9_13<-data.frame(y, x1, x2, x3)

model13<-lm(y~x1+x2+x3, data=data9_13); model13

##9_14
new9_14<-data.frame(x1=c(95,50,50,1), x2=c(35,20,6,2), x3=c(18,10,11,15))
predict(model13, new9_14, interval="confidence")
predict(model13, new9_14, interval="prediction")
```

연습문제 10장

1. 상관계수란 무엇인가?

sol) 두 변수 X, Y 사이의 선형성의 강도를 의미한다.
상관계수가 양수이면 두 변수는 양의 상관관계에 있다고 말하고, 이는 한 변수의 값이 증가하면 다른 변수의 값도 증가함을 의미한다. 상관계수가 음수이면 두 변수는 음의 상관관계가 있다고 말하고, 이는 한 변수의 값이 증가하면 다른 변수의 값은 감소함을 의미한다. 만약 0이면 두 변수 사이에 선형 관계가 없음을 의미한다.

2. 편상관계수란 무엇인가?

sol) 다른 변수들의 영향을 모두 제거 했을 때 두 변수 사이의 연관성의 크기를 의미한다.

3. 상관분석과 회귀분석의 공통정과 차이점을 서술하라.

sol)

공통점	두 변수 사이의 관계를 추론하고자 할 때 사용하는 분석방법이다.
차이점	-상관분석은 두 변수가 있을 때 종속변수를 선택할 필요가 없지만, 회귀분석은 어떠한 변수를 종속변수로 선택하는지에 따라 분석결과에 차이가 있다. -상관분석은 두 변수 X, Y를 모두 확률변수로 간주하는 반면, 회귀분석은 독립변수는 고정된 상수로, 종속변수는 확률표본으로 간주한다.

4. 상관분석으로 인과관계를 알 수 있는지 설명하라.

sol) 알 수 없다. 두 변수 사이의 연관성의 크기와 상관없이 연관성은 인과성을 의미하지 않기 때문이다.

5. 다중상관계수란 무엇인가?

sol) 두 변수 간의 상관계수를 3개 이상의 확률변수가 있는 경우로 확장한 개념으로 여러 변수 사이의 연관성의 크기를 의미한다.

6. 상관분석의 가정을 기술하라.

sol) ① X의 값이 주어졌을 때 Y의 부분모집단은 정규분포를 따른다.
② Y의 값이 주어졌을 때 X의 부분모집단은 정규분포를 따른다.
③ 두 변수 X, Y는 이변량 정규분포를 따른다.
④ Y의 부분모집단의 분산은 X와 상관없이 모두 같다.
⑤ X의 부분모집단의 분산은 Y와 상관없이 모두 같다.

7. 다음은 10명의 환자들에게 대한 간호사 평가(X)와 의사 평가(Y)에 대한 점수이다.

X	18	13	18	15	10	12	8	4	7	3
Y	23	20	18	16	14	11	10	7	6	4

두 변수의 상관계수를 구하고, 유의성을 검정하라.

sol)

```
> cor.test(X,Y, data=data10_7)

        Pearson's product-moment correlation

data:  X and Y
t = 6.2827, df = 8, p-value = 0.0002372
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6627116 0.9792608
sample estimates:
      cor
0.9118562
```

-표본 상관계수: $r = 0.912$
-가설: $H_0 : \rho = 0,\ H_A : \rho \neq 0$
-결론: t-값이 6.2827이고 p-값이 0.0002372이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, ρ는 유의하다.

8. 메틸 수은에 오염된 물고기를 먹은 12명을 대상으로 메틸 수은의 섭취량과 혈중 수은 농도를 조사한 결과 다음의 결과를 얻었다.

메틸수은 섭취량(x)	혈중 수은량(y)
180	90
200	120
230	125
410	290
600	310
550	290
275	170
580	375
105	70
250	105
460	205
650	480

sol)

1) 두 변수 간의 관계를 설명하는 회귀 방정식을 구하고 결정계수를 계산하라.

```
> model8<-lm(y~x, data=data10_8)
> coef(model8)
(Intercept)           x
-20.5790583   0.6407458
> summary(model8)$r.sq
[1] 0.8830834
```

추정된 회귀 방정식은 $Y = -20.5790583 + 0.6407458X$이고,
이때의 결정계수는 약 0.8831이다.

2) 1)을 이용하여 상관계수를 구하라.

```
> sqrt(summary(model8)$r.sq)
[1] 0.9397252
```

추정된 상관계수는 0.9397252이다.

3) 2)가 유의한지 검정하라.

