부록 02

연습문제 해답

부 록
02

연습문제 해답

1장 연습문제 해답

01 (1) ③, ④
(2) 아니오
(3) 새로운 방법이 적용되기 이전의 성적을 알지 못하므로 적용 이전과 이후를 비교할 수 없다. 따라서 새로운 수업 방법이 성공적이라고 말할 수 없다.

02 (1) ④, ⑤
(2) 새로운 진통 주사제를 맞는 것
(3) 허리 통증 부위에 새로운 진통 주사제를 맞은 집단
(4) 아니오. 효과가 있다/없다고 말하기 어렵다.
(5) 이중맹검, 플라세보가 없으므로 실험자 또는 실험대상자에서 편이가 발생할 수 있다.

03 관측연구

04 알 수 없다. 백신 접종하지 않은 고위험군의 사망률과 백신 접종한 고위험군의 사망률을 비교해야 판단할 수 있다.

05 (1) 모집단: 서울 시내 직장인들
(2) 표본: 매일 7시~8시에 이대역에 내리는 직장인들
(3) 표본이 모집단을 대표하지 않는다. 이대역만 조사할 경우 시청역, 역삼역 등 서울 내 다양한 지역의 직장인들을 포함하지 못하게 된다. 그리고 매일 7시~8시 이외에 새벽 6시나 오후 9시에 출근하는 사람들도 있기에 다른 시간대에 출근하는 직장인들을 포함하지 못하게 된다.

06 (1) 모집단: 한국대학에 이번 학기 등록한 학생들

(2) 영희의 방법이 가장 적절하다.

(3) 철수의 방법으로 뽑을 경우, 선택편이가 발생할 수 있다. 점심시간에 수업이 있는 학생들이 빠질 수 있고 또는 휴학생이지만 학교에 동기를 만나러 온 경우에 표본으로 선택될 수 있다. 민수의 경우 응답을 거부하지 않는 학생들만 표본집단이 되므로 무응답편이가 발생할 수 있다.

07 (1) $$\frac{(A\text{회사 남자 합격자 수})+(B\text{회사 남자 합격자 수})}{(\text{전체 남자 지원자 수})}$$

$$=\frac{1000\times 0.6+1000\times 0.3}{2000}=0.45$$

(2) $$\frac{(A\text{회사 여자 합격자 수})+(B\text{회사 여자 합격자 수})}{(\text{전체 여자 지원자 수})}$$

$$=\frac{100\times 0.6+1000\times 0.3}{1100}=0.32727$$

(3) 남녀 차별이 있다고 보기 어렵다. 2개 회사를 합하여 계산한 합격률에서는 남성 합격률이 높지만, 개별 회사의 합격률을 살펴보면 남녀의 합격률이 동일하다(심슨의 패러독스). 이는 남성은 회사 A와 B에 50%씩 지원한 반면, 여성은 회사 A에는 10% 미만이 지원하고 회사 B에는 90% 이상이 지원했는데, 회사 B는 합격률이 낮기 때문이다. 즉, 여성의 합격률이 낮은 것은 여성 대부분이 합격률이 낮은 회사에 지원했기 때문이다.

08 오전 9시~ 오후 6시에는 직장에 출근하는 사람들을 집에서 만나기 어려우므로, 선택편이와 무응답편이가 발생할 수 있다.

09 대표성이 있다.

10 (1) 성별과 나이에 따라 건강상태가 다르다는 합성요인을 제어하기 위해서이다.

(2) 아니오.

(3) 과거 얼마나 오랫동안 흡연을 하였는지 조사되지 않았다. 과거 오랜 기간 흡연자는 이미 흡연의 폐해가 누적되어 있어서 현재는 금연이라도 건강상태가 이미 나빠졌을 수 있다. 또는 흡연자가 건강상 문제가 발생하여 흡연을 중단했을 수도 있다.

11 사용하기 어렵다. 비교, 제어화, 랜덤화 등을 만족하는 임상시험을 설계하여 효과를 확인해야 한다. 10명으로는 치료의 안정성을 확보하기 어렵고 부작용에 대한 연구도 이루어지지 않았다.

12 사용하면 안 된다. 원숭이와 인간은 다르므로 원숭이에 대한 실험 결과를 인간에게 적용하기 어렵다.

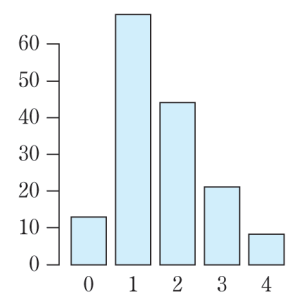
13 ②, ③, ④, ⑤

14 효과가 있다고 보기 어렵다. 수술법을 실시한 환자와 실시하지 않은 환자의 예후를 비교해보아야 한다.

2장 연습문제 해답

01 (1) 막대그림을 그리는 R 코드는 다음과 같다.

```
> Class <- c(0, 1, 2, 3, 4)
> Freq <- c(13, 68, 44, 21, 8)
>
> df <- data.frame(Class, Freq)
> df
  Class Freq
1     0   13
2     1   68
3     2   44
4     3   21
5     4    8
>
> barplot(df$Freq, names.arg=df$Class)
```



(2) 아니오

(3) 도수가 가장 많은 범주는 1(희미한 얼룩)이다.

02 (1) 상대도수를 구하는 R 코드는 다음과 같다. (직접 계산하여도 된다.)

```
> year1990 <- c(26, 47, 17)
> prop.table(year1990)
[1] 0.2888889 0.5222222 0.1888889
>
> year2000 <- c(25, 67, 18)
> prop.table(year2000)
[1] 0.2272727 0.6090909 0.1636364
```

상대도수분포표로 나타내면 다음과 같다.

	1990년	2000년
제조업	0.2888889	0.2272727
서비스업	0.5222222	0.6090909
사무직	0.1888889	0.1636364
합계	1	1

(2) 제조업, 사무직의 비율은 감소하고 서비스업의 비율은 증가했다.

03 (1) 도수분포표를 작성하는 R 코드는 다음과 같다.

```
> titan <- scan("ch2_titan_c_s.txt")
Read 891 items
> df <- table(titan)
> df
titan
  1   2   3   4   5   6
136  87 119  80  97 372
```

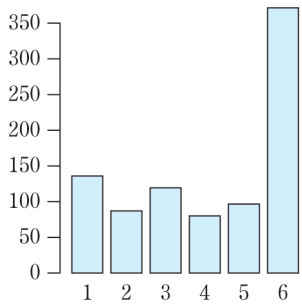
도수분포표로 나타내면 다음과 같다.

범주	도수
1	136
2	87
3	119
4	80
5	97
6	372

(2) 막대그림을 그리는 R 코드는 다음과 같다.

```
〉 barplot(df)
```

막대그림은 다음과 같다.



(3) 6번 범주의 도수가 가장 높다.

(4) class=1인 사람들 중 생존한 사람들의 비율은
1, 4번 범주 중에서 1번 범주의 비율과 같다.
이를 구하는 R 코드는 다음과 같다.

```
〉 136/(136+80)
[1] 0.6296296
```

따라서 class=1인 사람들 중 생존한 사람들의 비율은 약 0.6296296이다.

04 (1) $(\text{평균}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6+8+10+6+10}{5} = \frac{40}{5} = 8$

(2) 자료를 순서대로 나타내면 6, 6, 8, 10, 10이므로 중앙값은 8이다.

(3) 편차의 합은 항상 0이다.

(4) 편차, 편차의 제곱을 구하면 다음과 같다.

x	6	6	8	10	10
$x-\overline{x}$	-2	-2	0	2	2
$(x-\overline{x})^2$	4	4	0	4	4

$(\text{분산}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{4+4+0+4+4}{5} = \frac{16}{5} = 3.2$

$(\text{표준편차}) = \sqrt{\text{분산}} = \sqrt{3.2}$

05 (1) $(\text{평균}) = \dfrac{1+3+5+1+5}{5} = \dfrac{15}{5} = 3$

각각의 자료에서 5를 빼니, 평균도 5만큼 작아졌다.

(2) 자료를 순서대로 나타내면 1, 1, 3, 5, 5이므로 중앙값은 3이다.

각각의 자료에서 5를 빼니, 중앙값도 5만큼 작아졌다.

(3) 편차, 편차의 제곱을 구하면 다음과 같다.

x	1	1	3	5	5
$x-\overline{x}$	-2	-2	0	2	2
$(x-\overline{x})^2$	4	4	0	4	4

$$(\text{분산}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{4+4+0+4+4}{5} = \frac{16}{5} = 3.2$$

각각의 자료에서 5를 빼도, 분산은 같다.

