표본론 2022년 중간시험

[1] (20점) 모집단이 {1, 3, 4, 6, 7, 10, 11, 15, 16, 20}으로서 10개의 원소로 이루어져 있다고 하자.

[1-1] 5개를 단순확률추출(simple random sampling)한 표본이 {1, 4, 7, 15, 16}이라고 하자. 표본평균을 구하고 표본평균의 분산을 추정하시오. 이때 모집단을 구성하는 10개의 값을 모른다고 가정하시오.

[1-2] 5개를 단순확률추출해서 얻은 표본평균을 $\bar{y}$라고 할 때 $\bar{y}$의 기댓값 $E(\bar{y})$와 참분산 $V(\bar{y})$를 각각 구하시오.

[2] (30점) 전화를 이용한 선거예측조사에서 무선전화면접 방식과 무선전화 자동응답시스템(ARS) 방식 중에서 어떤 방식으로 조사했느냐에 따라 조사결과가 달라진다는 주장이 있다. 이를 확인하기 위해 2022년 2월 교통방송(TBS)의 의뢰를 받아 한국사회여론연구소(KSOI)가 동일한 시점에 동일한 설문지를 이용해 두 가지 다른 방식으로 조사했다. 아래 자료를 보고 물음에 답하시오.

	무선전화면접	무선ARS
표본크기	1005	1000
이재명 후보의 지지율	43.8%	43.2%
윤석열 후보의 지지율	36.1%	45.0%

[2-1] 무선전화면접 조사에서 이재명 후보와 윤석열 후보의 지지율 차이에 대한 95% 신뢰구간을 구하시오. 1005명의 표본을 단순확률추출 했다고 가정하시오.

[2-2] 두 조사방식에서 윤석열 후보의 지지율에 유의한 차이가 있다고 할 수 있는가? 이때 두 조사에서 각각 단순확률표본을 추출했다고 가정하시오. 유의수준은 5%이다.

[2-3] 조사방식에 따라 이재명 후보와 윤석열 후보의 지지율 차이가 달라진다고 할 만한 증거가 있는가? 즉, $0.077(=0.438-0.361)$과 $-0.018(=0.432-0.450)$이 다르다고 할 만한 충분한 증거가 있는가? 유의수준은 5%이다.

[3] (20점) "모집단 크기가 커지면 표본을 더 많이 추출해야 한다. 예를 들어 '기후위기'라는 주제에 대해 국민의식조사를 할 때, 인구 5천만인 우리나라에서 조사할 때보다 인구 3억인 미국에서 조사할 때 표본을 더 많이 추출해야 한다"라는 주장이 타당한가를 논하시오.

[4] (30점) 크기 N인 모집단을 L개의 층(strata)으로 나누어 층화추출(stratified random sampling)을 하려고 한다. N_i는 i번째 층의 크기를 나타내며 $N=N_1+N_2+\cdots+N_L$ 이다. 각 층에서 추출하는 표본의 크기 n_i를 층의 크기 N_i에 비례해서 할당(proportional allocation)했을 때,

[4-1] i번째 층에 있는 임의의 한 원소가 추출될 확률이 단순확률추출(simple random sampling) 했을 때의 확률과 같음을 증명하시오.

[4-2] 모평균에 대한 추정량의 형태가, 마치 단순확률추출을 한 것으로 간주하고 구한 추정량과 같아짐을 증명하시오. 즉, $\sum_{i=1}^{L} N_i\bar{y}_i/N=\sum_{i=1}^{L}\sum_{j=1}^{n_i} y_{ij}/n$ 임을 증명하시오. 단, y_{ij}는 i층에서 추출한 j번째 관측값이고, $\bar{y}_i=\sum_{j=1}^{n_i} y_{ij}/n_i$ 이다.

[4-3] 그렇다면 '비례배분에 의한 층화추출'과 '단순확률추출'은 아무런 차이가 없는 것인가?

표본론 2022년 기말시험

[1] (20점) 무명산에 살고 있는 노루의 총 수 N을 추정하기 위하여 한 생태학자는 다음과 같은 실험을 하였다. 먼저 그 지역의 노루 r마리를 잡아 표시를 한 다음 놓아주었다. 표시된 노루들이 이 지역에 고루 퍼질 수 있도록 시간을 둔 다음, 다시 n마리의 노루를 잡았다. s를 n마리 중에서 표시가 되어 있는 노루의 수라고 하자. 만약 무명산에 살고 있는 노루의 수가 이 실험 기간동안에 변하지 않았고 각 노루가 잡힐 가능성이 같다고 가정할 때

[1-1] (10점) s가 i일 확률, $P(s=i)$를 구하여라. 단 $n \le N-r$ 이다.

[1-2] (10점) $r=40, n=30, s=3$이 관측되었을 때, N을 추정하고 추정량의 오차의 한계를 구하시오.

[2] (20점) 산림관리자가 300 헥타르 면적의 산림에 식재된 전나무 중에서 죽은 전나무의 총수를 추정하고자 한다. 항공사진을 촬영한 다음 산림을 300개의 단위 헥타르 구역으로 나누었다. 항공사진의 정확성을 확인하기 위해 300개의 구역 중에서 6개 구역을 단순확률추출(simple random sampling)하여 실사를 하였다. 사진으로 확인한 죽은 전나무 수(x)와 실사 결과 정확히 조사한 죽은 전나무 수(y)에 대한 자료는 아래와 같다. 사진으로 확인된 죽은 전나무의 총 수(τ_x)가 6,200그루일 때, 실제 죽은 전나무의 총 수 τ_y에 대한 비추정량(ratio estimator)의 값을 구하고, 오차의 한계를 구하시오. 참고: $s_x^2 = 59.1$, $s_y^2 = 83.367$, $\hat{\rho} = \widehat{corr}(x,y) = 0.99014$.

