출석수업 과제물(평가결과물) 표지(온라인제출용)

교과목명 : 통계로 세상 읽기

학 번 : 202135-367895

성 명 : 김태정

강 의 실 : 지역대학 호

연 락 처 : 010-4172-47516

- 이하 과제 작성

1번문제

개인 - 개인의 삶속에서 선택 및 판단의 순간에서 논리적이고 합리적인 판단의 근거로 이용한다.

기업 - 기업의 운영을 위한 시장분석과 기업전략 수립의 기본자료로 활용한다.

정부 - 정부가 정책을 운영하면서 국가 현황을 파악하고 정책 기획 수립의 기초자료로 활용한다. 정책에 대한 국민 합의 도출과 정책 유효성 평가에 이용한다.

2번문제

자료의 수집 - 관심 대상에 대한 객관적 자료를 수집한다

자료의 정리 요약 및 설명 - 수집한 자료를 정리 요약하여 새로운 정보를 얻어낸다.

자료로부터 결론 도출 - 정보를 기초로 하여 자료구조를 설명하고 분석한 자료를 바탕으로 추측 및 의사결정 방법을 제시한다.

3번문제

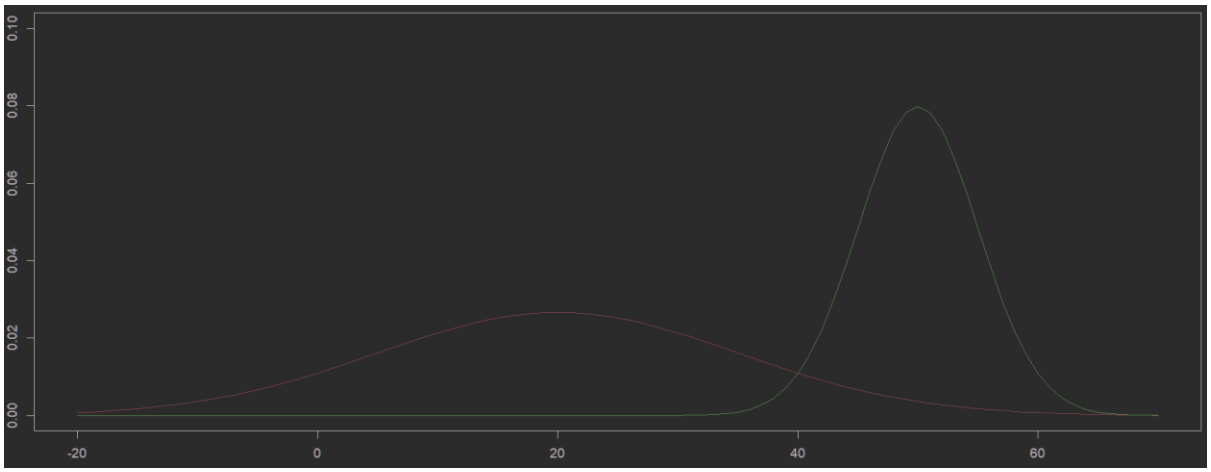
기대값의 공식은 X * P(X)이다. 따라서 아래와 같이 표현된다.

0 * 0.2 + 1 * 0.4 + 2 * 0.2 + 3 * 0.2 = 1.4

4번문제

```
x <- seq(-20, 70)
ylim <- c(0, 0.1)
plot(x, dnorm(x, mean = 20, sd = 15), type = 'l', col = 2, ylim = ylim)
par(new = T)
plot(x, dnorm(x, mean = 50, sd = 5), type = 'l', col = 3, ylim = ylim)
```

그래프는 R로 그렸다



정규분포는 면적의 합계가 1이므로 분산이 커지면 높이는 낮아진다. 반대로 분산이 커질수록 폭은 넓어진다. 평균은 개형에는 영향을 미치지 못하고 좌우로 이동에만 영향을 미친다. 평균이 20에서 50으로 바뀐다면 정규분포의 중심은 20에서 30만큼 이동하게 될 것이고 분산이 15에서 5로 줄어들면 산포도가 줄어들었으니 폭이 좁아진다.

