

숙제 1

숙제에 대한 답을 pdf로 변환하여 블랙보드에 제출하시오. 답변에 필요한 코드를 포함하여 하나의 pdf 파일로 만들어 제출하시오. (각 문제 10점, 분석문제 50점)

1. n 개의 값, x_1, x_2, \dots, x_n 을 하나의 값 m 으로 대표하고자 한다. 대표값과 실제값들의 제곱오차의 합, $\sum_{i=1}^n (x_i - m)^2$ 을 최소화하는 m 으로 대표하고자 한다. 이때, m 의 값은 어떻게 주어지는가?

2. n 개의 관측 데이터 X_1, X_2, \dots, X_n 은 독립적으로 같은 모집단에서 추출되었다. 모집단의 평균과 분산은 각각 μ, σ^2 라고 하자. 데이터의 관측된 평균을 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 라고 하자.

(1) $E[\bar{X}]$ 는 얼마인가?

(2) $E[\sum_{i=1}^n (X_i - \bar{X})^2]$ 는 얼마인가?

(3) $E[k \cdot \sum_{i=1}^n (X_i - \bar{X})^2] = \sigma^2$ 이 되도록 k 를 정한다면, k 의 값은 얼마인가?

(4) 위의 사실을 이용하여 표본분산의 계산에서 n 이 아니라 $n-1$ 로 나누는 이유를 설명하시오.

3. 다음의 물음에 답하시오.

(1) 0에서 1사이에 n 개의 점이 임의로 분포하고 있다. 이때, 한 점에서 가장 가까운 점까지의 평균적인 거리는 대략적으로 $1/n$ 에 비례함을 설명하시오. 여기서는 n 이 충분히 크다고 생각한다. 엄밀하게 수학적으로 증명할 필요는 없음.

(2) $(0,1)^p$ 로 주어지는 p 차원 공간(한 변의 길이가 1인 p 차원 입방체)에 n 개의 점이 임의로 분포하고 있다. 이때, 한 점에서 가장 가까운 점까지의 평균적인 거리는 얼마인가?

(3) $(0,1)^p$ 로 주어지는 p 차원 공간에 n 개의 점이 임의로 분포하고 있다. 만일, 차원이 두 배가 된다면, 가장 가까운 점까지의 평균적인 거리를 같게 만들기 위해서는 몇 개의 점이 필요한가?

(4) 위의 사실을 이용하여 차원의 저주를 설명하시오.

4. 재원이는 앞면이 나오면 이기고 뒷면이 나오면 지는 동전 던지기 게임을 하고 있다. 10번 던진 결과 앞면이 1번 나오고 뒷면이 9번 나왔다. 재원이는 이 동전이 공정하지 않은 조작된 동전이 아닐까 의심한다. 다음의 물음에 답하시오.

- (1) 공정한 동전이라고 가정할 때, 앞면이 1번 나오고 뒷면이 9번 나올 확률은 얼마인가?
- (2) 공정한 동전이라고 가정할 때, 지금 내가 관측한 상황과 같거나 더 극단적인 결과가 나올 확률은 얼마인가? 다시 말하면, (앞면, 뒷면)이 (1,9), (9,1), (10,0), (0,10)으로 나올 확률은 얼마인가? 이 확률을 바탕으로 이 동전은 공정하지 않은 동전이라고 말할 수 있는가?
- (3) 앞면 1번 뒷면 9번이라는 관측결과에 대하여, 이 동전이 공정한 동전인지를 통계적으로 확인하고자 한다. 동전이 앞면이 나올 확률을 θ 라고 할 때, 귀무가설과 대립가설을 세우고 p값을 구하시오. 유의수준 0.05에서 이 가설을 검정하시오.

5. 직종(White/Blue)과 소비성향(A/B/C)에 대한 크로스테이블에 대하여 카이제곱 검정을 수행하고자 한다.

	A	B	C	Total
White	90	70	140	300
Blue	60	80	60	200
Total	150	150	200	500

다음의 물음에 답하시오.

- (1) 직종과 소비성향이 완전히 독립적이라고 가정할 때, White이면서 C인 사람들이 평균적으로 관측되는 숫자는 얼마인가?
- (2) 카이제곱 검정을 위한 Q값을 계산하여, 소수점이하 둘째자리까지 반올림하여 나타내시오.

6. (분석문제) data99_churn_train.csv는 통신망 가입자의 서비스탈퇴(churn) 여부에 대한 데이터이다. 데이터에서 가장 마지막 열에 있는 churn 변수가 관심 변수이고, 나머지는 이를 설명하기 위한 변수이다. 이 데이터를 이용하여 아래의 질문에 답하시오. 답변에 필요한 코드를 첨부하여 제출하시오.

(1) 이 데이터는 모두 몇 개의 변수와 몇 개의 표본으로 이루어져 있는가?

(2) 각 변수에 대하여 변수의 이름을 이용하여 변수의 의미를 설명하고, 각 변수가 수치형인지 범주형인지 확인하시오.

(3) 결측치를 확인하시오. 총 몇 개의 결측치가 존재하는가? 결측치를 갖고 있는 변수는 몇 개인가? 결측치를 포함하는 표본은 몇 개인가? 결측치를 갖는 표본을 제거하시오.

(4) 각 변수에 대하여 요약하시오. 수치형 변수는 평균과 분산을 제공하시오. 범주형 변수에 대해서는 도수분포를 제공하시오. 수치형 변수에 대해서는 히스토그램을 제공하시오.

(5) (2)와 (4)의 결과를 바탕으로 customerID를 후속 분석에서 제외해야하는 이유를 설명하시오.

(6) Churn와 다른 모든 변수(customerID 제외) 사이에서, 수치-범주의 경우에는 boxplot을 그리고, 범주-범주인 경우에는 크로스테이블을 구하시오.

(7) Churn와 다른 모든 변수(customerID 제외) 사이에서 적절한 관계검정을 수행하고 p값을 구하시오. p값 기준으로 가장 중요한 변수 3개와 가장 중요하지 않은 변수 3개를 각각 선정하시오.