```
> cor.test(x, y, data=data10_8)

        Pearson's product-moment correlation

data:  x and y
t = 8.6909, df = 10, p-value = 5.66e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7940772 0.9833153
sample estimates:
      cor
0.9397252
```

-가설: $H_0 : \rho = 0,\ H_A : \rho \neq 0$

-결론: t-값이 8.6909이고 p-값이 5.66e-06이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, ρ는 유의하다.

9. X는 설명변수, Y는 반응변수이다. 다음의 물음에 답하라.

$$\sum x_i = 82{,}500,\ \sum y_i^2 = 871,\ \sum x_i^2 = 523{,}750{,}000,\ \sum x_i y_i = 510{,}500,$$
$$\sum y_i = 107,\ n = 100$$

sol)

1) 표본상관계수를 구하라.

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} = \frac{100\times 510500 - (82500)\times(107)}{\sqrt{100\times 523750000 - (82500)^2}\sqrt{100\times 871 - (107)^2}}$$
$$= 0.7191228$$

2) $H_0 : \rho = 0$를 검정하고 결론을 내려라.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.719228\times\sqrt{\frac{100-2}{1-(0.719228)^2}} = 10.24482$$

$\Rightarrow$ 유의수준 α를 0.05라고 가정하면 자유도가 $100-2$인 $t-$분포의 임계값은 1.984467이다. 검정통계량이 임계값보다 크므로 영가설 $H_0 : \rho = 0$을 기각한다.

3) 그 검정에 대한 $p-$값을 결정하라.

```
> pt(10.24482, df=98, lower.tail = F)*2
[1] 3.562863e-17
```

$p-$값은 자유도가 98인 $t-$분포가 10.24482보다 크거나 –10.24482보다 작을 확률과 같으므로 R함수를 실행하면, $p-$값은 3.562863e-17이다.

4) $p-$값에 대한 95% 신뢰구간을 구하라.

i) $z_r = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2} ln\left(\frac{1+0.7191228}{1-0.7191228}\right) = 0.9058259$

ii) $z_r \pm z_{1-\frac{\alpha}{2}}\left(\frac{1}{\sqrt{n-3}}\right) = 0.9058259 \pm 1.96\left(\frac{1}{\sqrt{100-97}}\right)$

$\Rightarrow (0.7068217,\ 1.1048301)$

iii) 위 신뢰구간의 양 끝 값을 z_r 변환의 역변환으로 변환하면 $r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$ 이므로

$\Rightarrow (0.6086800,\ 0.8022273)$

10. 다음의 데이터를 이용하여 질문에 답하라.

Y	X_1	X_2	X_3
115	0.38	899	56
86	1.65	167	158
19	0.16	73	152
6	0.08	146	35
23	0.02	82	60
147	1.98	2483	1993
27	0.15	404	30
140	0.25	2438	72
345	0.55	780	12
92	0.22	517	5
85	0.09	346	5
24	0.17	82	12
185	0.28	2932	108
30	0.19	658	33
9	0.03	103	5
76	0.21	2339	5
51	0.09	31	36
73	0.06	158	5
47	0.08	773	5
48	0.12	545	67
16	0.03	5	5

sol)

1) 가능한 모든 변수 쌍들의 단순상관계수를 계산하라.

```
> cor(data10_10)
           y         x1        x2        x3
y  1.0000000 0.3739267 0.5132981 0.2034367
x1 0.3739267 1.0000000 0.3368211 0.7755871
x2 0.5132981 0.3368211 1.0000000 0.4338756
x3 0.2034367 0.7755871 0.4338756 1.0000000
```

2) 1)의 유의성을 검정하라.

```
> rcorr(as.matrix(data10_10))$P
             y          x1          x2          x3
y           NA 9.495276e-02  0.01732181 3.764373e-01
x1  0.09495276          NA  0.13543166 3.614821e-05
x2  0.01732181 0.103543166          NA 4.940368e-02
x3  0.37643734 3.614821e-05  0.04940368          NA
```

-가설: $H_0 : \rho_{i,j} = 0,\ H_A : \rho_{i,j} \neq 0,\ (i = 1,2,3, j = 0,1,2,\ i \neq j)$

-결론: 유의수준이 0.05라고 가정할 때, 각 $\rho_{1,0}, \rho_{2,0}, \rho_{2,1}, \rho_{3,0}, \rho_{3,1}, \rho_{3,2}$에 대하여 가설 검정을 한 결과 $p-$값이 0.05보다 작은 경우가 $\rho_{2,0}, \rho_{3,1}, \rho_{3,2}$이다. 따라서 이 경우 영가설을 기각할 수 있으므로 $\rho_{2,0}, \rho_{3,1}, \rho_{3,2}$는 유의하다.

3) 네 변수들 간의 다중상관계수를 계산하라.

```
> sqrt(summary(lm(y~x1+x2+x3, data10_10))$r.sq)
[1] 0.6172522
```

네 변수들 간의 추정된 다중상관계수는 0.6172522이다.

4) 3)의 다중상관계수의 유의성을 검정하라.