(4) $(\text{표준편차}) = \sqrt{\text{분산}} = \sqrt{3.2}$

각각의 자료에서 5를 빼도, 표준편차는 같다.

06 (1) R 코드는 다음과 같다.

```
> chol <- scan("ch2_cholesterol.txt")
Read 40 items
> summary(chol)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
  30.00  53.38  57.75  58.67  66.00  79.50
> b <- c(20, 30, 40, 50, 60, 70, 80)
> chol_class <- cut(chol, b, right=TRUE)
> t <- table(chol_class)
> t
chol_class
(20,30] (30,40] (40,50] (50,60] (60,70] (70,80]
    2       0       6      16       7       9
> prop.table(t)
chol_class
(20,30] (30,40] (40,50] (50,60] (60,70] (70,80]
 0.050   0.000   0.150   0.400   0.175   0.225
```

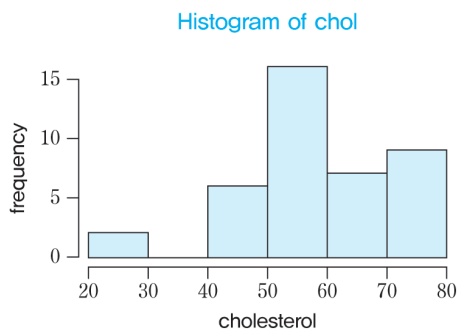
도수분포표로 나타내면 다음과 같다.

구간	도수	상대도수
(20,30]	2	0.050
(30,40]	0	0.000
(40,50]	6	0.150
(50,60]	16	0.400
(60,70]	7	0.175
(70,80]	9	0.225

(2) 히스토그램을 그리는 R 코드는 다음과 같다.

```
> hist(chol, b, xlab="cholesterol", ylab="frequency")
```

히스토그램은 다음과 같다.



(30, 40]에는 데이터가 존재하지 않으며 대부분의 데이터는 40보다 큰 값을 갖는다.

(50, 60]에 가장 많은 데이터가 존재한다.

(3) 줄기－잎 그림을 그리는 R 코드는 다음과 같다.

```
> stem(chol)
The decimal point is 1 digit(s) to the right of the |
   3 | 00
   3 |
   4 | 111
   4 | 999
   5 | 234444
   5 | 55668888
   6 | 0024444
   6 | 59
   7 | 334
   7 | 66666
   8 | 0
```

3 | 00

4 | 111999

5 | 23444455668888

6 | 002444459

7 | 33466666

8 | 0

(4) 평균과 표준편차를 구하는 R 코드는 다음과 같다.

```
> mean(chol)
[1] 58.675
> sd(chol)
[1] 12.37944
```

평균: 58.675

표준편차: 12.37944

(5) 중앙값과 사분위 범위를 구하는 R 코드는 다음과 같다.

```
> median(chol)
[1] 57.75
> q <- quantile(chol)
> q[4]-q[2]
   75%
12.625
```

중앙값: 57.75

사분위 범위: 12.625

(6) 하위 25% 값을 구하는 R 코드는 다음과 같다.

```
> q <- quantile(chol)
> q[2]
   25%
53.375
```

하위 25% 값: 53.375

07 (1) 평균과 표준편차를 구하는 R 코드는 다음과 같다.

```
> bus <- scan("ch2_bus.txt")
Read 25 items
>
> mean(bus)
[1] 15.36
> sd(bus)
[1] 5.536846
```

평균: 15.36

표준편차: 5.536846

(2) 중앙값과 사분위범위를 구하는 R 코드는 다음과 같다.

```
> median(bus)
[1] 15
> q <- quantile(bus)
> q[4]-q[2]
75%
  8
```

중앙값: 15

사분위범위: 8

(3) 평균과 표준편차를 구하는 R 코드는 다음과 같다.

```
> bus2 <- bus*1000
> mean(bus2)
[1] 15360
> sd(bus2)
[1] 5536.846
```

평균: 15360

표준편차: 5536.846

(4) 중앙값과 사분위범위를 구하는 R 코드는 다음과 같다.

```
> median(bus2)
[1] 15000
> q <- quantile(bus2)
> q[4]-q[2]
 75%
8000
```

중앙값: 15000

사분위범위: 8000

08 오른쪽 꼬리가 긴(positively skewed) 그래프이므로 평균이 중앙값보다 크다.

09 (1) (평균) $= \dfrac{3+3+3+4+4+4+4+5+5+5}{10} = \dfrac{40}{10} = 4$

자료의 개수가 10개이므로 중앙값은 자료를 순서대로 나열했을 때, 5번째 값과 6번째 값의 평균이다.

(중앙값) $\dfrac{4+4}{2} = 4$

(2) (평균) $= \dfrac{5+5+5+6+6+6+6+7+7+7}{10} = \dfrac{60}{10} = 6$

자료의 개수가 10개이므로 중앙값은 자료를 순서대로 나열했을 때, 5번째 값과 6번째 값의 평균이다.

(중앙값) $\dfrac{6+6}{2} = 6$

(3) 각각의 자료에서 2를 더하니, 평균도 2만큼 커졌다.

(2의 평균) = (1의 평균) + 2

(4) 각각의 자료에서 2를 더하니, 중앙값도 2만큼 커졌다.

(2의 중앙값) = (1의 중앙값) + 2

10 ①

11 ①, ②

12 ④

13 32

3장 연습문제 해답

01 (1) $\dfrac{(\text{십의 자리 홀수, 일의 자리 짝수})+(\text{십의 자리 짝수, 일의 자리 짝수})}{(\text{2자리 숫자를 만드는 경우의 수})}$

$= \dfrac{2\times 2+2\times 1}{4\times 3} = \dfrac{1}{2}$

(2) 카드로 만들 수 있는 숫자 중에서 22와 30 사이인 숫자는 23, 24로 2가지이다.

$\therefore \dfrac{2}{4\times 3} = \dfrac{1}{6}$

02 신생아의 유전자형은 AA이므로 친부와 친모 모두에게서 A를 물려받았다. 따라서 친부의 유전자형에는 적어도 한 개의 A가 포함되어야 하고, 이는 AA, AO, AB 총 3가지가 가능하다.

$\therefore P(f=AA)P(b=AA|f=AA)+P(f=AO)P(b=AA|f=AO)$
$+P(f=AB)P(b=AA|f=AB)$
$=0.5\times 1+0.25\times 0.5+0.11\times 0.5=0.68$

여기에서 f와 b는 각각 아버지와 신생아의 유전자형을 나타낸다.

03 (1) $P(\text{여자}) = \dfrac{140+70}{210+80+140+70} = \dfrac{210}{500} = 0.42$

(2) $P(\text{연금가입 } yes) = \dfrac{210+140}{210+80+140+70} = \dfrac{350}{500} = 0.7$

(3) $P(\text{여자}|\text{연금가입 } no) = \dfrac{70}{80+70} = \dfrac{70}{150} \approx 0.4667$

04 (1) 배반 아님 $\because P(\text{남자}\cap\text{연금가입 } yes)\neq 0$

(2) 독립 아님

$\because P(\text{여자})=0.42,\ P(\text{여자}|\text{연금가입 } yes) = \dfrac{140}{210+140} = \dfrac{140}{350} = 0.4,$

$P(\text{여자}) \neq P(\text{여자}|\text{연금가입 } yes)$

05 (1) $P(\text{60분 미만 지연}) = \dfrac{410+400+380+360}{410+400+80+380+360+150} = \dfrac{1550}{1780} \approx 0.8708$

(2) $P(\text{항공사 } B) = \dfrac{380+360+150}{410+400+80+380+360+150} = \dfrac{890}{1780} = 0.5$

(3) $P(30\sim 60\text{분 지연}|\text{항공사}A)=\dfrac{400}{410+400+80}=\dfrac{400}{890}\approx 0.4494$

(4) $P(\text{항공사}A|30\sim 60\text{분 지연})=\dfrac{400}{400+360}=\dfrac{400}{760}\approx 0.5263$

06 (1) 배반 아님 $\because P(\text{항공사}A\cap 60\text{분}\sim \text{지연})\neq 0$

(2) 독립 아님

$\because P(\text{항공사}A)=0.5,\ P(\text{항공사}A|60\text{분}\sim \text{지연})=\dfrac{80}{80+150}=0.3478,$

$P(\text{항공사}A)\neq P(\text{항공사}A|60\text{분}\sim \text{지연})$

07 (1) $P(A\cap B^c)=P(B^c)P(A|B^c)=(1-\dfrac{1}{3})\times\dfrac{3}{7}=\dfrac{2}{7}$

(2) 분할에 의해, $P(A)=P(A\cap B)+P(A\cap B^c)=\dfrac{2}{9}+\dfrac{2}{7}=\dfrac{32}{63}$

(3) $P(B^c|A)=\dfrac{P(A\cap B^c)}{P(A)}=\dfrac{2/7}{32/63}=\dfrac{9}{16}$

08 (1) $P(A)=P(A\cap B)+P(A\cap B^c)=0+P(A\cap B^c)=P(A\cap B^c)$

(2) $P(A\cap B)=P(A)P(B)=0.3$ 이면 A와 B는 독립

(3) 서로 배반이면 P(A∪B)=P(A)+P(B)=1.1이어야 하는데 확률은 1보다 클 수 없으므로 배반일 수 없다.

