19101185 강동현 데이터 분석 Assignment 02

1. 다음 테이블을 보고 , 물음에 답하는 문제이다.

(a). 항공기 사고 표본의 평균을 찾는 문제이다. 사고가 0번인 경우 1개 , 1번인 경우 3개 , 2번인 경우 6개 , 3번인 경우 4개 , 4번인 경우 5개 , 6번인 경우 2개 , 11번인 경우 1개가 존재 하므로 평균은 약 3.18번 이다.

(b). 항공기 사고 표본에서 중앙값을 찾는 문제이다. 즉 데이터를 사고수를 기준으로 오름차순으로 배열 했을 때 , 해당하는 데이터를 찾으면 된다. 22개의 표본중 중간값의 인덱스는 11번과 12번 이다. 따라서 중앙값은 3 이다.

(c). 항공기 사고 표본으로 표준편차를 구하는 문제이다. 분산을 구한후 루트를 씌우면 표준편차를 구할 수 있다. 따라서 제곱의 평균에서 평균의 제곱을 빼 표준편차를 구할 수 있다. 따라서 약 15.272727 – 약 10.123966 = 5.148761 가 분산이고 , 루트를 씌운 값은 약 2.269088가 표준편차이다.

2. A마을의 여자 몸무게 평균을 E[AW] , B마을은 E[BW] 이라 하자. 또한 A마을의 남자 몸무게 평균을 E[AM] , B마을은 E[BM]이라 하자. A마을 성인의 몸무게 평균은 ( E[AW] + E[AM] ) / 2 이고 , B마을은 ( E[BW] +E[BM] ) / 2이다. E[AW] > E[BW] 이고 , E[AM] > E[BM] 이므로 A마을 성인의 몸무게 평균이 B마을 성인의 몸무게 평균보다 큰 것을 알 수 있다.

3. 키와 초봉에 대한 correlation coefficient를 찾는 문제이다. 따라서 키와 초봉의 평균을 구할 필요가 있다. 키의 평균은 약 70.3 이고 , 초봉의 평균은 97.5 이다. 따라서 각각의 표본 값에서 평균을 뺸 값들을 활용하여 correlation coefficient를 구할 수 있다. 계산결과 식의 분자부분은 272 , 분모부분은 186.68 \* 1693에 루트를 씌운 값으로 나왔다. 따라서 분모부분은 약 562.18 이고 계산하게 되면 r은 약 0.483 으로 나오게 된다. R이 0.483의 값을 가지므로 이 표본에서 키와 초봉은 보통의 양적 선형관계를 가진다고 볼 수 있다.

4. 조건부 확률에 관한 문제이다. 먼저 2개의 전자레인지를 동일한 공장에서 구입 하므로 구입한 전자레인지가 A공장 제품일 확률은 1/2 , B공장 제품일 확률은 1/2 이다. A공장과 B공장 모두 전자레인지 불량인 사건은 다른 전자레인지가 불량인 것에 영향을 받지 않으므로 독립 사건이다. 따라서 A공장에서 첫번째 전자레인지전 불량일 때 , 두번째 전자레인지가 불량일 확률은 0.05 , B공장에서 첫번째 전자레인지전 불량일 때 , 두번째 전자레인지가 불량일 확률은 0.01 이다. 따라서 정답은 0.05 / 2 + 0.01 / 2 = 0.03 이다.

5. 모든 경우에 대하여 수학적으로 , 조건부 확률을 이용하여 살펴보면 죄수의 논리에 대하여 추리가 가능하다. 맨 먼저 죄수의 번호를 A , B , C 라 하자. 죄수 A는 이들 중 2명이 사형 한다는 사실과 한명은 풀려난다는 사실을 알고 있다. 따라서 가능한 사형수의 조합은 AB조합 , AC조합 , BC조합 모두 3가지다. 일단 먼저 A가 간수에게 누군가에 대한 정보를 듣기 전까지는 A가 사형당할 확률은 2/3다.

