권이태

Problem. 1. *저가커피 브랜드에서 원두의 볶는 시간(A)과 원산지(B)에 따른 고객만족도를 확인하고자 한다. 다음은 반복이 없는 이원배치법을 이용한 실험결과이다. 데이터를 옮기는 과정에서 조교의 실수로 한 수준의 조합에 누락이 발생하였다.*

	A_1	A_2	A_3	계
B_1	?	9	12	?
B_2	3	6	8	17
B_3	3	4	11	18
계	?	19	31	?

(a) *수준 A_1B_1 에서의 결측치 ?를 y로 쓰자. 주어진 데이터를 이용하여, 오차제곱합 S_E를 y에 대한 함수로 표현하여라.*

(b) *S_E를 최소화하는 y의 값을 구하고, 이 값이 결측치를 대체하는 값으로 사용될 때의 S_E를 구하여라.*

(c) (b)에서 얻은 y로 결측치를 대체하였을 때의, 아래의 분산분석표를 작성하여라. -표시가 되어있는 칸은 제외하고 작성하여도 괜찮다.

요인	S	ϕ(df)	V	F_0
A				
B				
E				-
T			-	-

(d) F_0를 통하여 인자 A와 B가 유의수준 0.05에서 유의한지 확인하고, 그 결과를 참고하여 결측치가 있는 인자수준 A_1B_1에서의 모평균의 95퍼센트 신뢰구간을 구하여라. 이때 $F_{0.05}(2,3) = 9.552, F_{0.05}(2,4) = 6.944, t_{0.025}(3) = 2.354, t_{0.025}(4) = 2.132$임을 이용하여도 괜찮다.

Problem. 2. *A, B, C 모두 2수준 모수인자이며 이들이 합금의 강도에 미치는 영향을 검토하고자 한다. 실험은 하루에 4번밖에 할 수 없어서, 실험일을 블록으로 취한 뒤 3회 반복실험하였으며, 각 반복에서 ABC, AC, BC를 블록과 교락시켜* 2^3 *인자 부분교락실험을 진행하였다. 이때 (1)은 모든 인자수준이 0인 상태이다.*

반복	I (ABC 교락)		II (AC 교락)		III (BC 교락)	
블럭	1	2	3	4	5	6
배치=결과	a=50 b=67 abc=82 c=62	ab=78 ac=82 bc=70 (1)=38	(1)=40 b=51 ac=74 abc=61	a=69 c=72 bc=91 ab=59	bc=82 a=74 (1)=41 abc=67	ab=40 c=55 ac=79 b=60
계	261	268	226	291	264	234
	529		517		498	
	1544					

(a) *위의 원자료표를 바탕으로 아래의 분산분석표를 채워라.*

요인	S	ϕ(df)	V	F_0
블럭			-	-
반복				
반복내 블럭				-
A				
B				
C				
$A \times B$				
$A \times C$				
$B \times C$				
$A \times B \times C$				
E				-
T			-	-

(b) *모든 요인이 존재함을 가정하고 최적수준을 찾은 뒤, 최적수준에서 모평균의 95퍼센트 신뢰구간을 구하시오.* $t_{0.025}(11) = 2.201, t_{0.025}(13) = 2.161$ *임을 이용하여도 좋다.*

(c) F_0*값이 1 미만인 교호작용을 오차항에 풀링한 뒤 다시 분석을 진행하였다. 최적수준이 (b)와 동일한지 논하고, 같다면 최적수준에서 모평균의 95퍼센트 신뢰구간을 새로 구하시오.*

Problem. 3. *특정 함수의 최대값을 찾기 위한 MM(miniorization-maximization) 알고리즘을 고려하자.*

(a) *최대화하고자 하는 함수를* $f(\mathbf{x})$, *대리함수(surrogate function)를* $g(\mathbf{x}|\mathbf{x}^{(t)})$*라고 할 때,* t*번째 반복에서 얻은* $\mathbf{x}^{(t)}$*에 대하여* $g(\mathbf{x}|\mathbf{x}^{t})$*가 만족해야 하는 두 조건을 쓰고 식으로 표현하여라.*