(4) $P(A\cap B)=P(A)+P(B)-P(A\cap B)=0.5+0.5-0.8=0.2$

$\neq P(A)P(B)=0.25$

∴ A와 B가 서로 독립 아님

4장 연습문제 해답

01 (1) $\dfrac{({}_{10}C_2\times{}_{10}C_2)+({}_{10}C_3\times{}_{10}C_1)+({}_{10}C_4\times{}_{10}C_0)}{{}_{20}C_4}=0.709$

(2) $\dfrac{({}_{5}C_2\times{}_{10}C_2)+({}_{5}C_3\times{}_{10}C_1)+({}_{5}C_4\times{}_{10}C_0)}{{}_{15}C_4}=0.4066$

(3) $\dfrac{({}_3C_2 \times {}_9C_2) + ({}_3C_3 \times {}_9C_1)}{{}_{12}C_4} = 0.2364$

(4) $\dfrac{{}_{10}C_2 \times {}_{10}C_2}{{}_{20}C_4} \times \dfrac{{}_5C_2 \times {}_{10}C_2}{{}_{15}C_4} \times \dfrac{{}_3C_2 \times {}_9C_2}{{}_{12}C_4} = 0.0301$

02 (1) $\mathrm{P}(2 \le X \le 3) = \mathrm{P}(X=2) + \mathrm{P}(X=3) = 0.35 + 0.30 = 0.65$

(2) $\mathrm{E}(X) = 0 \times 0.15 + 1 \times 0.2 + 2 \times 0.35 + 3 \times 0.3 = 1.8$

$\mathrm{E}(X^2) = 0^2 \times 0.15 + 1^2 \times 0.2 + 2^2 \times 0.35 + 3^2 \times 0.3 = 4.3$

$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 4.3 - 1.8^2 = 1.06$

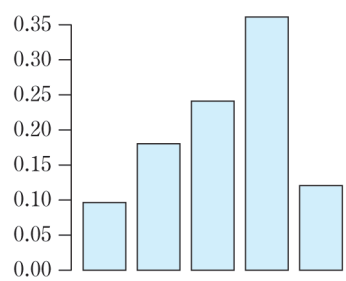
기대치는 1.8, 분산은 1.06이다.

03 (1) X의 확률질량함수는

$$f(x) = \begin{cases} 8/83 & x=1 \\ 15/83 & x=2 \\ 20/83 & x=3 \\ 30/83 & x=4 \\ 10/83 & x=5 \end{cases}$$

막대그림을 구하는 R 코드는 다음과 같다.

```
> y=c(8,15,20,30,10)
> n=sum(y)
> fx=y/n
> barplot(fx)
```



(2) (a) $P(X=3) = \dfrac{20}{83}$

(b) $P(X \ge 3) = \dfrac{20}{83} + \dfrac{30}{83} + \dfrac{10}{83} = \dfrac{60}{83}$

(c) $P(X < 4) = \dfrac{8}{83} + \dfrac{15}{83} + \dfrac{20}{83} = \dfrac{43}{83}$

(3) $\mathrm{E}(X)=1\times\frac{8}{83}+2\times\frac{15}{83}+3\times\frac{20}{83}+4\times\frac{30}{83}+5\times\frac{10}{83}=\frac{268}{83}(=3.229)$

$\mathrm{E}(X^2)=1^2\times\frac{8}{83}+2^2\times\frac{15}{83}+3^2\times\frac{20}{83}+4^2\times\frac{30}{83}+5^2\times\frac{10}{83}=\frac{978}{83}$

$\mathrm{Var}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2=\frac{978}{83}-\left(\frac{268}{83}\right)^2=1.357$

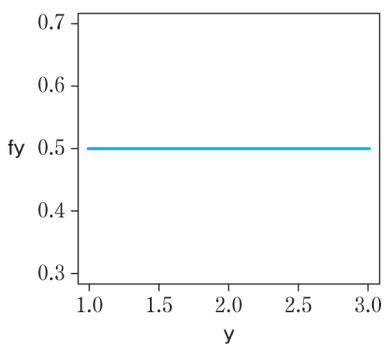
기대치는 3.229, 분산은 1.357이다.

04 (1) $0<X<1$이므로 $1<Y<3$이고, Y의 확률밀도함수는

$$f(y)=\begin{cases}\frac{1}{2} & 1<y<3\\ 0 & y\le 1,\ y\ge 3\end{cases}$$

$f(y)$를 그림으로 그리는 R 코드는 다음과 같다.

```
> x=seq(0, 1, by=0.001)
> y=2*x+1
> fy=dunif(y, 1, 3)
> plot(y, fy)
```



(2) $\mathrm{E}(Y)=\int_1^3\frac{1}{2}y\,dy=\left[\frac{1}{4}y^2\right]_1^3=\frac{1}{4}(3^2-1^2)=2$

$\mathrm{E}(Y^2)=\int_1^3\frac{1}{2}y^2\,dy=\left[\frac{1}{6}y^3\right]_1^3=\frac{1}{6}(3^3-1^3)=\frac{13}{3}$

$\mathrm{Var}(Y)=\mathrm{E}(Y^2)-\mathrm{E}(Y)^2=\frac{13}{3}-2^2=\frac{1}{3}=0.333$

Y의 기대치는 2, 분산은 0.333이다.

05 (1) 각 위치에서 주어진 숫자와 일치하지 않을 확률: 0.9

$X=0$일 때 확률 $\mathrm{P}(X=0)=0.9^3=0.729$

$X=1$일 때 확률 $\mathrm{P}(X=1)=(0.9^2\times0.1)\times{}_3\mathrm{C}_1=0.243$

$X=2$일 때 확률 $\mathrm{P}(X=2)=(0.9^1\times0.1^2)\times{}_3\mathrm{C}_2=0.027$

$X=3$일 때 확률 $\mathrm{P}(X=3)=0.1^3=0.001$

$$f(x)=\begin{cases}0.729 & x=0\\0.243 & x=1\\0.027 & x=2\\0.001 & x=3\end{cases}$$

(2) $\mathrm{E}(X)=0\times\mathrm{P}(X=0)+1\times\mathrm{P}(X=1)+2\times\mathrm{P}(X=2)+3\times\mathrm{P}(X=3)=0.3$

(3) $\mathrm{Var}(X)=(0-0.3)^2\times\mathrm{P}(X=0)+(1-0.3)^2\times\mathrm{P}(X=1)+(2-0.3)^2$
$\times\mathrm{P}(X=2)+(3-0.3)^2\times\mathrm{P}(X=3)=0.27$

$\mathrm{SD}(X)=\sqrt{0.27}=0.5196$

(4) 상금을 변수 Y라 하면 $Y=10000X$

$\mathrm{E}(Y)=10000\times\mathrm{E}(X)=10000\times0.3=3000$(원)

(5) $\mathrm{Var}(Y)=10000^2\times\mathrm{Var}(X)=27000000$

$\mathrm{SD}(Y)=10000\times\mathrm{SD}(X)=5196.15$

06 $\mathrm{E}(X)=1\times0.6+2\times0.3+3\times0.08+4\times0.02=1.52$

$\mathrm{E}(X^2)=1^2\times0.6+2^2\times0.3+3^2\times0.08+4^2\times0.02=2.84$

$\mathrm{Var}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2=2.84-1.52^2=0.5296$

X의 기대치는 1.52, 분산은 0.5296이다.

