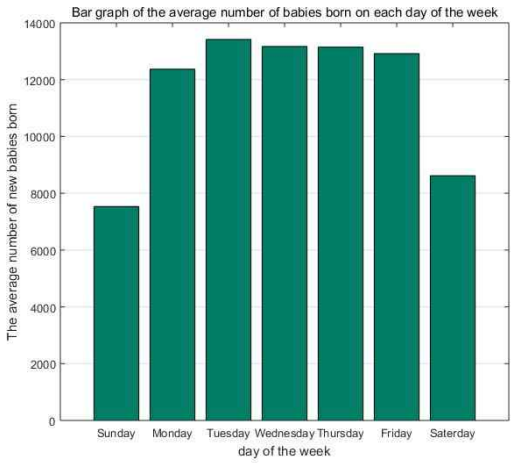
Statistics I

Homework 1

학 과	응용수학과
교수님	김경수 교수님
학 번	2014110374
이 름	정상만
제출일	2018. 4. 18.

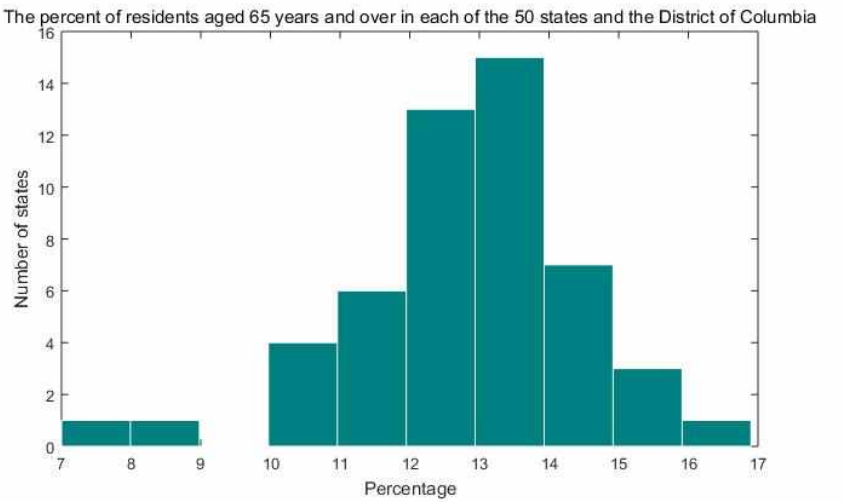
1.5 exercise)



[각 요일별 출생한 평균 신생아 수]

2008년도 각 요일마다 출생한 평균 신생아 수에 대한 **원 그래프는 적절하지 않다.**
원 그래프를 적용함으로써 요일별 분포는 확연히 알 수 있으나 주중과 주말에 관한 분포는 파악하기 어렵다. 하지만 막대그래프에서 주중과 주말에 관한 분포를 쉽게 알 수 있고, 각각의 요일별 신생아 수도 수치적이고 직관적으로 알 수 있다. 위의 그래프는 각 요일마다 출생한 평균 신생아 수에 관한 막대그래프이다. (Matlab 프로그램 이용)

1.8 exercise)



[미국 50개 주 및 콜롬비아 특별지구에서의 65세이상 노인인구의 백분율]

연습문제 1.6의 표 1.2를 참고하여 히스토그램으로 나타내었다(Matlab 프로그램 이용).
분포의 형태는 왼쪽으로 기울어졌으며, 비대칭적이다. 위 자료의 **중앙값**(분포의 중간점)은 13.0 (캔자스)이며 분포의 **퍼진 정도**는 최소값인 7(알래스카)부터 최대값인 16.9(플로리다)이다.

2.4 exercise)

문제에서 판매된 신규 주택가격이 평균값=213000, 중앙값=268700라 주어졌는데, 이 중 어느 것이 평균값과 중앙값인지 묻는 것은 의도가 적절하지 않다. 평균값과 중앙값이 위와 같을 때 분포를 설명하는 것을 묻는 문제로 생각하면, 우선, 평균값과 중앙값이 동일하거나 매우 유사한 값을 갖지 않으므로 **분포가 비대칭적**임을 암시하고, 평균값<중앙값 의 관계로 떨어져있으므로 **분포는 왼쪽으로 기울어진 형태**임을 예상할 수 있다.