```
> anova(lm(y~1, data10_10), lm(y~x1+x2+x3, data10_10))
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     20 123754
2     17  76604  3     47150 3.4879 0.0388 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

-가설: $H_0 : \rho = 0,\ H_A : \rho \neq 0$

-결론: F-값이 3.4879이고 p-값이 0.0388이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, ρ는 유의하다.

사용된 R Code

```
##10_7
X<-c(18, 13, 18, 15, 10, 12, 8, 4, 7, 3)
Y<-c(23, 20, 18, 16, 14, 11, 10, 7, 6, 4)
data10_7<-data.frame(Y, X)
cor.test(X,Y, data=data10_7)

##10_8
x<-c(180, 200, 230, 410, 600, 550, 275, 580, 105, 250, 460, 650)
y<-c(90, 120, 125, 290, 310, 290, 170, 375, 70, 105, 205, 480)
data10_8<-data.frame(y, x)
model8<-lm(y~x, data=data10_8)
coef(model8)
summary(model8)$r.sq

sqrt(summary(model8)$r.sq)

cor.test(x, y, data=data10_8)

##10_9
qt(0.975, df=98)
pt(10.24482, df=98, lower.tail = F)*2

r<-0.7191228
zr <- 0.5*log((1+r)/(1-r))
zCI <- zr+qnorm(0.975)*(1/sqrt(100-3))*c(-1,1)
(exp(2*zCI)-1)/(exp(2*zCI)+1)

##10_10
y<-c(115, 86, 19, 6, 23, 147, 27, 140, 345, 92, 85, 24, 185, 30, 9, 76, 51, 73, 47, 48, 16)
x1<-c(0.38, 1.65, 0.16, 0.08, 0.02, 1.98, 0.15, 0.25, 0.55, 0.22, 0.09, 0.17, 0.28, 0.19,
      0.03, 0.21, 0.09, 0.06, 0.08, 0.12, 0.03)
x2<-c(899, 167, 73, 146, 82, 2483, 404, 2438, 780, 517, 346, 82, 2932, 658, 103, 2339, 31,
      158, 773, 545, 5)
x3<-c(56, 158, 152, 35, 60, 1993, 30, 72, 12, 5, 5, 12, 108, 33, 5, 5, 36, 5, 5, 67, 5)
data10_10<-data.frame(y, x1, x2, x3)

cor(data10_10)

install.packages("Hmisc")
library(Hmisc)
rcorr(as.matrix(data10_10))$P

sqrt(summary(lm(y~x1+x2+x3, data10_10))$r.sq)

anova(lm(y~1, data10_10), lm(y~x1+x2+x3, data10_10))
```

연습문제 11장

1. 자유도가 n인 카이제곱분포와 정규분포 사이의 관계를 설명하라.

sol) $z_1, \cdots, z_n$이 각각 독립적이고, 동일하게 표준정규분포를 따를 때, $\sum_{i=1}^{n} z_i^2$는 자유도가 n인 카이제곱분포를 따른다.

2. 자유도가 n인 카이제곱분포의 평균과 분산을 구하라.

sol) 자유도가 n인 카이제곱분포의 평균은 n이고 분산은 $2n$이다.

3. 적합도 검정에서 카이제곱분포의 자유도 계산 방법을 설명하라.

sol) k가 범주의 개수이고 r은 제약조건의 수 일 때, 카이제곱분포의 자유도는 $k-r$이다.

4. 적합도 검정에서 기대도수의 크기에 대한 Cochran의 법칙을 설명하라.

sol) 모집단이 정규분포와 같이 종형의 분포를 따르는 경우에도 각 범주별 기대도수는 최소한 1보다 커야 한다.

5. 적합도 검정에서 기대도수의 크기가 작을 때 해결하는 방법을 설명하라.

sol) 하나 이상의 범주가 1보다 작은 기대도수를 갖는다면 해당 범주를 모두 합친 뒤에 카이제곱 검정을 수행하면 된다. 이때, 여러 범주를 하나로 합치면 범주의 개수가 줄어들기 때문에 자유도도 감소함을 주의해야 한다.

6. 분할표로 X^2 값을 구할 때 자유도의 계산 방법을 설명하라.

sol) 두 범주형의 변수가 가질 수 있는 값의 수준이 각각 r, c라고 할 때, 검정통계량 X^2의 자유도는 $(r-1)\times(c-1)$이다.

7. 독립성 검정에서 기대도수의 계산방법을 설명하라.

sol) 각 칸의 기대도수는 칸의 행의 주변합과 열의 주변합을 곱한 뒤에 전체 표본의 크기인 n으로 나누면 된다.

8. 독립성 검정과 동질성 검정의 차이를 설명하라.

sol) 두 검정은 표본의 추출방식과 해석이 서로 다르다. 독립성 검정의 경우 결과를 해석할 때, 두 범주형 변수가 서로 독립적으로 관측됨을 의미하고 동질성 검정의 경우 결과를 해석할 때, 한 범주형 변수의 수준별로 다른 범주형 변수의 분포가 서로 동일함을 의미하기 때문에 개념적 차이가 있다.