5번문제

1955년 - 1955년부터 1963년은 베이비 붐세대로 출생아 수가 크게 증가하였다. 1943~1953년생은 한국전쟁 등으로 출생아 수가 많이 감소하였던 세대이다. 그리하여 인구 분포는 피라미드 형태를 나타낸다.

2005년 - 지속적으로 출산율이 지속적으로 감소하여 40대 50대가 제일 많고 양쪽으로는 인구가 적은 피라미드 형태가 아닌 항아리 형태이 인구 분포를 보여준다.

2067년 - 인구가 지속적으로 감소하여 2026년에 이미 초고령 사회에 진입하였고 인구 구조는 종형 분포를 보여준다.

6번문제

출생성비 - 여아 100명당 남아 수를 의미한다.

2008년에는 출생성비가 100.43이지만시간이 지날수록 출생성비가 감소하여 2015년을 기점으로 출생성비가 100이하가 되어 여초사회로 진입하게되어 계속해서 출생성비가 감소추세에 있다.

7번문제

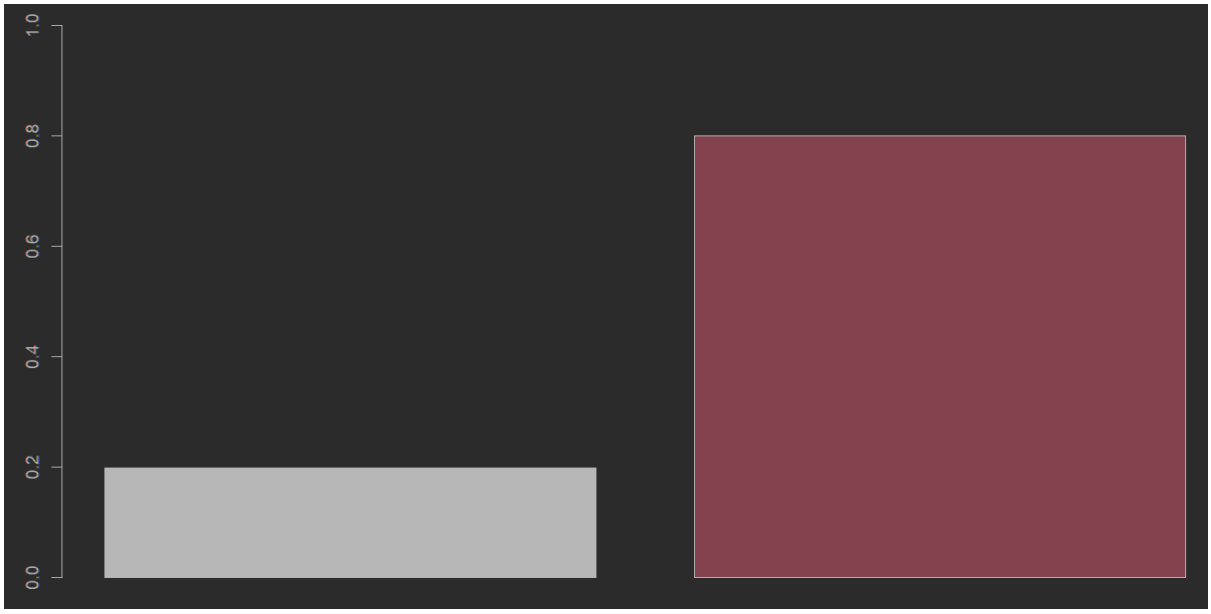
이상치를 구할 때는 보통 상자그림을 사용한다. 자료들로 사분위수를 구한 후 사분위수를 구한다. 그 후 안울타리의 값을 구한다. 안 울타리 안의 값을 정상치로 가정하고 안 울타리 밖의 값을 이상치(outlier) 로 규정한다.