2.6 exercise)

a)

28	0 5
29	8
30	0 5 5
32	5

스템플롯은 위와 같이 작성할 수 있다.
방어선에 있는 선수들의 체중을 순서대로 배열하면

280 285 298 300 305 305 325

이고 중앙값은 $\frac{7+1}{2}=\frac{8}{2}=4$ 번째 위치인 300 이다. 중앙값을 기준으로 왼쪽에 위치한 관찰값들의 중앙값인 제 1 사분위수 Q_1은 $\frac{3+1}{2}=2$ 번째 위치인 285가 된다.
마찬가지로 중앙값을 기준으로 오른쪽에 위치한 관찰값들의 중앙값인 제 3 사분위수 Q_3 은 Q_1의 위치와 마찬가지로 중앙값 기준 오른쪽의 관찰값들의 2번째인 305가 된다. 최대값은 325이고 최소값은 280이므로 다섯 개 숫자로 나타낸 개요는 다음과 같다.

Min	Q_1	Median	Q_3	Max
280	285	300	305	325

b)

30	4	4
31	5	8 9
32	4	5
33	8	
34	4	

스템플롯은 위와 같이 작성할 수 있다.
공격선에 있는 선수들의 체중을 순서대로 배열하면

304 304 315 318 319 324 325 338 344

이고 중앙값은 $\frac{9+1}{2}=5$번째 위치에 있으므로 319 이다. a)와 마찬가지로 사분위수들을 구하면, Q_1의 위치는 $\frac{4+1}{2}=2.5$이므로 304와 315 사이에 위치한다. 따라서 $Q_1=\frac{304+315}{2}=309.5$ 이다. Q_3도 마찬가지로 관찰값의 개수가 짝수이므로 $Q_3=\frac{325+338}{2}=331.5$ 가 된다. 최소값은 304 이고 최대값은 334 이므로 다섯 개 숫자로 나타낸 개요는 다음과 같다.

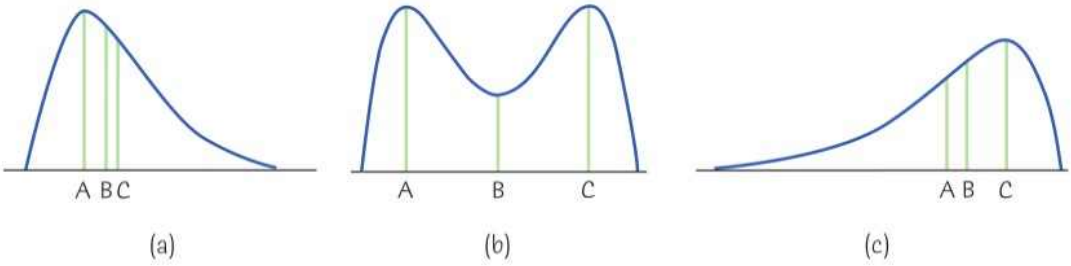
Min	Q_1	Median	Q_3	Max
304	309.5	300	331.5	334

c) 위의 두 그룹 모두 이탈값을 갖지 않는다. 공격선에 있는 선수들의 체중이 더 무거운 경향이 있다. 이는 두 그룹의 각각의 스템플롯을 비교하면 곧바로 확인할 수 있다.

2.8 exercise)

예제 2.1, 2.3, 2.6과 $1.5\times IQR$ 법칙에 의하면 $Q_3+(1.5\times IQR)=83.75$ 이고, 이를 초과하는 이탈값은 '85' 분이 유일하다. 또한 $Q_1-(1.5\times IQR)=-26.25$ 에도 당연히 초과하여 위치하지 않으므로 의심쩍은 이탈값으로 간주되지 않는다.

3.4 exercise)



a)
B 가 중앙값, C가 평균값이다. (∵ 곡선이 비대칭적이고 오른쪽으로 기울어졌으므로, 평균값은 긴 꼬리 쪽을 향하여 중앙값으로부터 떨어져 위치한다.)

b)
곡선이 대칭적이므로 B가 중앙값이고, 균형점도 B이므로 평균값도 B이다.

c)
분포가 비대칭적이므로 등적점은 B, 즉 중앙값이고, 분포가 왼쪽으로 기울어졌으므로 B에서 긴꼬리 쪽으로 향하여 떨어져 위치하는 A가 평균값이 된다.

