결과자료 통계분석과 활용 방법

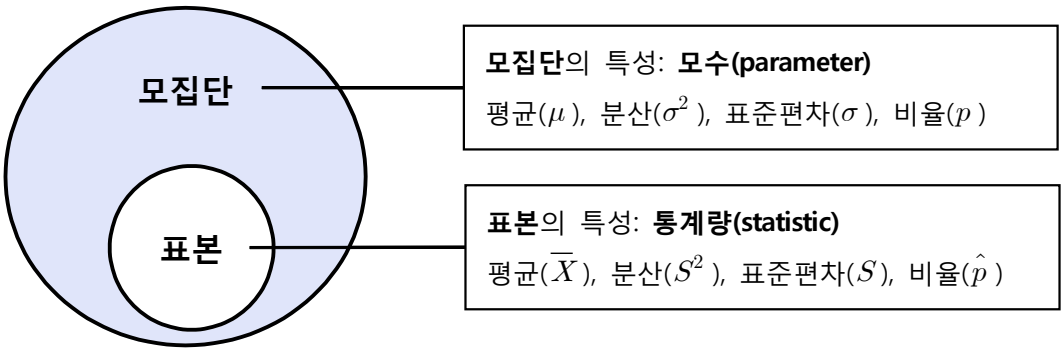
조 미 정
김영일교육컨설팅 연구위원

- 목 차 -

1. 통계학 개요

가. 통계학의 정의

관심의 대상 전체(모집단, population)에서 그 일부분인 표본자료(sample data)를 **수집**하여, 이 자료를 **정리 요약**하여 자료의 특성을 쉽게 파악할 수 있도록 묘사하고, 표면적인 자료의 특성 외에 내재되어 있는 자료의 성질과 규칙성과 같은 정보를 **분석**하여 표본자료의 모체인 관심의 대상에 대하여 특성을 **추론**하는 학문이다.



[그림 1-1] 모수와 통계량

나. 통계학 분류

1) 기술통계학(descriptive Statistics): 통계학의 기초적인 분야로 수집된 표본자료를 정리 요약하고 자료의 특성을 파악하기 쉽게 할 수 있도록 묘사하는 분야
   - 표(tables), 그래프(graphs) 그리고 요약값으로서 자료를 정리·도표화 그리고 기술(describing)하는 방법에 관한 통계학
2) 추측통계학(inferential Statistics): 표본자료로부터 분석된 성질과 규칙성과 같은 정보를 이용하여 관심의 대상전체인 모집단의 특성을 추론하는 분야
   - 추정(estimation)과 가설 검증(hypotheses testing) 그리고 미래에 대한 예측(forecasting) 수행 가능
   - 자료의 구조와 규칙에 관련된 모형을 설정(modeling)하므로 합리적인 의사 결정(decision making) 가능

다. 기본적인 통계 용어

다음의 [표 1-1]은 대학입시자료 예제로서 기본적인 통계용어를 설명해 보자.

[표 1-1] 대학입시자료

지원자	모집단위	전형유형	학생부(교과)		수능	
			표준점수	석차등급	언어	수리
김 **	인문과학	학교생활우수자	8.73	2.19	125	137
이 **	사회과학	글로벌리더1	10.88	4.82	127	134
박 **	공학	리더십	7.92	2.07	127	140
조 **	자연과학	논술우수자	13.78	2.39	123	137
고 **	의예과	일반전형	3.31	1.19	122	140

1) 구성요소(element/member): 지원자
2) 변수(variable): 모집단위, 전형유형, 학생부(표준점수, 석차등급), 수능(언어, 수리)
3) 관측값(observation) 또는 측정값(measurement): 각 변수에 해당되는 값(value)

(참고) **변수의 종류**

1. **질적변수(qualitative variable)**: 범주변수(categorical variable)라고도 하며 계산가능한 숫자적(numeric) 의미보다는 명목적(nominal) 또는 서열적(ordinal) 의미를 갖는 변수
2. **양적변수(quantitative variable)**: 숫자적 계산과 측정이 가능한 변수로써 우리가 가장 많이 접하는 변수

(참고) **계수가능성(countability)에 의한 분류**

1. **이산변수(discrete variable)**: 한개, 두개 등 개수로서 정확히 셀 수 있는 변수로 측정단위 이하로는 셀 수 없는 변수
2. **연속변수(continuous variable)**: 특정한 값뿐만 아니라 그 구간까지도 측정할 수 있는 변수

라. 척도의 유형

척도(scale)는 변수를 측정하는 도구를 말한다. 모든 척도는 그 척도가 담고 있는 정보의 양에 따라 명목척도, 서열척도, 간격척도, 비율척도 등 네 가지로 분류될 수 있다. 이 중 명목척도에 의해 측정된 자료가 가장 적은 정보를 갖는다. [표 1-2]는 이를 요약하여 보여주면, 구체적인 내용은 뒤에서 서술된다.