9. 동질성 검정에서 기대도수의 계산방법을 설명하라.

sol) 각 모집단별 수준에 따른 기대도수는 해당 수준의 개수를 전체 표본의 크기인 n으로 나누고, 모집단에서 뽑힌 표본의 개수를 곱하면 된다.

10. 어떤 경우에 카이제곱 검정보다 Fisher의 정확 검정이 바람직한지 설명하라.

sol) $n<20$이거나 혹은 $20 \le n < 40$이나 1개 이상의 칸의 기대도수가 5보다 작을 경우에는 카이제곱 검정을 사용하는 것이 바람직하지 않으며, Fisher의 정확 검정을 사용하는 것이 바람직하다.

11. 다음의 용어를 정의하라.

sol) 1) 관측연구: 연구 대상자의 관심변수의 관측값을 이용하는 연구를 의미하다. 다시 말해 실험이 아닌 조사를 통하여 얻은 데이터를 연구하는 방법론을 의미한다.

2) 위험요소: 결과변수와 관련이 있을 것으로 예상되는 변수를 의미한다.

3) 결과변수: 위험요소의 결과로 예상되어지는 변수를 의미한다.

4) 후향적 연구: 결과변수의 범주에 따라 서로 다른 확률표본을 뽑는다. 그런 다음 각 확률표본에 포함되는 개체의 과거를 조사하여, 위험요소에 노출된 적이 있는지 없는지 조사하는 연구방법론을 의미한다.

5) 전향적 연구: 2개 이상의 확률 표본을 뽑아 수행되는 관측연구이다. 예를 들어 위험요소에 노출된 표본과 위험요소에 노출되지 않은 표본을 뽑는다. 그런 다음 각 개체를 추적하여 일정한 시간이 흐른 뒤에 결과변수를 관측하여, 위험요소와 결과변수 사이의 연관성을 검정하는 연구방법론을 의미한다.

6) 상대위험도: 위험요소에 노출된 사람이 병에 걸릴 확률이 위험요소에 노출되지 않은 사람이 병에 걸릴 확률에 비해 얼마나 큰지 알려주는 측도이다.

7) 오즈: 실패 확률에 대한 성공 확률의 비를 의미한다.

8) 오즈비: 환자집단의 오즈와 정상집단의 오즈의 비를 의미한다.

9) 교란변수: 질병과 위험요소에 대한 연구를 수행할 때, 질병과 연관성이 있고 동시에 위험요소와도 연관성이 있는 변수를 의미한다.

12. Mantel-Haenszel 검정을 사용하는 경우가 적절한 조건에 대하여 설명해보라.

sol) 교란변수의 효과를 보정한 뒤에 질병과 위험요소 사이의 관계를 검정하는 것으로 모든 변수가 범주형 데이터인 경우에만 적용 가능하다.

13. 다음은 성별에 따른 정당 선호의 빈도표이다.

성별/정당	A	B	C	D
여자	600	34	16	18
남자	307	10	17	10

위 데이터를 이용하여 성별과 정당 선호도가 서로 독립인지 $\alpha = 0.05$에서 검정하라.

sol)

```
> chisq.test(data11_13)

        Pearson's Chi-squared test

data:  data11_13
X-squared = 7.0499, df = 3, p-value = 0.07032
```

-가설: H_0 : 성별과 정당 선호도는 서로 독립이다.
H_A : 두 변수는 독립이 아니다.

-결론: X^2-값이 7.0499이고 p-값이 0.07032이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 성별과 정당 선호도는 서로 독립이다.

14. 다음 데이터를 이용하여 인종과 정당 선호도가 서로 독립인지 $\alpha = 0.05$에서 검정하라.

성별/정당	A	B	C	D	합
백인	356	20	3	9	388
미국-흑인	226	11	10	9	256
합	582	31	13	18	644

sol)

```
> chisq.test(data11_14)

        Pearson's Chi-squared test

data:  data11_14
X-squared = 8.7308, df = 3, p-value = 0.03309
```

-가설: H_0 : 인종과 정당 선호도는 서로 독립이다.
H_A : 두 변수는 독립이 아니다.

-결론: X^2-값이 8.7308이고 p-값이 0.03309이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, 성별과 정당 선호도는 서로 독립이 아니다.

15. 다음은 150명의 만성보균자와 500명의 대조군의 혈액형을 조사한 결과이다.

혈액형	보균자	비보균자	합
O	72	230	302
A	54	192	246
B	16	63	79
AB	8	15	23
합	150	500	650

항원 만성보균자와 대조군의 혈액형 분포가 동질한지 유의수준 $\alpha = 0.05$에서 검정하고, $p-$값을 구하라.

sol)

```
> chisq.test(data11_15)

        Pearson's Chi-squared test

data:  data11_15
X-squared = 2.4052, df = 3, p-value = 0.4927
```

-가설: H_0 : 항원 만성보균자와 대조군의 혈액형 분포는 동일하다.
H_A : 항원 만성보균자와 대조군의 혈액형 분포는 동일하지 않다.