3.7 exercise)

a)
68-95-99.7 법칙에 의하면, 95%는 평균값 μ 에 대하여 2σ 범위에 위치한다. 이 때 $\mu = 852$, $\sigma = 82$ 이므로 몬순 강우량의 구간 $(\mu - 2\sigma, \mu + 2\sigma) = (852 - 164, 852 + 164) = (688, 1016)$ 범위에 위치한다.

b)
μ를 기준으로 분포에서 95%에 해당하는 면적을 제하면 양 끝에 남는 면적을 고려하여 2.5%+2.5%=5%의 면적이 남는다. 따라서 95%를 제외한 나머지 면적의 $\frac{1}{2}$을 구하면 된다. 이 때 가장 건조했던 2.5% 이므로, 결국 분포에서 음수 범위가 된다. 따라서 가장 건조했던 2.5%에 해당하는 몬순 강우량을 M 이라 두면 $M \leq 688$이다.

3.11 exercise)

a)
강우량 697밀리미터를 표준정규분포로 표준화하기 위해 다음 공식을 사용하자.

$$z = \frac{x-\mu}{\sigma},\ z : \text{표준화된 값},\ x\text{: } N(\mu,\sigma)\text{의 한 개 관찰값.}$$

$x \le 697, \mu = 852,\ \sigma = 82$ 이므로 x의 표준화된 값은 $z \le \frac{697-852}{82} = -1.89$이다.

∴ 표준정규분포표를 참조하면 0.0294=2.94% 가 백분율이 된다.

b)
위와 마찬가지로 표준화하여 백분율을 구하자.
우선 $683 < x < 1022$ 의 범위에 있으므로 이를 표준화 하면
$\frac{683-852}{82} = -2.06 < z < 2.07 = \frac{1022-852}{82}$ 이므로 표준정규분포표를 참조하면
∴ 0.9808-0.0197=0.9611=96.11% 가 백분율이 된다.

4.1 exercise)

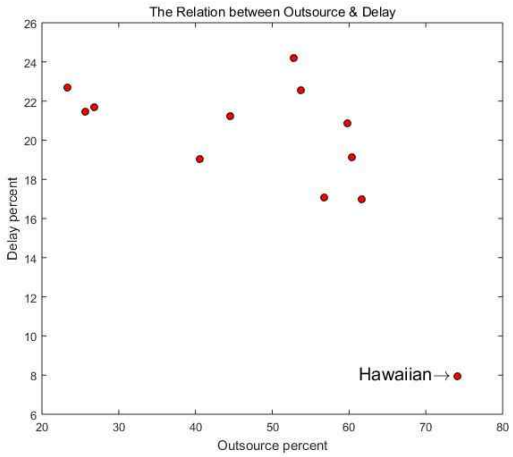
a) 설명변수 : 통계학 시험을 준비하는 데 사용한 시간의 양, 반응변수 : 통계학 시험의 학점.

b) 몸무게에 따라 키가 확실히 결정되는 것은 아니지만 키에 따른 몸무게는 어느정도 가늠할 수 있다. 따라서 키를 설명변수로 하고 몸무게를 반응변수로 하는 것이 적합하다.

c) 설명변수 : 페이스 북을 이용하기 위해 온라인에서 사용한 주당 시간, 반응변수 : 학점 평균.

d) 독해를 잘 한다고 해서 작문을 잘 할 수 있는 것은 보장할 수 없지만 작문을 잘한다면 독해능력이 좋다고 볼 수 있으므로 설명변수 : SAT 작문시험의 점수, 반응변수 : SAT 독해시험의 점수 로 두는 것이 적합하다.