5장 연습문제 해답

01 (1) $X\sim B(3,\,0.2)$

(2) $P(X=0)=\binom{3}{0}(0.2)^0(0.8)^3=0.512$

$P(X\geq1)=1-P(X=0)=1-0.512=0.488$

02 $X\sim Ber(p)$

$E(X)=1\times p+0\times(1-p)=p$

03 $X=$50세 이전에 사망하는 남자 수 $X \sim B(20, 0.05)$

(1) $20 \times 0.05 = 1$명

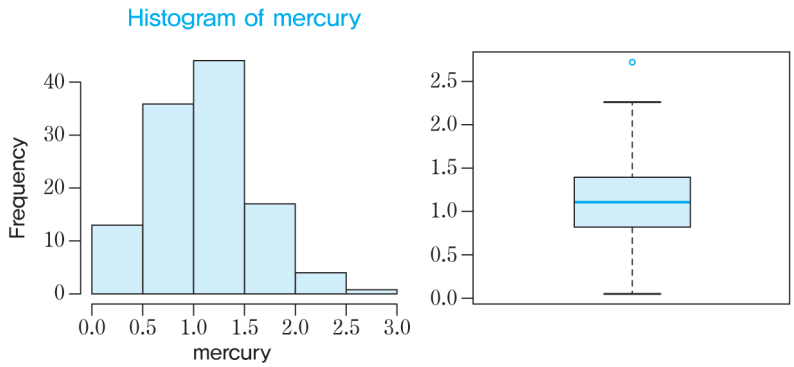
(2) $P(X=0) = \binom{20}{0}(0.05)^0(0.95)^{20} = 0.3585$

04 (1)

```
mercury=scan("ch2_mercury.txt")
〉 mean(mercury); var(mercury); sd(mercury)
[1] 1.100435
[1] 0.2505024
[1] 0.5005022
```

(2)

```
〉 hist(mercury)
〉 boxplot(mercury)
```



(3) 대략 대칭이고 종 모양의 분포를 가지므로 정규분포를 따른다고 가정할 수 있다.

(4) $X \sim N(1.1004,\ 0.5005^2)$

$$P(X>1) = P\left(Z > \frac{(1-1.1004)}{0.5005}\right)$$

$$= P(Z > -0.2006) = 0.5795 \qquad \therefore 57.95\%$$

05 미우라 부족 여성의 키(X) : $X \sim N(158, 4^2)$

(1) $P(155 < X < 164) = P\left(\frac{(155-158)}{4} < Z < \frac{(164-158)}{4}\right)$

$$= P(-0.75 < Z < 1.5)$$

$$= P(Z<1.5) - P(Z<-0.75) = 0.9332 - 0.2266 = 0.7066$$

(70.66%)

(2) $P(X>170)=P\left(Z>\dfrac{(170-158)}{4}\right)=P(Z>3.0)=1-P(Z<3.0)$

$=1-0.9987=0.0013$ (0.13%)

(3) $P(Z<q)=0.9$인 q

Z의 상위 10%는 1.28이므로

X의 상위 10%는 $1.28\times4+158=163.12$ (163.12cm)

06 (1) $X\sim B(100,\ 4/37)$

(2) $E(X)=100\times4/37=10.8108$,

$Var(X)=100\times(4/37)\times(33/37)=9.6421=3.1052^2$

(3) $X\approx N(10.8108,\ 3.1052^2)$

$P(X\geq15)=1-P(X<14.5)$

$=1-P\left(Z<\dfrac{(14.5-10.8108)}{3.1052}\right)=1-P(Z<1.1881)$

$=1-0.8826=0.1174$

(4) $P(X\leq5)=P(X<5.5)=P\left(Z<\dfrac{(5.5-10.8108)}{3.1052}\right)$

$=P(Z<-1.7103)=0.0436$

07 한 컵에 따라주는 콜라 양(X) : $X\sim N(200,\ 5^2)$

(1) $P(195<X<210)=P\left(\dfrac{(195-200)}{5}<Z<\dfrac{(210-200)}{5}\right)$

$=P(-1<Z<2)$

$P(Z<2)-P(Z<-1)=0.9773-0.1587=0.8186$

(2) $P(200<X<205)=P\left(\dfrac{(200-200)}{5}<Z<\dfrac{(205-200)}{5}\right)=P(0<Z<1)$

$P(Z<1)-P(Z<0)=0.8413-0.5=0.3413$

(3) $P(X\leq190)=P\left(Z<\dfrac{(190-200)}{5}\right)=P(Z<-2)=0.0228$

(4) $P(X>215)=P\left(Z>\dfrac{(215-200)}{5}\right)=P(Z>3)=1-P(Z<3)=0.00135$

(상위 약 0.135%)

08 (1) $p = 0.1$

(2) $X \sim B(60, 0.1)$

$E(X) = 60 \times 0.1 = 6$

$Var(X) = 60 \times 0.1 \times 0.9 = 5.4$

(3) n=60, p=0.1, np=6>=5, n(1−p)=54>=5

$P(X \geq 10) = P(X > 10.5) = P(Z > (10.5 - 6)/\sqrt{5.4}) = P(Z > 1.936)$

$= 0.026$

09 $X \sim \mathrm{B}(30, 0.2)$에서 $np = 30 \times 0.2 = 6$,

$np(1-p) = 30 \times 0.2 \times 0.8 = 4.8 = (2.19)^2$이므로

$X \sim \mathrm{N}(6, 2.19^2)$

$$\mathrm{P}(X \geq 8) = 1 - \mathrm{P}(X \leq 7) = 1 - \mathrm{P}(X < 7.5) = 1 - \mathrm{P}\left(Z < \frac{7.5 - 6}{2.19}\right)$$

$$= 1 - \mathrm{P}(Z < 0.6849)$$

$$= 0.2467$$

10 (1) X를 출퇴근 없는 것이 재택근무의 가장 큰 장점이라고 말하는 사람의 수라 하자.

$X \sim \mathrm{B}(100, 0.30)$에서

$np = 100 \times 0.3 = 30$, $np(1-p) = 100 \times 0.3 \times 0.7 = 21 = (4.58)^2$이므로

$X \sim \mathrm{N}(30, 4.58^2)$

$$\mathrm{P}(28 \leq X \leq 30) = \mathrm{P}(27.5 < X < 30.5)$$

$$= \mathrm{P}\left(\frac{27.5 - 30}{4.58} < Z < \frac{30.5 - 30}{4.58}\right)$$

$$= \mathrm{P}(-0.5459 < Z < 0.1092) = 0.2509$$

(2) $p = 0.3$, $p(1-p)/n = 0.3 \times 0.7/100 = 0.0021 = 0.0458^2$이므로

$\hat{p} \sim \mathrm{N}(0.3, 0.0458^2)$

$$\mathrm{P}(0.28 < \hat{p} < 0.3) = \mathrm{P}\left(\frac{0.28 - 0.3}{0.0458} < Z < \frac{0.3 - 0.3}{0.0458}\right)$$

$$= \mathrm{P}(-0.4367 < Z < 0) = 0.1688$$

(연속성 보정 안 함)

6장 연습문제 해답

01 $\mu = 1,\ \sigma^2 = 1$

변수 X를 25페이지 소설에서의 총 오타 수라고 하면

$\mathrm{E}(X) = n\mu = 25 \times 1 = 25,\ \mathrm{Var}(X) = n\sigma^2 = 25 \times 1 = 25 = 5^2$

X의 근사적 분포는 $\mathrm{N}(25,\ 5^2)$이므로

$$\mathrm{P}(X \geq 20) = \mathrm{P}(X > 19.5) = P\left(Z > \frac{19.5 - 25}{5}\right) = P(Z > 1.1)$$

$$= 1 - P(Z < -1.1) = 0.8643$$

02 (1) $\overline{X} \sim \mathrm{N}(\mu,\ 100^2/25) = \mathrm{N}(\mu,\ 20^2)$

$$\mathrm{P}(|\overline{X} - \mu| > 20) = \mathrm{P}\left(|Z| > \frac{20}{20}\right) = \mathrm{P}(Z < -1) + \mathrm{P}(Z > 1) = 2\mathrm{P}(Z > 1)$$

$$= 2 \times 0.1588 = 0.3173$$

(2) $\overline{X} \sim t_{24}(\mu,\ 100^2/25) = t_{24}(\mu, 20^2)$

$$\mathrm{P}(|\overline{X} - \mu| > 20) = \mathrm{P}\left(|T| > \frac{20}{20}\right) = \mathrm{P}(T < -1) + \mathrm{P}(T > 1) = 2\mathrm{P}(T > 1)$$

이 확률을 구하는 R 코드는 다음과 같다.

```
> 2*(1-pt(1, df=24))
[1] 0.3272869
```

03 $\overline{X} \sim \mathrm{N}(\mu,\ 1^2/4) = \mathrm{N}(\mu,\ 0.5^2)$

$$\mathrm{P}(|\overline{X} - \mu| < 1) = \mathrm{P}\left(|Z| < \frac{1}{0.5}\right) = \mathrm{P}(-2 < Z < 2) = \mathrm{P}(Z < 2) - \mathrm{P}(Z < -2)$$

$$= 0.9545$$

04 $\overline{X} \sim \mathrm{N}(\mu,\ 8/100) = \mathrm{N}(\mu,\ 0.2828^2)$

$$\mathrm{P}(|\overline{X} - \mu| > 1) = \mathrm{P}\left(|Z| > \frac{1}{0.2828}\right) = \mathrm{P}(|Z| > 3.5361) = 2\mathrm{P}(Z > 3.5361)$$

$$= 2 \times 0.0002 = 0.0004$$

05 변수 X를 총 승리 횟수라고 하자.