-결론: X^2-값이 2.4052이고 p-값이 0.4927이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 항원 만성보균자와 대조군의 혈액형 분포는 동일하다.

16. 대도시에 거주하고 있는 15세 미만의 아이들의 인종과 헤모글로빈 수치를 조사한 결과는 다음과 같다.

	헤모글로빈 수치			
인종	**>10.0**	**9.0-9.9**	**<9.0**	**합**
A	54	192	246	200
B	16	63	79	385
C	8	15	23	110
합	150	500	650	695

$\alpha = 0.05$에서 두 변수가 서로 독립인지 검정하고, $p-$값을 구하라.

sol)

```
> chisq.test(data11_16)

        Pearson's Chi-squared test

data:  data11_16
X-squared = 2.2658, df = 4, p-value = 0.687
```

-가설: H_0 : 15세 미만의 아이들의 인종과 헤모글로빈 수치는 독립이다.
H_A : 두 변수는 독립이 아니다.
-결론: X^2-값이 2.2658이고 p-값이 0.687이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 15세 미만의 아이들의 인종과 헤모글로빈 수치는 독립이다

※연습문제 17~20번의 상황이 동질성 검정에 대한 것인지, 아니면 독립성 검정에 대한 것인지 답하라.

17. 세 도시에서 소아마비에 대한 면역성이 있는 미취학 아동의 비율에 차이가 있는지 확인하기 위하여, 세 도시에서 각각 미취학 아동 표본을 뽑아 면역성이 있는 아동의 비율을 비교하였다.

sol) 동질성 검정

18. 흡연과 호흡기 질환 사이의 연관성을 조사하기 위하여, 성인의 확률표본을 담배 소비량과 호흡기 질환의 정도에 따라 분류하였다.

sol) 독립성 검정

19. 산모의 흡연과 태아의 선척적 장애에 대한 연관성을 검정하기 위하여, 엄마와 자녀로 이루어진 표본으로부터 엄마의 흡연 여부, 자녀의 선척적 장애 유무를 조사하였다.

sol) 독립성 검정

20. 저혈압인 사람이 그렇지 않은 사람에 비하여 우울증에 걸릴 확률이 더 높다고 알려져 있다. 이를 확인하기 위하여, 저혈압 환자와 대조군을 각각 500명 뽑아, 우울증 환자의 비율을 계산하였다.

sol) 동질성 검정

사용된 R Code

```
##11_13
data11_13 <- matrix(c(600,307,34,10,16,17,18,10), nrow=2)
dimnames(data11_13)<-list(sex=c("female","male"), party=c("A", "B", "C", "D"))
chisq.test(data11_13)



##11_14
data11_14 <- matrix(c(356,226,20,11,3,10,9,9), nrow=2)
dimnames(data11_14)<-list(race=c("white","black"), party=c("A", "B", "C", "D"))
data11_14
chisq.test(data11_14)

##11_15
data11_15 <- matrix(c(72,54,16,8,230,192,63,15), nrow=4)
dimnames(data11_15)<-list(blood=c("O","A", "B", "AB"),disease=c("Y", "N"))
data11_15
chisq.test(data11_15)

##11_16
data11_16 <- matrix(c(54,16,8,192,63,15,246,79,23), nrow=3)
dimnames(data11_16)<-list(race=c("A", "B", "C"),hemo=c(">10", "9~9.9","<9"))
data11_16
chisq.test(data11_16)
```

연습문제 12장

1. 비모수 통계학은 무엇인지 설명하라.

sol) 모집단의 형태와 관계없이 데이터로부터 직접 확률을 계산하여 검정하는 방법을 의미한다.

2. 비모수 분석의 장단점을 설명하라.

sol)

장점	①모집단의 모수에 대한 설명이 필요 없다. ②비모수적 검정은 모집단의 확률분포함수에 대한 가정을 할 수 없는 경우에도 이용할 수 있다. ③일부 비모수적 검정은 모수적 방법에 비해 계산이 간단하고 적용이 쉽다. ④관측값이 부정확하여 단순히 관측값의 순서를 정하거나 분류하는 경우에는 비모수적 방법을 사용하는 것도 바람직할 수 있다.
단점	①모수적 방법으로서 처리할 수 있는 데이터를 비모수적 방법으로 분석하는 것은 좋지 않으며, 따라서 모수적 방법으로 검정할 수 있는 데이터는 모수적 방법으로 분석해야 한다. ②일부 비모수적 방법은 계산량이 많기 때문에 표본의 크기가 큰 경우 비모수적 방법은 비효율적이다.