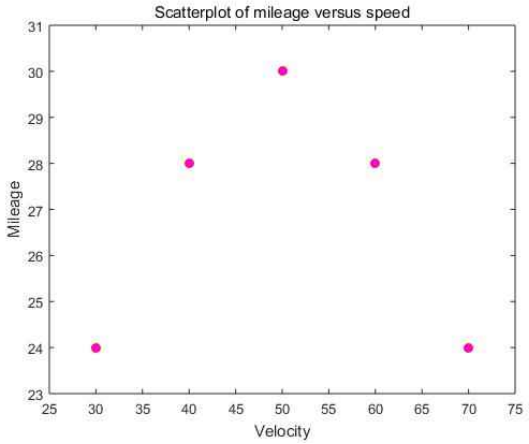
4.7 exercise)



[아웃소싱과 운항지연에 대한 산포도]

연습문제 4.5에서의 산포도를 그려야 연습문제 4.7에 대한 설명이 편리해지므로 직접 산포도를 그리면 위와 같은 그래프를 얻는다(Matlab 프로그램 이용). 위 산포도를 관찰하면, **아웃소싱과 운항지연 사이엔 어떤 관계도 존재하지 않는다.** 즉, 아웃소싱의 평균은 48.34 이고 운항지연의 평균은 19.57 인데, 각 항공사에 대한 대부분의 설명변수와 반응변수들이 양과 음의 관계 모두 해당되지 않는 값을 보인다. 예를 들면 Hawaiian 항공사의 경우, 아웃소싱 백분율이 74.1로 평균값을 훨씬 웃도는 수치이지만, 운항지연은 평균값보다 한참 아래인 7.94에 그친다. 이 둘은 서로 어떠한 관계에도 있지 않다. 또한 **Hawaiian 항공사는 위 산포도에서 이탈값**이다. 그러나 이 유일한 **이탈값을 제외**하더라도, 산포도는 **매우 약한 음의 선형관계를 가지게 된다.**

4.13 exercise)



[속도와 마일리지에 관한 산포도]

x를 속도, y를 마일리지로 두면 각각의 평균은$\bar{x}=50$, $\bar{y}=26.8$이고 x에 관한 표준편차는 $S_x=15.8114$, y에 관한 표준편차 $S_y=2.6833$ 이다. 이를 가지고 각 개체의 각각의 변수의 표준화된 값을 구하여 표로 나타내면 다음과 같다. (여기서 표준화는 $z=\frac{x-\mu}{\sigma}$ 이다.)

z_x	z_y	z_xz_y
-1.2649	-1.0435	1.3199
-0.6325	0.4472	-0.2828
0	1.1926	0
0.6325	0.4472	0.2828
1.2649	-1.0435	-1.3199

위 표에서 $\sum z_xz_y=0$ 이 되고, 따라서 상관 $r=\frac{1}{5-1}\sum z_xz_y=\frac{0}{5-1}=0$ 이므로 이들 변수들은 직선관계가 없으며 음 또는 양의 관계도 존재하지 않는다.

5.2 exercise)

새 비누의 중량이 80그램이므로, 이를 일수 $t=0$일 때 비누의 중량 $f(t)|_{t=0}=80$이라 하자. 비누의 중량이 매일 평균적으로 5그램씩 감소하므로 $t=1$일 때 $f(1)=75$이다. 이는 직교좌표상의 두 점으로 표현할 수 있으므로 $(0,80),(1,75)$를 이용하여 직선의 방정식을 구하면 $f(t)=80-5t$이고, 이는 회귀선의 식이 된다.

5.3 exercise)

a) & b)

해수면 온도의 평균값은 $\bar{x}=30.28$, 성장의 평균값 $\bar{y}=2.5157$, 각각의 표준편차는 $S_x=0.4296$, $S_y=0.1508$이다. 이를 이용하여 각 개체의 변수들을 표준화하여 정리하면

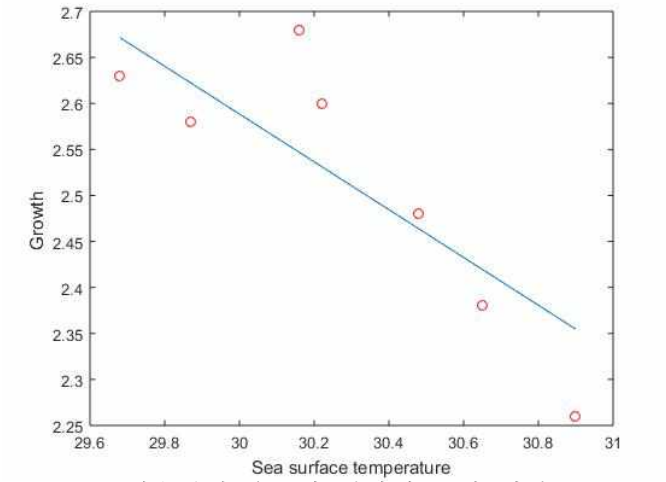
z_x	z_y	z_xz_y
-1.3966	0.7581	-1.0587
-0.9543	0.4264	-0.4069
-0.2793	1.0897	-0.3044
-0.1397	0.5591	-0.0781
0.4655	-0.2369	-0.1103
0.8612	-0.9002	-0.7753
1.4432	-1.6962	-2.4478