조사구역	x_i	y_i	조사구역	x_i	y_i
1	15	18	4	30	36
2	28	35	5	10	14
3	22	26	6	18	20

[3] (30점) PPS (probabilities proportional to size) 추정량의 분산과 단순확률추출 추정량의 분산을 간단한 예제로 비교해보는 문제이다. 모집단의 원소를 {10, 2, 3}이라고 하자.

[3-1] (20점) 각 원소를 추출할 확률을 (0.6, 0.2, 0.2)로 정했다고 하자. 크기 2인 표본을 **복원추출** (sampling with replacement) 했을 때 얻어지는 총합추정량 $\hat{\tau}_1$의 모분산(population variance) $V(\hat{\tau}_1)$을 구하시오.

[3-2] (10점) 이번에는 단순확률표본(simple random sample)을 추출했다고 하자. 즉 크기 2인 표본을 **비복원추출**(sampling **without** replacement) 했을 때 얻어지는 총합추정량 $\hat{\tau}_2$의 모분산 $V(\hat{\tau}_2)$을 구하시오.

[4] (30점) 모집단을 층(strata)으로 나눈 다음, 각 층에서 군집(clusters)을 추출하고, 추출된 군집에서 원소를 추출한다고 하자. 즉, 모집단을 L개의 층으로 나누고, 층 i에서 m_i, $i=1,\dots,L$개의 군집을 단순확률추출한(simple random sampling) 다음, 군집 j에서 n_{ij}, $j=1,\dots,m_i$개의 원소(element)를 단순확률추출하기로 표집설계(sampling design)를 하였다. 각 층의 크기 N_i와 각 층의 군집의 수 M_i, 군집의 크기 N_{ij}, $i=1,\dots,L$, $j=1,\dots,M_i$를 모두 알고 있다고 가정하시오. 모집단의 크기를 N이라고 할 때

$$N = \sum_{i=1}^{L} N_i = \sum_{i=1}^{L} \sum_{j=1}^{M_i} N_{ij}$$

이다. (교재 9장에 있는 기호와 이 문제에서 주어진 기호의 정의가 다름에 주의하시오.) 이 표집설계에 따라 추출한 표본을 y_{ijk}, $i=1,\dots,L$; $j=1,\dots,m_i$; $k=1,\dots,n_{ij}$라고 표시할 때 모평균(population mean) μ를 어떻게 추정할 수 있는가? 모평균의 추정량과 추정량의 분산을 구하는 식을 각각 제시하고 이유를 설명하시오.

[1] 전화를 이용한 선거예측조사에서 무선전화면접 방식과 무선전화 자동응답시스템(ARS) 방식 중에서 어떤 방식으로 조사했느냐에 따라 조사결과가 달라진다는 주장이 있다. 이를 확인하기 위해 2022년 2월 교통방송(TBS)의 의뢰를 받아 한국사회여론연구소(KSOI)가 동일한 시점에 동일한 설문지를 이용해 두 가지 다른 방식으로 조사했다. 아래 자료를 보고 물음에 답하시오.

	무선전화면접	무선ARS
표본크기	1005	1000
이재명 후보의 지지율	43.8%	43.2%
윤석열 후보의 지지율	36.1%	45.0%

[1-1] 무선전화면접 조사에서 얻어진 이재명 후보와 윤석열 후보의 지지율 차이에 대한 95% 신뢰구간을 구하시오. 1005명의 표본을 단순확률추출 했다고 가정하시오.

[1-2] 두 조사방식에서 얻어진 윤석열 후보의 지지율에 유의한 차이가 있다고 할 수 있는가? 이때 두 조사에서 각각 단순확률표본을 추출했다고 가정하시오. 유의수준은 5%이다.

[2] 선거예측조사는 전화조사로 실시하는데 자동응답시스템(ARS) 방식인 경우 응답률이 5% 내외이고, 면접방식인 경우 응답률이 15% 내외이다. (이 때 응답률은 결번이거나 전화를 받지 않은 경우를 제외하고 전화연결이 된 통화 중에서 응답을 얻은 비율을 의미한다.) 선거예측조사에서 "ARS 방식과 같이 응답률이 5% 정도로 낮은 조사는 믿을 수가 없다"는 주장에 대해 논하시오.

[3] 변수 y에 대한 모집단 총합을 추정하고자 할 때, y의 관측값만 이용하여 추정하지 않고 또 다른 변수 x를 이용하는 방법을 쓸 수 있다. 비추정(ratio estimation)이나 회귀추정(regression estimation)을 이용하는 방법이 y의 관측값만 이용하는 방법보다 더 나은 방법이 되기 위한 조건(들)은 무엇인가? 그리고 이러한 조건(들)이 만족될 것 같은 구체적 예를 들어 보시오.

[4] 층화표집(stratified random sampling)과 군집표집(cluster sampling)을 비교하여 설명하고, 이 두 방법이 혼합된 표집설계(sampling design)의 예를 들어 보시오.

[5] 이번 학기에 재학 중인 숭실대학교 학생 전체를 대상으로 '비대면방식 수업에 대한 만족도'를 조사하고자 한다. 모집단을 대표하는 표본 500명을 추출하는 방법을 제시하시오. 표집틀(sampling frame), 표집단위(sampling unit), 층(strata), 군집(cluster) 등의 용어를 설명에 반드시 포함시키시오.