8번문제

```
set.seed(367895)
parameter <- sample(x = 1:1000, size = 1000)
set.seed(NULL)
criteria <- 200
lower <- parameter[parameter < criteria]
upper <- parameter[parameter >= criteria]
lower_rate <- length(lower) / length(parameter)
upper_rate <- length(upper) / length(parameter)
barplot(c(lower_rate, 1 - lower_rate), col = 1:2, ylim = 0:1)

par(mfrow = c(3, 2))

getSample <- function(cnt) {
    sample_data <- sample(x = parameter, cnt)

    sample_data_lower <- sample_data[sample_data < criteria]
    sample_data_lower_rate <- length(sample_data_lower) /
length(sample_data)
    print(sample_data_lower_rate)
    barplot(c(sample_data_lower_rate, 1 - sample_data_lower_rate), col =
1:2, ylim = 0:1, main = cnt)

}

getSample(25)
getSample(50)
getSample(100)
getSample(200)
getSample(300)
getSample(500)
```

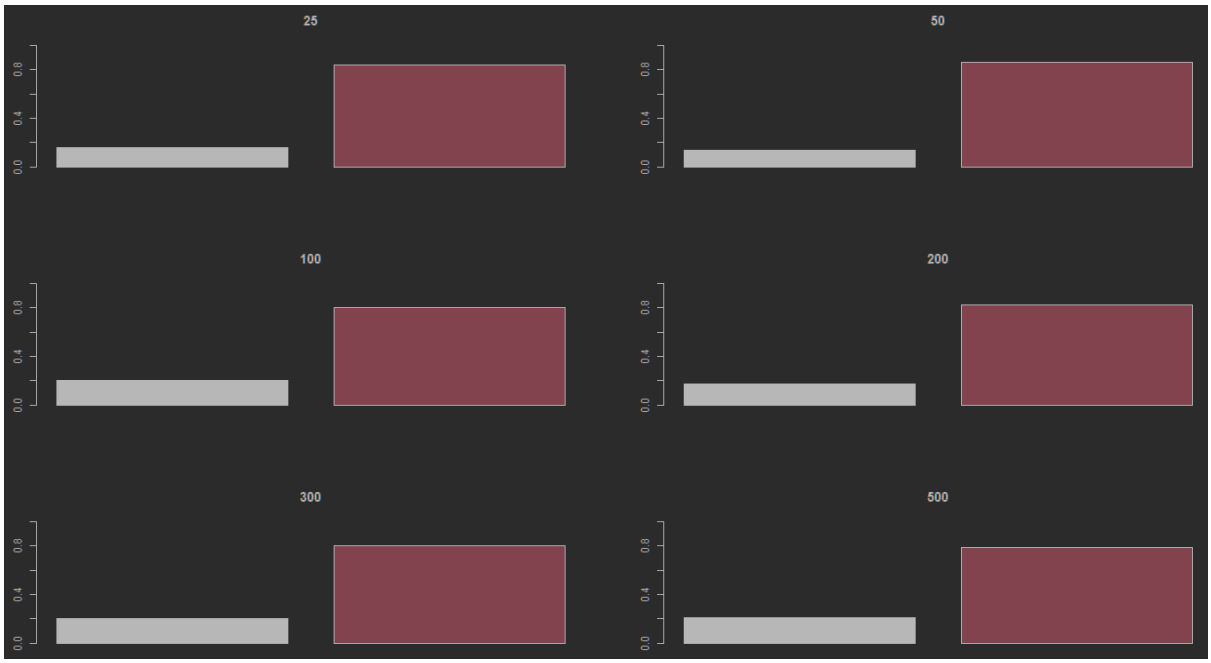
R함수를 사용해서 도표를 그렸다.

1000개의 고정된 샘플(난수 고정)에서 200보다 적은 숫자의 비율을 먼저 막대그래프로 출력하였다.