3. 15명의 학생의 권위주의에 대한 성향을 측정한 결과가 다음과 같다.

일련번호	점수	일련번호	점수
1	75	9	82
2	90	10	104
3	85	11	88
4	110	12	124
5	115	13	110
6	95	14	76
7	132	15	98
8	74		

모집단의 중위수가 100인지 아닌지 유의수준 $\alpha = 0.05$에서 검정하고 $p-$값을 구하라.

sol)

```
> SIGN.test(score, md=100, alternative = "two.sided", conf.level = 0.95)

        One-sample Sign-Test

data:  score
s = 6, p-value = 0.6072
alternative hypothesis: true median is not equal to 100
95 percent confidence interval:
  82.53451 110.00000
sample estimates:
median of x
        95
```

-가설: H_0 : 모집단의 중위수는 100이다.
H_A : 모집단의 중위수는 100이 아니다.
-결론: 부호검정통계량이 6이고, $p-$값이 0.6072이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 모집단의 중위수는 100이다.

4. 16마리의 실험실 동물에게 생후 12주까지 특별히 선별된 사료만을 먹였다. 12주 뒤에 동물의 몸무게(g)는 다음과 같다고 한다.

63 68 79 65 64 63 65 64 76 74 66 66 67 73 69 76

평균 몸무게가 70g보다 작다고 결론내릴 수 있는지, 유의수준 $\alpha = 0.05$에서 검정하라.

sol)

```
> wilcox.exact(data12_4, mu=70, alternative = "less", conf.level = 0.95)

        Exact Wilcoxon signed rank test

data:  data12_4
V = 48.5, p-value = 0.1632
alternative hypothesis: true mu is less than 70
#동점이 있는 경우에는 wilcox.test()를 쓰면 p-value가 정확하지 않으므로
  wilcox.exact()를 사용해야함.
```

-가설: H_0 : 모평균은 70이다.
H_A : 모평균은 70이 아니다.
-결론: Wilcoxon 검정통계량이 48.5이고, $p-$값이 0.1632이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 모평균은 70이다.

5. 다음은 8명의 환자의 치료 전후 신체계측치이다. 다음 질문에 답하라.

대상	1	2	3	4	5	6	7	8
전	214	362	202	158	403	219	307	331
후	232	276	224	412	562	203	340	313

sol)

1) $t-$검정을 사용하여 치료 전후의 신체계측치의 평균에 차이가 있는지 검정하라.

```
> before<-c(214, 362, 202, 158, 403, 219, 307, 331)
> after<-c(232, 276, 224, 412, 562, 203, 340, 313)
> t.test(before, after, paired = T)

        Paired t-test

data:  before and after
t = -1.1889, df = 7, p-value = 0.2732
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -136.74388   45.24388
sample estimates:
mean of the differences
                 -45.75
```

-가설: H_0 : 치료 전후의 신체계측치의 평균에 차이가 없다.
H_A : 치료 전후의 신체계측치의 평균에 차이가 있다.
-결론: $t-$값이 -1.1889이고, $p-$값이 0.2732로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 치료 전후의 신체계측치의 평균에 차이가 없다.

2) 비모수적 방법을 사용하여 1)을 검정하라.

```
> wilcox.exact(before, after, mu=0, alternative = "two.sided", paired = T)

        Exact Wilcoxon signed rank test

data:  before and after
V = 9.5, p-value = 0.2656
alternative hypothesis: true mu is not equal to 0
```

-가설: H_0 : 치료 전후의 신체계측치의 평균에 차이가 없다.
H_A : 치료 전후의 신체계측치의 평균에 차이가 있다.
-결론: Wilcoxon 검정통계량이 9.5이고, $p-$값이 0.2656이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 치료 전후의 신체계측치의 평균에 차이가 없다.

3) 1)과 2)의 차이는 무엇인가? 또한 어떠한 검정이 더 적절한지 설명하라.
⇒1)은 자료가 정규분포를 따른다는 가정하에 검정하는 모수적 방법이고,
2)는 자료의 분포를 가정하지 않고 검정하는 비모수적 방법이다.
이 경우, 자료가 어떠한 분포를 따르는지 알 수 없기 때문에 비모수적 방법인 부호순위검정이 더 적절하다.

6. 두 병원에서 15명의 환자의 수술 후 체중을 측정한 결과는 다음과 같다.

A	99	85	73	98	83	88	99	80	74	91	80	94	94	98	80
B	78	74	69	79	57	78	79	68	59	91	89	55	60	55	79

sol)

1, 2) 두 모집단의 중위수가 다른지 유의수준 $\alpha = 0.05$에서 검정하고, $p-$값을 구하라.

```
> wilcox_test(val~group, data=data12_6, distribution="exact")

        Exact Wilcoxon-Mann-Whitney Test

data:  val by group (1, 2)
Z = 3.5108, p-value = 0.0001977
alternative hypothesis: true mu is not equal to 0
```

-가설: M_A는 A병원의 모중위수, M_B는 B병원의 모중위수라고 하면, 영가설과 대립가설은 다음과 같다.

$$H_0 : M_A = M_B,\ H_A : M_A \neq M_B$$

-결론: 검정통계량이 3.5108이고, $p-$값이 0.0001977이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, 두 모집단의 중위수는 다르다.