$X \sim \mathrm{B}\left(25,\ \frac{12}{37}\right)$이므로

$\mathrm{E}(X) = 25 \times \left(\frac{12}{37}\right) = 8.11,\ \mathrm{Var}(X) = 25 \times \left(\frac{12}{37}\right) \times \left(\frac{25}{37}\right) = 2.34^2$

$np = 25 \times \left(\frac{12}{37}\right) \geq 5,\ n(1-p) = 25 \times \left(\frac{25}{37}\right) \geq 5$이므로 $X \sim \mathrm{N}(8.11,\ 2.34^2)$

$\mathrm{P}(X \geq 10) = \mathrm{P}(X > 9.5) = \mathrm{P}\left(Z > \frac{9.5 - 8.11}{2.34}\right) = 1 - \mathrm{P}(Z < 0.594) = 0.2763$

06 $p = 0.63,\ n = 50,\ np \geq 5,\ n(1-p) \geq 5$이므로

$\hat{p} \sim \mathrm{N}(p,\ p(1-p)/n) = \mathrm{N}(0.63,\ 0.63 \times 0.37/50) = \mathrm{N}(0.63,\ 0.0683^2)$

$\mathrm{P}(\hat{p} > 0.7) = \mathrm{P}(Z > (0.7 - 0.63)/0.0683) = \mathrm{P}(Z > 1.0249) = 0.1527$

07 (1) $\mathrm{E}(X) = 1 \times 0.6 + 2 \times 0.3 + 3 \times 0.08 + 4 \times 0.02 = 1.52$

$\mathrm{E}(X^2) = 1^2 \times 0.6 + 2^2 \times 0.3 + 3^2 \times 0.08 + 4^2 \times 0.02 = 2.84$

$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 2.84 - 1.52^2 = 0.5296$

X의 기대치는 1.52, 분산은 0.5296이다.

(2) $\mu = \mathrm{E}(X) = 1.52,\ \sigma^2 = \mathrm{Var}(X) = 0.5296$이라 하면

$\mathrm{E}(\overline{X}) = \mu = 1.52$

$\mathrm{Var}(\overline{X}) = \sigma^2/n = 0.5296/60 = 0.0088$

(3) $n \geq 30$이므로 정규근사에 의하여 $\overline{X} \sim \mathrm{N}(1.52,\ 0.0088)$

$= \mathrm{N}(1.52,\ 0.0939^2)$이므로

$\mathrm{P}(\overline{X} \geq 1.5) = \mathrm{P}\left(Z \geq \frac{1.5 - 1.52}{0.0939}\right) = \mathrm{P}(Z \geq -0.21) = 1 - \mathrm{P}(Z \leq -0.21)$

$= 1 - 0.4168 = 0.5832$

08 (1) 100번의 게임에서 이기는 횟수를 X라 하면 $X \sim \mathrm{B}\left(100,\ \frac{18}{37}\right)$

$n = 100,\ p = \frac{18}{37}$에서 $np \geq 5,\ n(1-p) \geq 5$이므로 정규근사에 의하여

$X \sim \mathrm{N}(np,\ np(1-p)) = \mathrm{N}(48.65,\ 24.9817) \approx \mathrm{N}(48.65,\ 5^2)$

$\mathrm{P}(X \geq 60) = \mathrm{P}\left(Z \geq \frac{60 - 48.65}{5}\right) = \mathrm{P}(Z \geq 2.27) = 1 - \mathrm{P}(Z \leq 2.27)$

$= 1 - 0.9884 = 0.0116$

(2) $\mathrm{P}(50 \le X \le 55) = \mathrm{P}\left(\frac{50-48.65}{5} \le Z \le \frac{55-48.65}{5}\right)$

$= \mathrm{P}(0.27 \le Z \le 1.27)$

$= \mathrm{P}(Z \le 1.27) - \mathrm{P}(Z \le 0.27) = 0.8980 - 0.6064$

$= 0.2916$

7장 연습문제 해답

01 $n = 50,\ \overline{X} = 38,\ s = 10$

구매자의 평균나이에 대한 99% 신뢰구간: $\overline{X} \pm 2.58 \times \frac{s}{\sqrt{n}} = (34.35,\ 41.65)$

02 $n = 100,\ \overline{X} = 7.5,\ s = 1$

퓨즈가 끊어지는 순간 전류의 평균에 대한 95% 신뢰구간:

$\overline{X} \pm 1.96 \times \frac{s}{\sqrt{n}} = (7.304,\ 7.696)$

03

```
> x=c(60, 56, 80, 55, 74, 68, 75, 73, 68, 90)
> n=length(x); n
[1] 10
> xbar=mean(x); xbar
[1] 69.9
> s=sd(x); s
[1] 10.94887
> se=s/sqrt(n); se
[1] 3.462337
> alpha=0.05
> c=qt(1-alpha/2, n-1); c
[1] 2.262157
> xbar-c*se
[1] 62.06765
```

```
> xbar+c*se
[1] 77.73235
```

$n=10,\quad \overline{X}=69.9,\quad s=10.94887,\quad df=n-1=9,\quad \frac{s}{\sqrt{n}}=3.4623,$

$t_{9,0.025}=2.262$

임신 후기 임산부의 평균 공복혈당에 대한 95% 신뢰구간은

$69.9\pm2.262\times3.4623=69.9\pm7.8317=(62.0683,\ 77.7317)$

즉, 95% 신뢰수준에서 임신 후기 임산부의 평균 공복혈당은 62.0683~77.7317 $mg/100ml$이다.

04 $n=100$

$\hat{p}=\frac{15}{100}=0.15$

$\widehat{SE}(\hat{p})=\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}=\frac{\sqrt{0.15\times0.85}}{\sqrt{100}}=0.0357$

무지개 송어 알에 주입된 성장 호르몬 유전자가 부화된 무지개 송어의 DNA에 남을 확률에 대한 95% 신뢰구간은

$0.15\pm1.96\times0.0357=0.15\pm0.069972=(0.080028,\ 0.219972)$

05 $n=30$

$\hat{p}=\frac{21}{30}=0.7,\quad n\hat{p}=30\times0.7=21\geq5,\quad n(1-\hat{p})=30\times0.3=9\geq5$

$\widehat{SE}(\hat{p})=\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}=\frac{\sqrt{0.7\times0.3}}{\sqrt{30}}=0.0837$

따라서 똑같은 환경에서 냄새로 형제를 알아보는 확률에 대한 90% 신뢰구간은

$0.7\pm1.645\times0.0837=(0.5623,\ 0.8377)$

06

```
> x=c(0.0132, 0.0138, 0.0108, 0.0125, 0.0136, 0.0111, 0.0124, 0.0116,
      0.0126, 0.0131)
> n=length(x); n
[1] 10
> xbar=mean(x); xbar
```

```
[1] 0.01247
> s=sd(x); s
[1] 0.001023122
> se=s/sqrt(n); se
[1] 0.0003235395
> alpha=0.05
> c=qt(1-alpha/2, n-1); c
[1] 2.262157
> xbar-c*se
[1] 0.0117381
> xbar+c*se
[1] 0.0132019
```

강철빔의 평균에 대한 95% 신뢰구간 = (0.01174,0.01320)