위의 표에 근거하여 상관 $r=-0.7402$ 임을 얻는다. 이제 회귀선의 식을 구하여 보자.

회귀선의 기울기 $b=r\frac{S_y}{S_x}=(-0.7402)\times\frac{0.1508}{0.4296}=-0.2598$, 절편

$a=\bar{y}-b\bar{x}=2.5157-(-0.2598)\times(30.28)=10.3812$ 이다.

따라서 최종적인 최소제곱에 의한 회귀선의 식은 $\hat{y}=10.3812-0.2598x$ 이다.



[해수면의 온도와 성장에 관한 회귀]

5.17 exercise)

결혼한 남자들의 일반적인 연령대를 고려해보면 20대 초중반을 시작으로 20대 후반과 30대 초중반에서 가장 많을 것이고 30대 후반 이후로 적어질 것이다. 그런데 20대 후반과 30대 초중반은 사회적인 위치로 볼 때 평균소득이 20대 초중반보다 높을 것으로 기대한다. 왜냐하면 젊은 남자들보다 좀 더 연령이 있는 구간에서의 남자들은 직장 내에서 승진했을 시기이고, 승진함으로써 소득이 더 높아질 것으로 기대하기 때문이다. 이는 30대 이후의 결혼한 남성 혹은 이혼한 남성에 대에서도 마찬가지다.

6.6 exercise)

a)

제이미 루비스의 시도한 필드에서의 득점 슈팅 중 성공 비율= $\frac{155}{331}=0.4683$, 약 47%이다.

린지 쉬어러의 시도한 필드에서의 득점 슈팅 중 성공 비율= $\frac{91}{191}=0.4764$, 약 48%이다.

b)

제이미의 2점짜리 슈팅의 백분율= $\frac{119}{331}=0.3595$, 약 36%이고, 3점짜리 슈팅의 백분율은 $\frac{36}{331}=0.1088$, 약 11%이다.

린지의 2점짜리 슈팅의 백분율= $\frac{86}{191}=0.4503$이고, 3점짜리 슈팅의 백분율은 $\frac{5}{191}=0.0262$이다.

c)

린지의 2점, 3점짜리 슈팅의 성공 백분율 모두 전체적인 백분율 48%보다 낮은데, 이러한 현상은 백분율 계산 시 분모에 해당하는 모든 슈팅 시행횟수에 있다. 만약 성공한 슈팅만을 분모로(91) 2점짜리와 3점짜리 백분율을 계산하면 각각 0.9451, 0.0549, 즉 94%와 5%의 백분율을 가지고 이 둘의 평균치가 약 49%이다. 이는 전체 득점 슈팅 중 성공비율인 48%와 근사한다. 즉, 2점, 3점짜리 슈팅에 관한 백분율을 고려하지 않았기 때문에 이러한 결과가 나타난다.

8.7 exercise)

Adelaja	Draguljic	Huo	Modur
Ahmadiani	Fernandez	Ippolito	Rettiganti
Barnes	Fox	Jiang	Rodriguez
Bonds	Gao	Jung	Sanchez
Burke	Gemayel	Mani	Sgambellone
Deis	Gupta	Mazzeo	Yajima
Ding	Hernandez		

표 B의 라인 134를 사용하여 위 목록에서 인터뷰 대상자 5명을 선택하고자 한다.
우선 위 목록에서 01…26 까지 분류표기를 첨부한다.

