보건통계학 기말고사

이지은 222DBG04 빅데이터 분석학과

1. ‘표준편차(Standard Deviation, SD)’와 ‘표준오차(Standard Error, SE)’에 대하여 각각 설명하시오. [15 점]

표준편차(SD)는 개별 관측값의 표준 편차를 의미하는것이고, 평균으로부터 각 관측치가 얼마나 떨어졌는지를 의미하는 지표이다. 표준오차(SE)는 한 모집단에서 추출된 표본들간의 변동성을 의미하며 , 모집단의 모수로부터 표본 통계량이 얼마나 떨어져있는지를 파악하는 지표이다.

2. ‘95% 신뢰구간(95% Confidence Interval, 95% CI)’의 정의를 기술하고, 모집단으로부터 무작위 추출한 113 명의 평균 수축기 혈압이 123.6 mmHg 이고 95% 신뢰구간이 121.2 – 126.0 mmHg 일 때 표본평균 123.6 mmHg 와 95% CI 121.2 – 126.0 mmHg 의 관계가 가지는 의미를 해석하시오. [15 점]

95% 신뢰구간은 통계량의 95%가 위치하는 구간이다. 즉 모집단의 참값을 포함할 확률이 95%이고, 113개의 표본 신뢰구간 95%가 모집단의 참값을 포함한다는 의미이다. 즉 113명의 표본에서의 모집단의 평균 수축기 혈압 에 대한 추정치가 123.6mmHg이며, 신뢰구간인 121.2-126사이에 모집단의 평균이 위치할 확률이 95%이다.

3. 다중공선성(multicollinearity)이 무엇인지 설명하고, 다중공선성을 해결하기 위한 방법을 제시하시오. [15 점]

다중공선성이란 회귀분석에서 독립변수간에 강한 상관관계가 존재하여 모델의 해석이 어려워지고, 예측력이 감소하는 문제이다. 강한 상관성을 가지는 두 변수간 한가지 변수만 선택하거나 변수에 정규화나 로그 변환으로 이러한 문제를 해결할 수 있다.

4. 방사선 노출(x)과 백혈병 발생(Y)의 관련성에 대한 회귀분석을 시행하고자 한다. 종속변수는 백혈병 발생확률 p, 독립변수는 방사선 노출량 x로 한다. p는 확률로서 0 부터 1 까지의 값을 가지며, x는 continuous data 로서 단위는 Sv(시버트)이다. Linear regression 모델 $p = \alpha + \beta x$ 에서의 회귀계수 β, 그리고 Logistic regression 모델 $\ln\left(\frac{p}{1-p}\right) = \alpha' + \beta' x$ 에서의 회귀계수 β' 를 어떻게 해석할 수 있는지 각각 설명하시오 (β와 β'의 의미를 각각 설명하시오). [20 점]

B는 회귀직선의 기울기인 회귀계수이며 독립변수와 종속변수 사이의 관계이다. x(방사선노출량)이 1sv 증가할때, 백혈병 발생확률은 B만큼 변화한다. 정확히는 x가 1단위 상승할때 y의 변화량이다.
B'은 로지스틱 회귀계수이며 로짓 변환된 종속 변수와 독립 변수 간의 관계이다. x가 1sv 증가할때, ln (p/(1-p))가 B'만큼 변화한다. 정확히 말하면 x가 한단위 변화할때 ln (p/(1-p))의 차이이다.

5. 고혈압의 위험요인에 대한 환자-대조군 연구(case-control study)를 수행하여 다음과 같은 결과를 얻었다.

Table. Odds ratios from logistic regression analysis

Risk factor	Odds Ratio	95% CI	
		Lower	Upper
Milk consumption (mL/day)	0.69	0.57	0.82
Alcohol consumption (gram/day)	3.23	2.31	4.50
Smoking			
Non-smoker	1.00		
Smoker	1.06	0.87	1.30
Family history of hypertension			
No	1.00		
Yes	1.13	0.78	1.63
Salt intake (gram/day)	1.75	1.28	2.38
Body mass index (kg/m^2)	1.51	1.12	1.96

(1) 고혈압과 유의하게 관련된 변수는 무엇인지 나열하고 그 판단의 근거를 제시하시오. [10 점]

신뢰구간에서 1을 포함한다면 해당변수의 영향력이 유의하지 않다는 것을 의미한다. 고혈압과 유의하게 관련된 변수는 신뢰구간에서 1이 포함되지 않은 변수인 'Milk consumption', 'Alcohol consumption', 'Salt intake','Body mass index'이다.

(2) 고혈압과 유의하게 관련된 변수들의 Odds Ratio 값의 의미를 각각 해석하시오. [10 점]

우유소비가 1ml증가할 수록 고혈압 확률이 기존확률의 0.69배로 하락한다. 음주량이 1g 상승할수록 고혈압 확률이 기존의 3.23배로 증가한다. 염분섭취량이 1g 더 상승할 수록 고혈압 확률이 기존의 1.75배로 증가한다. BMI지수가 1 상승할 수록 고혈압 확률이 기존의 1.51배로 증가한다. (이때 '기존'은 변수들이 한 단위 변화하기 이전의 고혈압 발생 확률을 의미한다.)

6. 통계적으로 유의하다고 발표된 결과가 실제로는 참이 아닌 것으로 판명될 가능성을 높이는 요인을 5 가지 이상 나열하시오. [15 점]

1. 표본의 크기가 작다 2. 보고된 효과가 작다. 3. 통계적으로 뚜렷한 비교결과만 제시한다
4.데이터를 계획없이 유동적으로 분석한다. 5.금전적인 동기로 비뚤어진 연구자 6. 경쟁이 치열한 분야