07 alpha = 0.05

오차범위 c = 0.2

$$n \geq (Z_{\frac{\alpha}{2}} \times \frac{1}{c})^2 = (1.96 \times \frac{1}{0.2})^2 = 96.04$$

n = 97

08 alpha = 0.01

오차범위 c = 0.02

$$n \geq (Z_{\frac{\alpha}{2}} \times \frac{\sqrt{0.5 \times 0.5}}{c})^2 = (Z_{\frac{\alpha}{2}} \times \frac{\sqrt{p(1-p)}}{c})^2$$

$$(Z_{\frac{\alpha}{2}} \times \frac{\sqrt{0.5 \times 0.5}}{c})^2 = (2.58 \times \frac{0.5}{0.02})^2 = 4160.25$$

n = 4161

8장 연습문제 해답

01 (1) $H_0 : \mu \le 36 \qquad H_1 : \mu > 36$

(2) $n = 50,\ \bar{x} = 38,\ s = 10,\ \alpha = 0.05$

$$z = \frac{38 - 36}{10/\sqrt{50}} = 1.4142$$

(3) $c = z_{0.95} = 1.645$

(4) $P(Z > 1.4142) = 0.0786$

(5) $z < c$(임계값)이므로 H_0을 기각하지 않는다. 또는 유의확률이 0.05보다 크므로 H_0을 기각하지 않는다.

02 (1)

```
> x = c(60, 56, 80, 55, 74, 68, 75, 73, 68, 90)
> n=length(x); n
[1] 10
> xbar=mean(x); xbar
[1] 69.9
> s=sd(x); s
[1] 10.94887
> se=s/sqrt(n); se
[1] 3.462337
> t=(xbar-80)/se; t
[1] -2.917105
```

$H_0 : \mu = 80 \qquad H_1 : \mu \neq 80$

$n = 10,\ \bar{x} = 69.9,\ s = 10.94887,\ \alpha = 0.05$

$$t = \frac{69.9 - 80}{10.94887/\sqrt{10}} = -2.917105$$

(2)

```
> alpha=0.05
> c=qt(1-alpha/2, n-1); c
[1] 2.262157
```

$c = t_{9,0.025} = 2.26$

(3)
```
> pvalue = 2 * pt(t, n-1); pvalue
[1] 0.01711224
```

$2 \times P(T < -2.92) = 2 \times 0.0085 = 0.017$

(4) 유의확률이 0.05보다 작으므로 H_0를 기각한다.

03 (1) $H_0 : p \geq 0.23 \qquad H_1 : p < 0.23$

$n = 100,\ \hat{p} = \dfrac{15}{100} = 0.15,\ \alpha = 0.05$

$z = \dfrac{0.15 - 0.23}{\sqrt{0.23(1-0.23)} / \sqrt{100}} = -1.9010$

(2) $P(Z < -1.9010) = 0.0287$

(3) 유의확률이 0.05보다 작으므로 H_0를 기각한다.

04 $H_0 : \mu \geq 11.5 \qquad H_1 : \mu < 11.5$

$n = 40,\ \bar{x} = 11.4,\ s = 0.6,\ \alpha = 0.1$

$z = \dfrac{11.4 - 11.5}{0.6 / \sqrt{40}} = -1.0541$

c(임계값) $= -z_{0.1} = -1.28$

$z > c$이므로 H_0를 기각하지 않는다.

05 $H_0 : p \leq 0.6 \qquad H_1 : p > 0.6$

$n = 30,\ \hat{p} = 21/30 = 0.7,\ \alpha = 0.05$

$z = \dfrac{0.7 - 0.6}{\sqrt{0.6(1-0.6)} / \sqrt{30}} = 1.1180$

유의확률: $P(Z > 1.1180) = 0.1318$

유의확률이 0.05보다 크므로 H_0를 기각하지 않는다.

06 $H_0 : \mu = 7.6 \qquad H_1 : \mu \neq 7.6$

$n = 100,\ \bar{x} = 7.5,\ s = 1.0$

$z = \dfrac{7.5 - 7.6}{1 / \sqrt{100}} = -1$

유의확률: $2 \times P(Z < -1) = 2 \times P(Z > 1) = 2 \times 0.1586 = 0.3172$

유의확률이 0.05, 0.01보다 크므로 유의수준 5%, 1% 둘 다 H_0를 기각하지 않는다.

07 $H_0 : p \geq 0.73 \qquad H_1 : p < 0.73$

$n = 50,\ \hat{p} = 34/50 = 0.68,\ \alpha = 0.05$

$$z = \frac{0.68 - 0.73}{\sqrt{0.73(1-0.73)}/\sqrt{50}} = -0.7964$$

유의확률: $P(Z < -0.7964) = 0.2129$

유의확률이 0.05보다 크므로 H_0를 기각하지 않는다.

유의수준 5%로, 새로운 프로그램이 테스트를 통과하지 못한 직원의 비율을 작년에 비하여 낮추는 효과가 있다.

08 $H_0 : p \geq 0.15 \qquad H_1 : p < 0.15$

$n = 140,\ \hat{p} = 19/140 = 0.1357,\ \alpha = 0.01$

$$z = \frac{0.1357 - 0.15}{\sqrt{0.15(1-0.15)}/\sqrt{140}} = -0.4739$$

유의확률: $P(Z < -0.4739) = 0.3178$

유의확률이 0.01보다 크므로 H_0를 기각하지 않는다.

1% 유의수준으로, 경영진이 도입한 공정 이후 독성화학물질로 인하여 건강에 문제를 가진 직원의 비율이 기존의 15%보다 유의하게 줄지 않았다.

09

```
〉 x=c(0.0132, 0.0138, 0.0108, 0.0125, 0.0136, 0.0111, 0.0124, 0.0116,
       0.0126, 0.0131)
〉 n=length(x); n
[1] 10
〉 xbar=mean(x); xbar
[1] 0.01247
〉 s=sd(x); s
[1] 0.001023122
〉 se=s/sqrt(n); se
[1] 0.0003235395
〉 t=(xbar-0.012)/se; t
[1] 1.452682
〉 pvalue=1-pt(t,n-1); pvalue   #P(T>t)
[1] 0.09013443
```

$H_0 : \mu \le 0.012 \qquad H_1 : \mu > 0.012$

$n = 10,\ \overline{x} = 0.01247,\ s = 0.001023,\ \alpha = 0.01$

$t = \dfrac{0.01247 - 0.012}{0.001023/\sqrt{10}} = 1.4528$

유의확률: $P(T > 1.4528) = 0.0901$

유의확률이 0.01보다 크므로 H_0를 기각하지 않는다.

9장 연습문제 해답

01 측정치(표본)는 정규분포를 따른다. → Z 검정

$H_0 : \mu_1 = \mu_2,\ H_1 : \mu_1 \neq \mu_2$

$n_1 = 12,\ \overline{X} = 17.325,\ s_1 = 12.62974$

$n_2 = 12,\ \overline{Y} = 11.88,\ s_2 = 10.10376$

$\overline{X} - \overline{Y} = 17.325 - 11.88 = 5.445$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(12-1)(12.62974)^2 + (12-1)(10.10376)^2}{12+12-2}}$$

$= 11.4367$

$SE = SE(\overline{X} - \overline{Y}) = s_p\sqrt{1/n_1 + 1/n_2} = 4.669013$

$Z = \dfrac{5.445}{4.669013} = 1.1662$

p−value $= 2\{1 - \text{pnorm}(1.1662)\} = 0.244$

p−value $>$ 0.05이므로 H_0가 기각되지 않는다.

02 $H_0 : \mu_1 = \mu_2,\ H_1 : \mu_1 > \mu_2$

$D = X - Y$

$\overline{D} = 1.88,\ s_d = 0.89368$

$T = \dfrac{\overline{D}}{s_d} = \dfrac{1.88}{0.89368} = 2.1036$

$\text{p}-\text{value}=1-\text{pt}(2.1036,\ 9)=0.032$

$\text{p}-\text{value}\ <\ 0.05$이므로 H_0가 기각된다.

03 $H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2$

$\hat{p_1}=0.3,\ n_1=1000$

$\hat{p_2}=0.34,\ n_2=1500$

$\hat{p}=\dfrac{300+510}{1000+1500}=0.324$

$\hat{p_1}-\hat{p_2}=0.3-0.34=-0.04$

$SE=\sqrt{\hat{p}(1-\hat{p})}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}=0.0191$

$Z=\dfrac{-0.04}{0.0191}=-2.09$

$\text{p}-\text{value}=2*\text{pnorm}(-2.09)=0.0366$

$\text{p}-\text{value}\ <\ 0.05$이므로 H_0가 기각된다.