(b) *EM 알고리즘은 MM 알고리즘의 특정한 예시로, 로그가능도를 최대화하는 알고리즘이다. 관측된 자료를* $\mathbf{O}$, *관측되지 않은 잠재변수를* $\mathbf{Z}$, *모수를* θ*라 할 때 로그가능도* $l(\theta)$*는*

$$l(\theta) = \log \int p_\theta(\mathbf{o}, \mathbf{z}) d\mathbf{z}$$

로 쓸 수 있다. 이때 $p_\theta(\cdot)$*는 식별가능성을 만족하여* θ*와* $p_\theta(\cdot)$*가 일대일로 대응될 수 있는 확률밀도함수이며, 모수에 무관하게 동일한 분포의 토대를 가진다고 하자. 이를 이용하여*

$$l(\theta) \geq \mathbb{E}^{(t)}_{\mathbf{Z}|\mathbf{O}}[\log p_\theta(\mathbf{o}, \mathbf{z})] - \int \log[p_{\theta^{(t)}}(\mathbf{z}|\mathbf{o})] p_{\theta^{(t)}}(\mathbf{z}|\mathbf{o}) d\mathbf{z}$$

임을 보이고, 우변의 식이 $l(\theta)$*에 대한 (a)의 대리함수 조건을 만족함을 보여라.*

(c) *(b)에서,*

$$l(\theta) - (\mathbb{E}^{(t)}_{\mathbf{Z}|\mathbf{O}}[\log p_\theta(\mathbf{o}, \mathbf{z})] - \int \log[p_{\theta^{(t)}}(\mathbf{z}|\mathbf{o})] p_{\theta^{(t)}}(\mathbf{z}|\mathbf{o}) d\mathbf{z}) \geq 0$$

가 $KL_{\mathbf{Z}|\mathbf{O}}(\theta^{(t)}, \theta)$*와 동일함을 밝혀라. 이를 통해 (b)의 부등식 동치조건과* $KL(\theta^{(t)}, \theta) = 0$*일 조건이 동일함을 확인하여라.* $p_\theta(\cdot)$*에 적절한 조건을 가하여* θ*와* $\theta^{(t)}$*만을 이용해 표현하여라. 해당 조건을 상세히 밝힐 필요는 없다.(즉, '좋은 조건 하에서'라고 표현하여도 괜찮다.)*

(d) $Q(\theta|\theta^{(t)}) = \mathbb{E}^{(t)}_{\mathbf{Z}|\mathbf{O}}[\log p_\theta(\mathbf{o}, \mathbf{z})]$*로 정의할 때,* Q*는 (a)의 조건을 만족하지 못할 수 있음에도 EM 알고리즘에서*

$$\theta^{(t+1)} = \mathrm{argmax}_\theta Q(\theta|\theta^{(t)})$$

로 θ*를 업데이트하는, 혹은 할 수 있는 이유를 말하시오. 또한 이로써 얻은* $\theta^{(t+1)}$*이*

$$l(\theta^{(t+1)}) \geq l(\theta^{(t)})$$

를 만족함을 보이시오. 이때 Q*가 (a)의 조건을 만족하지 못할 수 있음을 보이지는 않아도 괜찮다.*

Problem. 4. *한국이는 최소값이 0, 최대값이 1인 균등분포로부터 임의의 수를 뽑아 더해나간다. 이때 그 합이 1이 넘는 순간 시행을 멈추기로 하고 이를 기록하려 한다. 기록되는 수는 연속확률변수* V*로 취급할 수 있다. 예를 들어 한국이가 처음 뽑기에서 0.6을, 둘째 뽑기에서 0.5를 뽑았다면 시행을 멈추고* $V = 2$*를 기록한다. 처음 뽑기에서 0.2를, 둘째 뽑기에서 0.5를, 셋째 뽑기에서 0.9를 뽑았다면 시행을 멈추고* $V = 3$*을 기록한다.* $\mathbb{E}[V]$*를 구하여라.*
(Hint: 합이 a*가 넘는 순간 시행을 멈추는 상황에서 기록되는 수를* V_a*라 할 때,* V_a*의 기댓값은* a*의 함수이다. 해당 함수에 대한 방정식을 풀어 그를 얻고,* $a = 1$*을 대입하라.)*