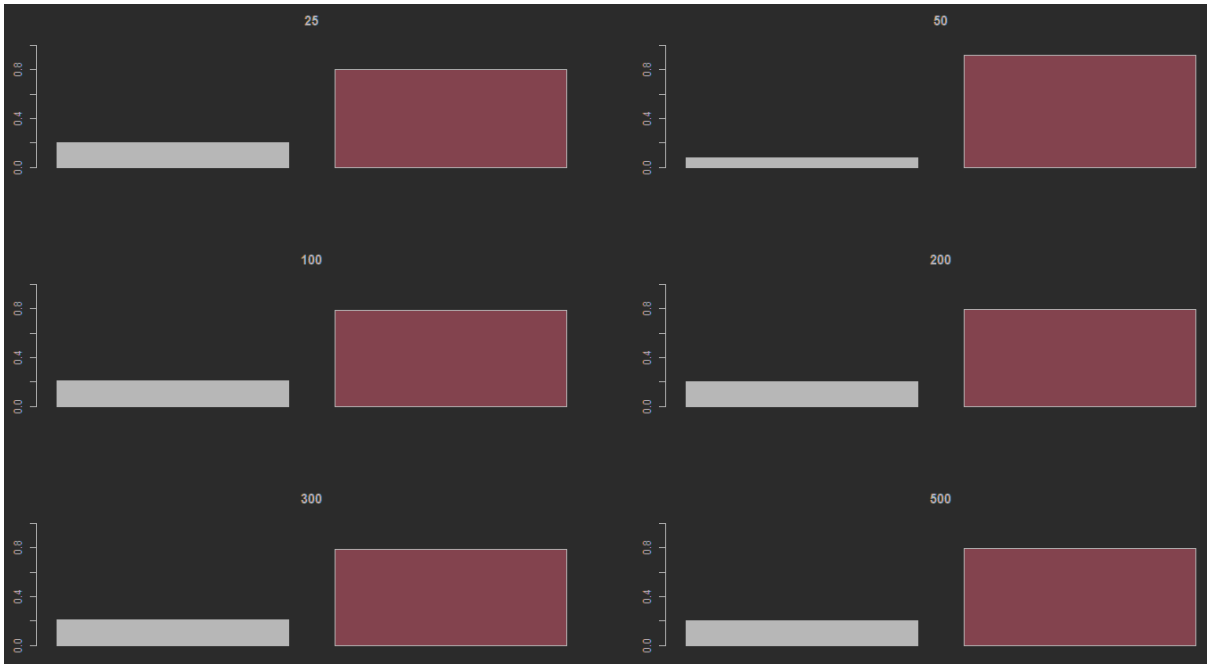


이 경우 1000개에서 복원 추출을 하였기에 (R에서 sample은 기본이 복원추출) 비율은 2:8나온다.

getSample은 그 중에서 다시 특정 개수의 샘플을 뽑아서 비율을 볼 수 있다. 여기서 난수를 다시 시간으로 돌렸으므로 매 시행마다 다른 값이 나올 것이다.



첫번째 시행결과 각각의 차이가 있지만 비율이 2:8에 근사하게 유지되는걸 확인할 수 있다.



두번째 시행에서는 50개를 뽑았을 때 다소 1:9에 가깝게 나왔으나 전체적으로 2:8비율이 유지됨을 확인할 수 있다.

랜덤샘플링의 경우 값이 고루 퍼져있고 제대로 임의로 뽑는다면 적은 샘플로도 모집단을 추정할 수 있음을 알 수 있다.

9번문제

정규분포는 평균을 기점으로 양옆으로 대칭을 이룬다. 정규분포 확률밀도함수의 전체 면적은 1이며 x축의 값은 무한대이다. y의 값, 즉 확률은 절대 0을 가지지 않는다. 분산이 클수록 그래프가 더 퍼지지만 봉우리가 완만해지며 분산이 작을수록 그래프가 모여있지만 봉우리가 뾰족해지는 특징이 있다.

10번문제

자료의 그래프로 인한 표현은 다른 말로 시각화라고도 하며 시각화의 장점은 많은 양의 데이터를 한눈에 볼 수 있으며 데이터의 요약 및 특징을 전문적인 지식이 없어도 확인할 수 있다. 데이터 분석에 널리 사용되는 그래프는 줄기-잎 그래프, 상자그림, 산점도, 히스토그램등이 있으며 막대그래프와 원그래프등도 있다. 그래프들은 질적자료를 나타낼지 양적자료를 나타낼지에 따라서 혹은 산포를 나타낼건지 대표값을 나타낼건지에 따라서 잘 선택해야한다.