01 Adelaja	02 Draguljic	03 Huo	04 Modur
05 Ahmadiani	06 Fernandez	07 Ippolito	08 Rettiganti
09 Barnes	10 Fox	11 Jiang	12 Rodriguez
13 Bonds	14 Gao	15 Jung	16 Sanchez
17 Burke	18 Gemayel	19 Mani	20 Sgambellone
21 Deis	22 Gupta	23 Mazzeo	24 Yajima
25 Ding	26 Hernandez		

표 B의 라인 134는 다음과 같다.

27816 78416 18329 21337 35213 37741 04312 68508

첫 번째 숫자 그룹을 택하면 27, 81, 67 이고, 이는 위 목록의 분류표기로 사용되지 않았으므로 무시한다. 다시 그 다음 두 번째 숫자 그룹을 택하면 84, 16, 18 이고 84를 제외하고 16, 18의 분류표기를 갖는 사람을 인터뷰 대상자로 택한다. 이제 남은 3명을 뽑기 위하여 그 다음 계속하여 숫자 그룹을 택하면 32, 92, 13, 37, 35, 21, 33, 77, 41, 04 이고, 여기서 13, 21, 04의 분류표기를 갖는 사람을 택하면 된다. 따라서 인터뷰 대상자로 택한 5명은 다음과 같다.

16 Sanchez 18 Gemayel 13 Bonds 21 Deis 04 Modur

8.13 exercise)

첫 번째 기간은 1월 1일부터 부활절 (매년 3월 22일부터 4월 25일 사이 춘분 다음 첫 만월 직후의 일요일) 이다. 그리고 두 번째 기간은 7월 1일부터 8월 31일 까지의 기간이다.
무응답의 비율은 두 번째 기간이 더 높은데, 이는 여름 휴가나 외출이 잦은 시기라서 전화 조사에 응답하지 않았을 것이라 추측된다. 첫 번째 기간은 조사기간이 두 번째 기간보다 더 길었고, 계절의 영향으로 오후 7시부터 10시 사이에 가정 내에 있을 확률이 더 높다고 추측하므로, 무응답의 비율이 두 번째 기간보다 상대적으로 낮은 이유라고 생각할 수 있다.

9.5 exercise)

개체(피실험자) : 소나무 묘목, 요인 : 빛의 양, 처리 : 완전한 빛, 25%의 빛, 5%의 빛,
반응변수 : 연구가 끝난 후 개체인 소나무 묘목을 건조시킨 무게

9.15 exercise)

실험자가 실험을 할 때 이중맹검법을 고려하지 않았기 때문에 편의가 발생할 수 있다.
즉, 실험자는 피실험자가 명상을 하는 그룹과 휴식을 취하는 그룹 중 어느 그룹에 속하는지 알고 있었기 때문에 명상을 한 그룹에 근심 수준을 주관적으로 판단하여 등급을 주거나 무의식적인 기대를 할 수가 있다.

10.7 exercise)

사면체 주사위가 가지는 가장 작은 숫자는 1, 가장 큰 숫자는 4이고 이를 두 번 굴려서 얻은 합계에 1을 더한 총합의 최소치는 1+1+1=3 이 된다. 마찬가지로 총합의 최대치는 4+4+1=9 가 되고, 3부터 9 사이의 숫자는 2+1+1=4, 2+2+1=5, 4+1+1=6, 3+3+1=7, 4+3+1=8 과 같은 경우가 존재하기 때문에 주사위가 낼 수 있는 경우의 수는 3 4 5 6 7 8 9 이다.
사면체 주사위는 1번 던질 때 4가지의 경우의 수를 가지므로 2번을 연달아 던질 때 사면체 주사위가 낼 수 있는 모든 경우의 수는 $4\times4=16$ 이다. 물론 16가지 결과 각각이 무작위이므로 장기적으로 주사위를 굴리는 모든 경우에 $\frac{1}{16}$ 의 확률을 가짐을 가정한다.

우선 합계가 3일 때의 경우의 수를 세면, (1+1)+1인 경우가 유일하다. 따라서 확률은 $\frac{1}{16}$ 이다. 합계가 4일 때의 경우의 수를 세면, (2+1)+1 또는 (1+2)+1 두 가지 경우가 가능하다. 따라서 확률은 $\frac{2}{16}$ 이다. 이처럼 각 합계마다의 경우의 수를 세어 확률을 구하면 확률모형은 다음과 같다.