우리는 A가 간수에게 풀려날 한명을 들었다고 가정해보자. 이 경우 본인에 대한 정보는 알려주지 않으므로 , B 혹은 C 일텐데 둘중 한명이 선택되는 것은 1/2의 확률로 나뉠뿐 , 누군가 풀려나는 사실을 알 때 , A가 사형당할 확률은 같게 된다. 즉 한명의 예시만 가지고 나머지 경우에 대해서도 확률을 알 수 있다. 따라서 B를 예시로 들어 살펴보자.

만약 A가 간수에게 B에 대한 정보를 들은 이후 위의 사형수 조합에서 AC조합은 사라지게 된다.

따라서 A는 AB조합과 BC조합만을 고려하면 된다. 만약 AB조합의 경우 간수는 B혹은 C에 대한 정보를 알려주어야 하기 때문에 알려줄 수 있는 정보는 B의 정보 밖에 없다.

BC조합의 경우 간수는 B와 C에 대한 정보중 선택하여 A에게 알려줄수 있으므로 2가지의 선택지가 생긴다. 하지만 간수는 A에게 B에 대한 정보를 주었으므로 1/2의 확률을 통해 A에게 B에 대한 정보를 주었다는 것을 알 수 있다.

따라서 정리 해보면 각각 조합의 경우에서 B에 대한 정보를 들었을 때의 A가 사형당할 확률을 정리하여 보면 다음과 같다. AB의 경우 무조건 B에 대한 정보를 말할 수밖에 없으므로 1 , AC의 경우 , 일어나지 않으므로 0 이다. BC의 경우 1/2의 확률로 B를 택하였기 때문에 1/2이다. 이 3가지 경우중 B에 대한 정보를 듣고 , A가 죽는 조합은 AB조합 한가지 밖에 없다. 따라서 B에 대한정보를 들었을 때를 조건으로 하여 위 사건이 발생했을 때 A가 사형당하는 조건부 확률을 계산할 수 있다. 조건부 확률에 따라 계산하면 1 / (1 + 1/2 + 0) 이고 이는 2/3입니다. ( 각각의 3가지 조합이 나올 확률은 1/3로 모두 같아 계산과정에서 생략 )

따라 우리는 B의 예시를 통해 B의 정보를 들었을 때 , A가 사형당할 확률은 2/3이라는 것을 알았다. C의 경우 B의 예시와 같게 진행될 것이므로 C의 정보를 들었을 때도 2/3의 확률로 A가 사형당한다는 것을 알았다. B , C의 정보를 각각 선택하여 알려줄 확률은 1/2 이므로 정리하면 1/2 \* 2/3 + 1/2 \* 2/3 = 2/3이다.

즉 누구의 정보를 듣던지 간에 정보를 듣기전과 들은 후 모두 확률이 2/3으로 같으므로 , 죄수의 논리는 틀렸다고 볼 수 있다.

6. 기상학자들이 내일 비가 올지 , 오지 않을지에 대한 확률을 제시하고 이에 따라 점수를 얻게 된다. 따라서 점수를 극대화 하기 위해서는 점수의 기댓값을 높이면 된다. 점수에 대한 기댓값을 정리하면 ( 1 + 2p – 2p^2 ) / 2 이다. 이를 미분하게 되면 ( -4p + 2 ) / 2 이고 p가 0.5 일때 가장 기댓값이 높음을 알 수 있다. 따라서 점수를 가장 높게 받기 위해서는 확률을 늘 0.5로 추정하면 된다. 실제 예시를 통해서 보면 p가 0.7일 경우 점수의 기댓값은 0.71 , 0.3인 경우 점수의 기댓값은 0.71 , 0.5인 경우 점수의 기댓값은 0.75로 p가 0.5인 경우 0.3 , 0.7보다 높은 점수 기댓값을 보여줌을 알 수 있다.

7. E[X] 가 2 이고 , E[X^2] = 8 일때 , 식을 계산하는 문제이다.

(a). E[(2+4X)^2] 을 계산하는 문제이다. 안의 식을 전개하면 E[4+16X+16X^2]이고 이는 E[16X+4] + E[16X^2]과 같다. 따라서 E[16X+4]는 16E[X] + 4 즉 36이고 , E[16X^2]는 16E[X^2] , 즉 128이다. 답은 128 + 36 = 164이다.

