제 1 문 (15점)

현재 어떤 메일시스템에 수신되는 메일 중 40%가 스팸메일이고 나머지는 정상메일이라고 한다. 스팸메일 중 제목에 "A"라는 단어가 있는 메일은 25%이고 "A"와 "B" 두 단어가 모두 있는 메일은 20%라고 한다. 정상메일 중 제목에 "A"가 있는 경우는 5%이고 두 단어가 모두 있는 메일은 2%라고 한다면 아래의 물음에 답하여라.

(1) 제목에 "B" 단어는 없고 "A" 단어만 들어 있는 메일은 전체 메일 중 %인지 구하라. (5점)

(2) 제목에 "A"와 "B" 단어가 모두 있는 메일을 수신했다면 그 메일이 스팸메일일 확률을 구하라. (단, 분수형태로 표시할 것) (4점)

(3) 향후 20년 동안 전체메일에서 스팸메일이 차지하는 비율은 매년 2% 포인트씩 증가한다고 하자. 매년 스팸메일과 정상메일에서 제목에 "A"와 "B"가 들어가는 비율에는 변화가 없다고 할 때, 앞으로 몇 년 후부터 전체메일에서 제목에 "A" 단어가 들어가는 메일이 15% 이상 되는지를 구하라. (6점)

제 2 문 (10점)

서울지역의 초등학교를 다니는 6학년 학생들의 학업성취도를 알아보기 위해 다음과 같은 방법으로 자료를 수집하려고 한다. 이를 통해 초등학교의 평균학업성취도를 비교한다고 했을 때 통계적 관점에서 어떤 공통점과 차이점이 있는지를 기술하라.

방법 ①	서울지역 모든 초등학교에서 6학년 전체학생들을 대상으로 학업성취도 자료를 얻음
방법 ②	서울지역 모든 초등학교에서 몇 명의 6학년 학생들을 무작위로 추출하여 학업성취도 자료를 얻음
방법 ③	서울지역 초등학교 몇 곳을 무작위로 선택하고 선택된 학교에서 몇 명의 6학년 학생들을 무작위로 추출하여 자료를 얻음
방법 ④	조사자가 관심을 가지는 초등학교 몇 곳에서 몇 명의 6학년 학생들을 무작위로 추출하여 학업성취도 자료를 얻음

제 3 문 (10점)

A 지역단체장이 제안한 새로운 지역개발 계획에 대한 지역 주민의 지지율을 알아보기 위하여 한 여론조사업체가 유선전화를 통하여 표본을 추출하려 한다. 이 때 추출된 표본은 모두 지지여부에 대한 응답을 한다고 하자. 95% 신뢰수준(confidence level)에서 지역 주민의 지지율을 허용오차(allowable error) ±3.92% 이내에서 얻기 위한 적정 표본크기 결정방법에 대하여 논하고 표본의 크기를 구하라.
(여기서, $Z \sim N(0,1)$이면, $P(Z \geq z_\alpha) = \alpha$에서 $z_{0.05} = 1.645$, $z_{0.025} = 1.96$이다.)

제 4 문 (15점)

단순선형 회귀(simple linear regression) 모형을 고려하자.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ...n$$

β_0 와 β_1 의 통상최소제곱추정치(ordinary least squared estimate)는 다음과 같이 b_0 와 b_1으로 구해진다

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

기울기의 추정치 b_1과 표본상관계수(sample correlation coefficient) r이 같을 수 있는 상황에 대하여 논하라.

여기서 $\bar{x}$와 $\bar{y}$는 x 와 y의 표본평균이고, r은 다음과 같다.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$