7. 두 모집단의 모중위수가 같은지 확인하기 위하여, 다음과 같이 두 표본을 얻었다. Mann-Whitney 검정을 이용하여, 두 모중위수가 서로 다른지 유의수준 $\alpha = 0.05$에서 검정하라.

무게 (lb)					
그룹1			그룹2		
252	215	240	185	195	220
240	190	302	310	210	295
205	270	312	212	190	202
200	159	126	238	172	268
170	204	268	184	190	220
170	215	215	136	140	311
320	254	183	200	280	164
148	164	287	270	264	206
214	288	210	200	270	170
270	138	225	212	210	190
265	240	258	182	192	
203	217	221	225	126	

sol)

```
> wilcox_test(val~group, data=data12_7, distribution="exact")

        Exact Wilcoxon-Mann-Whitney Test

data:  val by group (1, 2)
Z = 1.1814, p-value = 0.2403
alternative hypothesis: true mu is not equal to 0
```

-가설: M_A는 그룹1의 모중위수, M_B는 그룹2의 모중위수라고 하면, 영가설과 대립가설은 다음과 같다.

$$H_0 : M_A = M_B,\ H_A : M_A \neq M_B$$

-결론: 검정통계량이 1.1814이고, $p-$값이 0.2403이므로 유의수준 0.05하에서 영가설을 기각할 수 없다. 즉, 두 모집단의 중위수는 같다.

8. 어떤 질병을 앓았던 25명의 부검 당시 뇌의 무게는 다음과 같다.

뇌의 무게 (g)				
859	1073	1041	1166	1117
962	1051	1064	1141	1202
973	1001	1016	1168	1255
904	1012	1002	1146	1233
920	1039	1086	1140	1348

모집단이 모평균 1050, 표준편차는 50인 정규분포를 따르는지 적합도 검정을 수행하라.

sol)

```
> ks.test(brain, 'pnorm', mean=1050, sd=50)

        One-sample Kolmogorov-Smirnov test

data:  brain
D = 0.32407, p-value = 0.007747
alternative hypothesis: two-sided
```

-가설: $H_0 : F(x) = F_T(x), -\infty < x < \infty$

H_A : 적어도 하나의 x에 대해 $F(x) \neq F_T(x)$이다.

-결론: 검정통계량이 0.32407이고, 근사적인 $p-$값이 0.007747이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, 모집단이 모평균 1050, 표준편차는 50인 정규분포를 따른다고 볼 수 없다.

9. 세 지역의 수술 비용에 차이가 있는지 조사하기 위하여, 각 지역의 병원을 방문한 외래환자의 수술 비용을 조사하였다. 세 지역의 수술 비용에 차이가 있다고 결론을 내릴 수 있는지 유의수준 $\alpha = 0.05$에서 검정하라.

Ⅰ	Ⅱ	Ⅲ
80.75	58.63	84.21
78.15	72.7	101.76
85.4	64.2	107.74
71.94	62.5	115.3
82.05	63.24	126.15

sol)

```
> kruskal.test(cost~group, data=data12_9)

        Kruskal-Wallis rank sum test

data:  cost by group
Kruskal-Wallis chi-squared = 11.52, df = 2, p-value = 0.003151
```

-가설: H_0 : 모집단의 중위수는 모두 같다.
H_A : 적어도 한 모집단의 중위수가 다르다.
-결론: 검정통계량이 11.52이고, $p-$값이 0.003151이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, 세 지역의 수술 비용에 차이가 있다고 결론을 내릴 수 있다.

10. 9명의 학생들의 3과목 기말 고사 점수를 조사하였다. 세 과목 성적 사이에 차이가 있는지 유의수준 $\alpha = 0.05$에서 검정하라.

번호	시험 과목		
	기초과목(A)	생리학(B)	해부학(C)
1	98	95	77
2	95	71	79
3	76	80	91
4	95	81	84
5	83	77	80
6	99	70	93
7	82	80	87
8	75	72	81
9	88	81	83

sol)

```
> friedman.test(final, group, id)

        Friedman rank sum test

data:  final, group and id
Friedman chi-squared = 8.6667, df = 2, p-value = 0.01312
```

-가설: H_0 : 세 과목의 기말고사 점수는 동일하다.
H_A : 적어도 하나의 기말고사 점수가 다르다.
-결론: 검정통계량이 8.6667이고, 정확한 $p-$값이 0.010이므로 유의수준 0.05하에서 영가설을 기각할 수 있다. 즉, 적어도 하나의 기말고사 점수의 중위수가 다르다고 볼 수 있다.