(b). E[X^2 + (X+1)^2] 을 계산하는 문제이다. 안의 식을 전개하면 E[2X^2 + 2X + 1] 이고 이는 E[2X^2] + E[2X + 1] 과 같다. 따라서 E[2X^2] = 16 , E[2X + 1] = 5 이므로 16 + 5 즉 21이 답이다.

8. COV(X1 – X2 , X1 + X2) = 0 을 보이는 문제이다. COV(X1 – X2 , X1 + X2)는 COV(X1 , X1 – X2) + COV(X2 ,X1 – X2) 로 나눌 수 있다. 또 이는 COV(X1 – X2 , X1) + COV(X1 – X2 , X2) 로도 바꿀 수 있다. 이는 다시 COV(X1 ,X1) + COV(-X2 , X1) + COV(X1 , X2) + COV(-X2 , X2) 로 정리 가능하다. 이를 다시 표현하면 COV(X1 ,X1) - COV(X2 , X1) + COV(X2 , X1) - COV(X2 , X2) 이고 , 정리하면 COV(X1 ,X1) - COV(X2 , X2) 이다. 이는 VAR(X1) – VAR(X2) 이다. 문제에서 X1 , X2는 same probability distribution function을 가진다고 하였으므로 VAR(X1) = VAR(X2) 이다. 따라서 . COV(X1 – X2 , X1 + X2) = 0 임을 알 수 있다.

9. 다음 관계식을 보고 물음에 대해 답하는 문제이다.

(a). E[Y]의 관점에서 E[X]를 표현하는 문제이다. Fx와 Fy의 관계가 다음과 같으므로 , E[X] = E[(X-a)/b] 이다. E[(X-a)/b]를 다시 (E[X] – a) / b 이고 이는 E[Y]와 같다. 따라서 식을 E[Y]의 관점에서 식을 정리하면 b \* E[Y] + a = E[X] 이다.

(b). Var[Y]의 관점에서 Var[X]를 표현하는 문제이다. Var[X] = Var[(X-a)/b] 이다. Var[(X-a)/b]를 다시 표현하면 , Var[X] / b^2 이고 이는 Var[Y]와 같다. 따라서 Var[Y]의 관점에서 식을 정리하면 ( b^2 ) \* Var[Y] = Var[X] 이다.

10. X1과 X2를 각각 독립적인 랜덤 변수로 두고 , 평균은 10 , 분산은 같도록 두었을 때 , 어느 확률이 더 큰지 판단하는 문제이다.

(a). X1 , X2가 독립변수 이므로 , E[X1+X2]는 20이라고 볼 수 있다. X1 , X2의 분산은 같으므로 같은 확률 분포를 가진다는 것을 알 수 있다. 따라서 P(X1 > 15) 와 P ( X1 + X2 > 25 ) 모두 평균에서 5만큼 떨어진 부분에서의 확률을 묻고 있으므로 두 값은 서로 같다.

(b). (a)번과 마찬가지로 풀면 , P(X1 > 15) 는 평균에서 5만큼 떨어진 부분부터의 확률을 묻고 있고 , P(X1+X2 > 30)은 평균에서부터 10만큼 떨어진 부분부터의 확률을 묻고 있다. 따라서 P(X1> 15)가 더 크다는 것을 알 수 있다.

(c). P(X1 + X2 > x) = P(X1 > 15) 를 만족하는 x를 구하는 문제이다. 우리는 (a)번에서 이미 구하였으므로 답이 25임을 알 수 있다.

11. Φ 이 standard normal distribution function을 따른다고 하자. 이때 X의 특징 , 즉 distribution , 평균 , 분산을 서술하는 문제이다. Φ 가 standard normal distribution function을 따르고 , 이 Φ에 linear한 가공이 이루어 졌으므로 , X역시 standard normal distribution function을 따른다. 이후 식을 보면 ( X – a ) / b의 방식으로 가공이 이루어 졌으므로 평균의 경우는 E[X]/b – a/b의 형태를 보일 것이고 , 분산의 경우는 Var(X) / b^2의 형태를 보여줄 것이다.