총합	3	4	5	6	7	8	9
확률	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

10.13 exercise)

a) 주어진 확률 모형의 모든 확률을 합산하여 1이 되는지 확인해본다. 그러면 $0.73+0.06+0.06+0.06+0.04+0.02+0.01+0.02=1.00$ 이므로 이것은 정당한 유한 확률 모형이다.

b) 사건 X는 일주일 중 역도의 웨이트 들기를 하는 일수에 해당하므로 $X<4$ 는 역도의 웨이트 들기를 하는 일수가 4일 미만일 사건을 의미한다. 또한 $P(X<4)$ 는 역도의 웨이트 들기를 하는 일수가 4일 미만일 확률을 의미하므로 이는 4일 미만의 모든 사건의 확률의 총합이다. 따라서 $0.73+0.06+0.06+0.06=0.91$ 이다.

c) 적어도 한번 역도의 웨이트를 드는 사건은 $\{X\geq 1\}$ 이고 확률은 $P(X\geq 1)=0.27$ 이다. 또한, 이는 다시 말해서 한 번도 역도의 웨이트를 들지 않는 사건을 제외한 모든 사건과 같은 의미이므로, 이에 따른 확률은 마찬가지로 $1-P(X=0)=1-0.73=0.27$ 이다.

10.17 exercise)

a) $P(X\geq 3.0)$: B 이상의 학점을 받을 확률, $P(X\geq 3.0)=0.14+0.10+0.08+0.09=0.41$.

b) $\{X<2.7\}$, $P(X<2.7)=P(X\leq 3.0)-0.14-0.13=0.46$.

11.5 exercise)

만약 12명에게만 그런 보험을 판매하게 된다면 12명 중 적어도 1명 이상이 아파트 화재가 일어날 확률이 존재하고, 그것이 일어난다면 보험회사는 12명분으로 얻는 수익보다 훨씬 큰 청구액을 지급해야하므로 보험회사는 손실이 매우 클 수 밖에 없다. 하지만 대수의 법칙을 따라 수천 명에게 보험을 판매하게 된다면 평균 보험 청구액은 $\mu=75$ 에 근사하게 된다. 그러면 화재가 일어났을 때 받은 보험료를 가지고도 거의 확실하게 청구액을 지불할 수 있다고 판단한다.

11.9 exercise)

a) $\bar{x}$ 의 표본분표는 $N(186,41/\sqrt{100})=N(186,41/10)=N(186,4.1)$ 이다. 단위는 mg/dl.

$P(183<x<189)=P(\frac{183-186}{4.1}<Z<\frac{189-186}{4.1})=P(-0.73<Z<0.73)$,

$P(-0.73<Z<0.73)=0.7673-0.2327=0.5346$ 이다.

b) 우선 $n=1000$ 일 때의 $\bar{x}$ 의 표본분표를 구하면 $N(186,1.2965)$ 이다.

$P(183<x<189)=P(\frac{183-186}{1.2965}<Z<\frac{189-186}{1.2965})=P(-2.31<Z<2.31)$,

$P(-2.31<Z<2.31)=0.9896-0.0104=0.9792$ 이다.

11.13 exercise)

평균 손실액 $\mu = 75$, 손실액의 표준편차 $\sigma = 300$ 일 때 10000명의 보험 가입자를 단순 무작위 표본으로 생각하자. 즉 $n = 10000$. 중심극한정리에 따르면 표본평균 $\overline{x}$ 의 표본분포는 $N(75, 300/\sqrt{10000}) = N(75, 300/100) = N(75, 3)$ 이다. 이는 모분포가 한쪽으로 강하게 기울어졌음에도 10000개의 보험에 대한 평균 손실액이 $N(75, 3)$ 으로 근사함을 알려준다. 이 때 평균 손실액이 85달러를 초과할 확률을 구해보면,

$P(x > 85) = P(Z > \frac{85-75}{3}) = P(Z > 3.33) = 1 - 0.9996 = 0.0004$ 이므로 매우 희박하다.

따라서 평균 손실액이 85달러보다 크지 않다는 가정하에 보험료를 안전하게 설정할 수 있다.