04

```
〉 market=read.csv(file="ch9_market.csv")
〉 x=market$market1
〉 y=market$market2
〉 d=x-y
〉 dbar=mean(d); dbar
[1] -0.1343478
〉 s=sd(d); s
[1] 0.432328
〉 n=length(d); n
[1] 23
〉 se=s/sqrt(n); se
[1] 0.09014662
〉 t= (dbar)/se; t
[1] -1.490326
〉 pvalue=pt(t,n-1); pvalue
[1] 0.0751687
```

유의확률 = 0.075

유의확률 > 0.05이므로 5% 유의수준으로 귀무가설을 기각하지 않는다.

('Priceless'가 'Cheap' 마켓보다 상품을 싸게 판매한다고 할 수 없다.)

그러나 유의확률 < 0.10이므로 10% 유의수준으로는 귀무가설을 기각한다.

05

```
〉 temp=read.csv(file="ch9_bodytemp.csv")
〉 x=temp$men; y=temp$women
〉 n1=length(x); n2=length(y)
〉 xbar=mean(x); xbar
[1] 36.72615
〉 ybar=mean(y); ybar
[1] 36.88923
〉 s1=sd(x); s2=sd(y)
〉 se=sqrt(s1**2/n1 + s2**2/n2); se
[1] 0.07028103
〉 z=(xbar-ybar)/se; z
[1] -2.320355
〉 pvalue=2*(1-pnorm(abs(z))); pvalue
[1] 0.02032169
```

유의확률 = 0.02

유의확률 < 0.05이므로 귀무가설을 기각한다.

즉, 유의수준 5%로 남녀간 평균 체온에 유의미한 차이가 있다.

06

```
〉 xbar=4; ybar=20
〉 s1=12; s2=15
〉 n1=10; n2=10
〉 sp=sqrt(((n1-1)*s1**2 + (n2-1)*s2**2)/(n1+n2-2)); sp
[1] 13.58308
〉 se=sp*sqrt(1/n1+1/n2); se
[1] 6.074537
〉 df=n1+n2-2
〉 t=(xbar-ybar)/se; t
```

```
[1] -2.633946
> pvalue=pt(t, df); pvalue
[1] 0.008425275
```

유의확률 = 0.008

유의확률 < 0.01이므로 1% 유의수준으로 하루는 잠이 부족하지만 다음날 잠을 보충하는 것이 연속 이틀 동안 잠이 부족한 것에 비하여 인지능력 향상에 유의한 효과가 있다.

07 $H_0 : \mu_1 = \mu_2,\ H_1 : \mu_1 > \mu_2$

```
> n1=50; n2=50
> xbar=85.8; ybar= 80.4
> s1=14.4; s2=16.5
> se=sqrt(s1**2/n1 + s2**2/n2); se
[1] 3.097128
> z=(xbar-ybar)/se; z
[1] 1.743551
> pvalue=1-pnorm(abs(z)); pvalue
[1] 0.04061871
```

유의확률 = 0.04

유의확률 < 0.05이므로 5% 유의수준으로 기초 통계학 과목에서 매주 퀴즈를 보는 것이 성적 향상(평균성적 증가)에 유의한 영향을 미친다.

10장 연습문제 해답

01 미식축구공 모양

02 (1) T, 상관계수는 단위에 의존하지 않으므로 서로 다른 쌍들을 비교할 수 있다.

(2) F, 기울기는 상관계수뿐만 아니라 X와 Y의 표준편차에 의존하므로 기울기가 크다고 해서 예측이 더 정확하다고 할 수 없다.

(3) F, 연관성은 있으나 인과관계는 알 수 없다.

(4) T, beta1 = r * s_y/s_x인데, s_y/s_x >0이므로 상관계수와 회귀직선의 기울기 추정치는 항상 부호가 같다.

(5) F, 비선형관계 존재 가능

03 Z와 W를 표준화하면 각각 X와 Y를 표준화한 것과 같으므로

$\mathrm{cor}(Z, W) = \mathrm{cor}(X, Y) = 0.1$

선형변환은 상관계수에 영향을 미치지 않는다.

04 (1) $r = S_{xy} / \sqrt{S_{xx} S_{yy}} = 1165 / \sqrt{(85.4 * 26053.3)} = 0.781$

(2) beta1 = Sxy/Sxx = 1165/85.4 = 13.64

05 (1) F, 산점도가 정확하게 직선이 아니고 오차가 있으므로 "반드시" 더 크지는 않다.

(2) F, 기울기=1.5(상관계수는 기울기가 아니다. 더욱이 상관계수는 1보다 클 수 없다)

(3) F, 관측범위를 벗어난 예측은 위험. age=0은 관측범위를 벗어난다(갓 태어난 아기는 윗몸일으키기를 못한다).

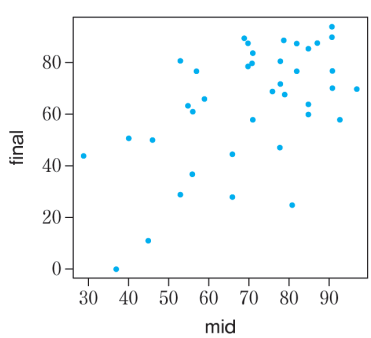
(4) T, 기울기와 상관계수는 부호가 같다.

06 (1) n=100 => 정규 분포, 0.72±1.96*0.40 = (−0.064, 1.504)

(2) 0.072

07 (1)

```
score=read.csv("ch10_testscore.csv")
mid=score$mid
final=score$final
plot(mid, final)
```



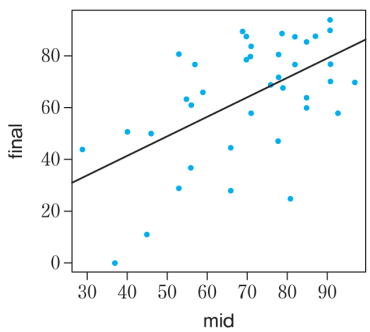
선형관계가 존재한다.

(2)

```
cor(mid, final)
[1] 0.5564777
```

(3)

```
out=lm(final ~ mid) # 또는 out=lm(final~mid, data=score)
abline(out)
```



(4)

```
summary(out)

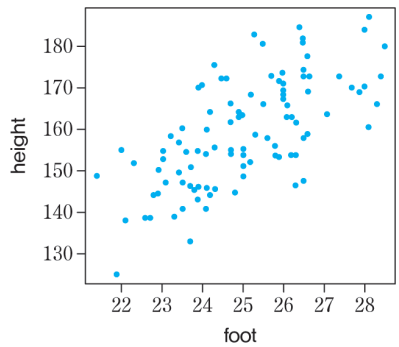
Coefficients:
             Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)  10.9280     13.3907      0.816       0.419675
mid           0.7526      0.1847      4.074       0.000234 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.52 on 37 degrees of freedom
Multiple R-squared:  0.3097,        Adjusted R-squared:  0.291
F-statistic:  16.6 on 1 and 37 DF,  p-value: 0.0002344
```

검정의 유의확률은 0.00023이다.

08 (1)

```
fh=read.csv("ch10_foot-height.csv")
foot=fh$foot; height=fh$height
plot(foot, height)
```



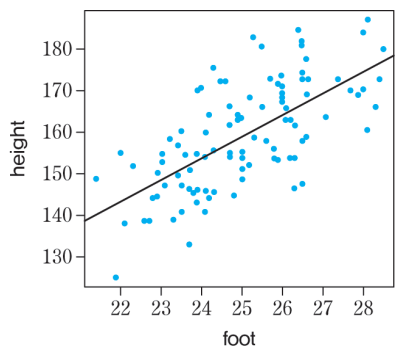
선형관계가 존재한다.

(2)

```
cor(foot, height)
[1] 0.6664676
```

(3)

```
out=lm(height~foot)
abline(out)
```



(4)

```
summary(out)
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  27.9708    14.8737     1.881    0.063 .
foot          5.2432     0.5925     8.850    3.7Se-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.672 on 98 degrees of freedom
Multiple R-squared:  0.4442,    Adjusted R-squared:  0.4385
F-statistic: 78.32 on 1 and 98 DF,  p-value: 3.784e-14
```

height의 전체 변동 중 회귀직선으로 설명되는 변동은 44.4%이다.