11. 10개의 지역에서 100명의 어린이의 영구치우식경험자 수와 공설 급소의 불소 농도(ppm)를 비교하여, 각 지역의 순위를 조사하였다.

지역	순위		지역	순위	
	영구치우식경험자 수 순위(X)	불소농도 순위(Y)		영구치우식경험자 수 순위(X)	불소농도 순위(Y)
1	8	1	6	4	7
2	9	3	7	1	10
3	7	4	8	5	6
4	3	9	9	6	5
5	2	8	10	10	2

100명당 영구치우식경험자의 수와 불소농도 사이에 양의 상관성이 있다고 할 수 있는지 유의수준 $\alpha = 0.05$에서 검정하라.

sol)

```
> r_s<-1-6*sum(d^2)/(n*(n^2-1))
> r_s
[1] -0.9515152
```

-가설: H_0 : 영구치우식경험자의 수와 불소농도는 독립이다.
H_A : 영구치우식경험자의 수와 불소농도는 양의 상관성이 없다.
-결론: 검정통계량이 –0.9515152이고, 검정통계량의 임계값은 0.7818이다.
따라서 검정통계량이 임계보다 작기 때문에 영가설을 기각할 수 없다.
즉, 영구치우식경험자의 수와 불소농도는 독립이다.

사용된 R Code

```
##12_3
library(BSDA)
data12_3<-c(75, 90, 85, 110, 115, 95, 132, 74, 82, 104, 88, 124, 110, 76, 98)
SIGN.test(data12_3, md=100, alternative = "two.sided", conf.level = 0.95)

##12_4
library(exactRankTests)
data12_4<-c(63, 68, 79, 65, 64, 63, 65, 64, 76, 74, 66, 66, 67, 73, 69, 76)
wilcox.exact(data12_4, mu=70, alternative = "less", conf.level = 0.95)

##12_5
before<-c(214, 362, 202, 158, 403, 219, 307, 331)
after<-c(232, 276, 224, 412, 562, 203, 340, 313)
t.test(before, after, paired = T)
wilcox.exact(before, after, mu=0, alternative = "two.sided", paired = T)

##12_6
library(coin)
A<-c(99, 85, 73, 98, 83, 88, 99, 80, 74, 91, 80, 94, 94, 98, 80)
B<-c(78, 74, 69, 79, 57, 78, 79, 68, 59, 91, 89, 55, 60, 55, 79)
data12_6<-data.frame(val=c(A, B), group=factor(rep(1:2, c(length(A),length(B)))))
wilcox_test(val~group, data=data12_6, distribution="exact")

##12_7
group1<-c(252, 215, 240, 240, 190, 302, 205, 270, 312, 200, 159, 126, 170, 204, 268, 170, 215, 215, 320, 254, 183, 148,
          164, 287, 214, 288, 210, 270, 138, 225, 265, 240, 258, 203, 217, 221)
group2<-c(185, 195, 220, 310, 210, 295, 212, 190, 202, 238, 172, 268, 184, 190, 220, 136, 140, 311, 200, 280, 164, 270,
          264, 206, 200, 270, 170, 212, 210, 190, 182, 192, 225, 126)
data12_7<-data.frame(val=c(group1, group2), group=factor(rep(1:2, c(length(group1),length(group2)))))
wilcox_test(val~group, data=data12_7, distribution="exact")

##12_8
brain<-c(859, 1073, 1041, 1166, 1117, 962, 1051, 1064, 1141, 1202, 973, 1001, 1016, 1168, 1255, 904, 1012, 1002, 1146,
         1233, 920, 1039, 1086, 1140, 1348)
ks.test(brain, 'pnorm', mean=1050, sd=50)

##12_9
cost<-c(80.75, 78.15, 85.4, 71.94, 82.05, 58.63, 72.7, 64.2, 62.5, 63.24, 84.21, 101.76, 107.74, 115.3, 126.15)
group<-factor(rep(1:3, each=5))
data12_9<-data.frame(cost, group)
kruskal.test(cost~group, data=data12_9)

##12_10
final<-c(98, 95, 77, 95, 71, 79, 76, 80, 91, 95, 81, 84, 83, 77, 80, 99, 70, 93, 82, 80, 87, 75, 72, 81, 88, 81, 83)
group<-factor(rep(c("A", "B", "C"), 9))
id<-factor(rep(1:9, each=3))
friedman.test(final, group, id)

##12_11
X<-c(8, 9, 7, 3, 2, 4, 1, 5, 6, 10); Y<-c(1, 3, 4, 9, 8, 7, 10, 6, 5, 2); n<-10
d<-(X-Y); r_s<-1-6*sum(d^2)/(n*(n^2-1)); r_s
```