[표 1-2] 네 가지 척도의 특성

척도	정보				통계기법
	범주	서열	거리	원점/비율	
명목척도	○	×	×	×	비모수/모수
서열척도	○	○	×	×	비모수
간격척도	○	○	○	×	모수
비율척도	○	○	○	○	모수

(1) 명목척도(nominal scale): 측정대상에 그들이 속한 범주나 종류에 따라 수나 번호를 부여하여 생긴 척도

[예-1] 성별: 남성 - 0, 여성 - 1

[예-2] 계열: 인문계 - 1, 자연계 - 2, 예체능계 - 3

(2) 서열척도(ordinal scale): 조사 대상들의 특성을 서열로 나타내어 주는 척도

[예-1] 학년: 1학년 - 1, 2학년 - 2, 3학년 - 3

[예-2] 선호도: 아주 좋다 - 1, 좋다 - 2, 보통이다 - 3, 나쁘다 - 4, 아주 나쁘다 - 5

(3) 구간척도(interval scale): 숫자간의 구간이 산술적 의미를 갖는 척도

[예] 등급구간

등급	1	2	3	4	5	6	7	8	9
누적	4	11	23	40	60	77	89	96	100

(4) 비율척도(ratio scale): 수치적 척도로 범주, 서열, 거리의 정보에 추가적으로 비율의 정보를 갖는 척도로 가장 상위의 척도

[예] 학생부 과목별 원점수, 수능 성적, 표준점수, Z값 등

2. 표본통계량

본장에서는 표본통계량을 계산하고 그 값을 이용하여 표본자료의 특징을 산술적으로 표현하는 방법을 서술한다. 표본을 산술적으로 요약할 때 우리는 표본의 중심(中心)과 중심으로부터 흩어져 있는 정도(퍼져있는 정도 또는 모여있는 정도)에 대하여 주된 관심을 갖는다.

표본자료의 중심을 대표(代表)값이라 하고 중심으로 산재되어 있는 정도를 산포도(散布度)라고 한다. 대표값과 산포도는 통계학에서 가장 중요한 개념으로 이 두 가지 표본통계량이 없으면 통계학이 존재하지 않을 정도로 매우 중요한 통계학의 핵심이다.

가. 대표값

표본자료의 중심을 나타내는 대표값에는 평균(mean 또는 average), 중앙값(median), 최빈값(mode), 그리고 절사평균(trimmed mean) 등의 종류가 있다.

평균

수집된 n 개의 표본자료의 관측값을 x_1, x_2, ..., x_n 으로 표시한다. 표본평균은 표본 관측값의 모든 합을 관측값의 개수로 나눈 것으로 (1-1)과 같이 정의되며 $\overline{X}$ 로 표시한다.

$$\overline{X} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \qquad (1\text{-}1)$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

[표 2-1]의 자료는 어떤 고등학교에서 남학생 8명과 여학생 7명을 무작위로 선정하여 100점을 만점으로 '체험활동 만족도'에 대하여 조사한 자료이다.

[표 2-1] 체험활동 만족도

남	52	99	43	82	92	60	39	78
여	92	62	95	74	54	75	85	

[표 2-1]의 자료에서 남학생들에 대한 체험활동 만족도의 평균값을 구하여 보자. 첫 번째 관측값 x_1=52 이고 두 번째 관측값 x_2=99 이며,..., 8번째 관측값 x_8=78 이므로 식 (1-1)에 의하여 표본평균은

$$\overline{X_{남}} = \frac{1}{8}(52 + 99 + \cdots + 78) = 68.1$$

이다. 그리고 여학생들에 대한 체험활동 만족도의 평균값은 다음과 같이 구한다.

$$\overline{X_{여}} = \frac{1}{7}(92 + 62 + \cdots + 85) = 76.7.$$

체험활동 만족도의 평균값만으로는 여학생들이 남학생들보다 체험활동에 대하여 더욱 만족한다고 볼 수 있다.

표본자료의 관측값을 수직선 위에 점으로 표시한 후 수직선을 무게가 없는 막대기로 가정하고 각 관측값 위에 동일한 무게의 추를 얹어 놓으면, 이 막대기는 표본 평균값에서 평형을 이루며 좌우로 쳐지지 않는다. 즉, 이러한 의미에서 평균은 무게중심 (center of gravity)이라고도 한다. 중심을 나타내는 평균은 각각의 관찰값의 크기에 민감하게 반응하며 특히 매우 큰 값 (또는 작은 값)과 같은 값들에는 매우 민감하게 반응하는 특징이 있다.