(5) (4)의 출력 결과에서 foot 계수(beta1)의 유의확률=3.78e−14 < 0.000001이므로 유의수준 1%로 height를 추정하는 데 foot이 유의한 영향을 미친다.

(6)

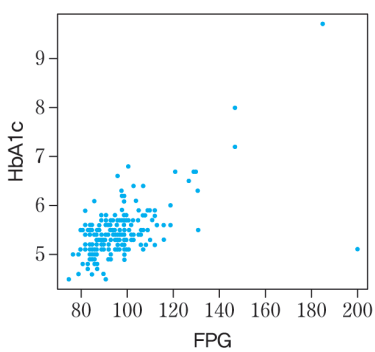
```
predict( out, data.frame(foot=24.0), se=T , interval="confidence")
$fit
fit        lwr        upr
1 153.8084 151.5257 156.0911
(151.53, 156.09)
```

(7)

```
predict( out, data.frame(foot=26.0),   interval="prediction")
fit        lwr        upr
1 164.2948 144.9739 183.6158
(144.97, 183.62)
```

09 (1)

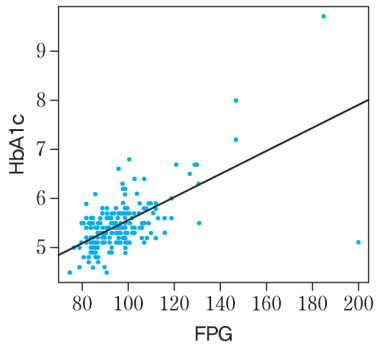
```
db=read.csv("ch10_Diabetes.csv")
HbA1c=db$HbA1c; FPG=db$FPG
plot(FPG, HbA1c)
```



선형관계가 존재한다.

(2)

```
out=lm(HbA1c~FPG)
abline(out)
```



(3)

```
summary(out)
Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  3.190759    0.201098     15.87      <2e-16 ***
FPG          0.023395    0.002059     11.36      <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4416 on 198 degrees of freedom
Multiple R-squared:  0.3947,    Adjusted R-squared:  0.3916
F-statistic: 129.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

출력 결과에서 FPG 계수(beta1)의 유의확률 < 2e−16이므로 유의수준 5%로 HbA1c를 추정하는 데 FPG가 유의한 영향을 미친다.

(4)

```
predict(out, data.frame(FPG=96), interval="confidence")
fit       lwr       upr
1 5.436654 5.375039 5.498268
(5.375, 5.498)
```

(5)

```
predict(out, data.frame(FPG=105), interval="predict")
fit       lwr       upr
1 5.647206 4.773433 6.520979
(4.773, 6.521)
```

10 (1)

```
chol=read.csv("ch10_Cholera.csv")
out=lm(death ~ ., data=chol)
summary(out)

Coefficients:
                Estimate     Std. Error   t value    Pr(>|t|)
(Intercept)   103.885019     22.091028     4.703     4.42e-05 ***
houseVal        0.385024      1.361394     0.283     0.779085
poorRate      735.166439    174.863887     4.204     0.000188 ***
elevation     -27.854200      3.714579    -7.499     1.27e-08 ***
density         0.009206      0.048160     0.191     0.849581
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.42 on 33 degrees of freedom
Multiple R-squared:  0.812,     Adjusted R-squared:  0.7892
F-statistic: 35.62 on 4 and 33 DF,  p-value: 1.525e-11
```

(2) 앞의 summary(out)에서 유의확률에 *가 하나 이상 표시된 변수는 elevation, poorRate이므로 death를 추정하는 데 5% 유의수준으로 유의한 영향을 미치는 변수는 elevation과 poorRate이다. 다른 변수들은 5% 유의수준에서 유의하지 않다.

(3) 다른 설명변수들이 고정된 상태에서 elevation이 1단위 증가하면 death는 평균적으로 약 27.85단위 감소한다.

(4) 다른 설명변수들이 고정된 상태에서 poorRate이 1단위 증가하면 death는 평균적으로 약 735.16단위 증가한다.

(5) 81%(Multiple R-squared: 0.812)

(6) F-statistic: 35.62 on 4 and 33 DF, p-value: 1.525e-11이므로, 회귀직선은 유의하다. 즉, 한 개 이상의 설병변수가 death 추정에 유의한 영향을 미친다.

11 (1)

```
out=lm(death ~ elevation+poorRate, data=chol)
summary(out)

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   107.587      17.783     6.050   6.63e-07 ***
elevation     -27.608       3.515    -7.855   3.12e-09 ***
poorRate      716.354     147.050     4.871   2.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.76 on 35 degrees of freedom
Multiple R-squared:  0.8113,    Adjusted R-squared:  0.8005
F-statistic: 75.22 on 2 and 35 DF,  p-value: 2.125e-13
```

(2)

```
yhat= 107.587 - 27.608*3.9 + 716.354*0.06 = 42.89704
```

(3)

```
Multiple R-squared:  0.8113 (2개 - elevation, poorRate 변수 사용)
Multiple R-squared:  0.812  (4개 변수 사용)
```

차이가 거의 없다.

(4) 전체 변동 중 회귀직선으로 설명되는 변동이 두 모형에서 거의 차이가 없다. 이는 2개의 변수를 추가함으로써 증가하는 설명력이 거의 없다는 뜻이므로 상대적으로 간단한 모형이 더 좋다.

12 (1)

```
cor(chol$death, chol$elevation);
[1] -0.8266172
cor(chol$death, chol$poorRate);
[1] 0.6917803
cor(chol$elevation, chol$poorRate)
[1] -0.4505271
```

(2) 위의 세 변수는 모두 서로 강한 상관관계를 가진다. 실은, 고도가 낮은 템스 강변은 하수처리 시설이 없던 당시 수질 오염이 극심하였고 가난한 사람들이 사는 지역이었으며 콜레라 사망률이 매우 높았다. 수질 오염이 콜레라를 유발하

는 요인인데, 이 요인이 고도, 빈곤도와 관련이 있으므로 고도, 빈곤도가 사망률과 관련이 있는 현상이 발생한다. 그러나 고도, 빈곤도가 콜레라 사망의 원인은 아니다.

13 (1)

```
health=read.csv("ch10_health.csv")
out=lm(HbA1c ~ ., data=health)
summary(out)

Coefficients:
             Estimate    Std. Error   t value   Pr(>|t|)
(Intercept) 2.5667261   0.3898507     6.584    4.14e-10 ***
sysBP       0.0019429   0.0018543     1.048    0.2960
TCHL        0.0021119   0.0008921     2.367    0.0189 *
FPG         0.0233224   0.0020660    11.289    < 2e-16 ***
BMI        -0.0006981   0.0101341    -0.069    0.9452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4372 on 195 degrees of freedom
Multiple R-squared:  0.4158,    Adjusted R-squared:  0.4039
F-statistic:  34.7 on 4 and 195 DF,  p-value: < 2.2e-16
```

(2) 앞의 summary(out)에서 유의확률에 *가 하나 이상 표시된 변수는 TCHL, FPG이므로 HbA1c를 추정하는 데 유의한 영향을 미치는 변수는 TCHL과 FPG이다. 다른 변수들은 유의하지 않다.

(3) 다른 설명변수들이 고정된 상태에서 FPG가 1단위 증가하면 HbA1c는 평균적으로 약 0.023단위 증가한다.

(4) 41.6%(Multiple R-squared: 0.4158)

(5) F-statistic: 34.7 on 4 and 195 DF, p-value: <2.2e-16이므로, 회귀직선은 유의하다. 즉, 한 개 이상의 설병변수가 HbA1c 추정에 유의한 영향을 미친다.

14 (1)

```
out=lm(HbA1c ~ TCHL+FPG, data=health)
summary(out)

Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  2.751746   0.267616     10.282     <2e-16 ***
TCHL         0.002166   0.000885      2.448      0.0153 *
FPG          0.023606   0.002035     11.598     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4362 on 197 degrees of freedom
Multiple R-squared:  0.4126,    Adjusted R-squared:  0.4066
F-statistic: 69.17 on 2 and 197 DF,  p-value: < 2.2e-16
```

(2) 41.3%(Multiple R-squared: 0.4126)

(3) 전체 변동 중 회귀직선으로 설명되는 변동이 두 모형(4개 설명변수 모형과 2개 설명변수 모형)에서 거의 차이가 없다. 이는 2개의 변수를 추가함으로써 증가하는 설명력이 거의 없다는 뜻이므로 상대적으로 간단한 2개 설명변수 모형이 더 좋다.