중앙값

[표 2-1]에서 나열된 남학생과 여학생의 만족도에 관한 관측값 8개, 7개를 크기 순서대로 나열하면 다음과 같다.

[표 2-2] 만족도 자료의 순서통계량 값

남	39	43	52	60	78	82	92	99
여	54	62	74	75	85	92	95	

위와 같이 크기 순서대로 나열된 관측값을 순서통계량(order statistic)이라고 하며, n 개의 관측값인 경우에 가장 작은 값을 가진 순서통계량을 최소 순서통계량 $x_{(1)}$으로 표시하고 가장 큰 값을 가진 최대 순서통계량을 $x_{(n)}$으로 표시한다. 따라서 여학생에 대하여 $x_{(1)} = 54$, $x_{(2)} = 58$, ... , $x_{(8)} = 95$ 가 된다.

표본중앙값은 순서통계량 중에서 가장 가운데 위치한 관측값으로 정의되며 $\tilde{X}$로 표시한다. 이때 관측값의 개수가 홀수$(n = 2k-1\ ;\ k = 1,\ 2,...)$인 경우에 가장 가운데 위치한 관측값은 k번째 순서통계량 $x_{(k)}$이지만, 관측값의 개수가 짝수$(n = 2k\ ;\ k = 1,\ 2,...)$인 경우에는 가운데 위치한 관측값이 $x_{(k)}$와 $x_{(k+1)}$이므로 표본중앙값은 이 두 값의 평균을 택한다. 즉 n 개의 순서통계량

$\{x_{(1)},\ x_{(2)},\ \cdots,\ x_{(n)}\}$에 대하여 표본중앙값 $\tilde{X}$는 다음과 같이 정의한다.

$$\tilde{X}=\begin{cases} x_{(k)} & , \quad \text{만약} \quad n=2k-1, \\ \dfrac{x_{(k)}+x_{(k+1)}}{2} & , \quad \text{만약} \quad n=2k\ , \end{cases} \qquad \text{여기서} \quad k=1,\ 2,\ \cdots. \qquad (1\text{-}2)$$

크기 순서대로 배열한 [표 2-2] 자료에서 남학생들의 표본중앙값은 $(x_{(4)}+x_{(5)})/2=(60+78)/2=69$ 이며 여학생들의 표본중앙값은 $x_{(4)}=75$ 이다. 표본중앙값에서도 남학생들보다 여학생들의 만족도가 높으며 표본평균보다 남학생들의 중앙값은 더욱 높고 $(\bar{x}_{(남)}=68.1\ <\ \tilde{x}_{(남)}=69)$ 여학생인 경우에는 표본평균이 높은 값을 가지고 있다 $(\bar{x}_{(여)}=76.7\ >\ \tilde{x}_{(여)}=75)$.

무게중심을 나타내는 표본평균은 관측값에 민감하게 반응하지만 표본중앙값은 관측값의 크기 순서에 의존하기 때문에 평균에 비하여 민감하지 않다. 특히 표본자료 중 다수의 관측값에 비하여 멀리 떨어져 있는 관측값(매우 큰 값 또는 매우 작은 값)을 통계학에서는 특이값(outlier)이라고 하는데 이러한 특이값에 표본평균의 값은 매우 민감하게 반응하는 한편 표본중앙값은 많은 영향을 받지 않는다는 장점이 있다. 특이값에 따라 많이 변하지 않는 통계량을 로버스트(robust)하다고 하는데, 이러한 관점에서 중앙값은 로버스트하다고 한다.

절사평균

표본평균과 표본중앙값의 장점을 이용하고 단점을 최소화하기 위한 절충안으로 절사평균을 고려해 보자. 절사(trimmed)라는 말의 의미대로 전체표본자료 중에 양쪽 끝부분의 일정한 자료를 제외하고 나머지 관측값들을 사용하여 평균을 구하는 방법으로, 0과 0.5 사이의 값을 갖는 임의의 α에 대하여 $100\times\alpha$ % 절사평균은 순서통계량 중에서 전체의 $100\times\alpha/2$ %에 해당하는 작은 관측값들과 $100\times\alpha/2$ %에 해당하는 큰 관측값들을 제외한 나머지 순서통계량의 평균으로 정의하며 이를 $\bar{X}_{\alpha}$로 표시한다.

예를 들어 [표 2-2] 남학생의 자료에서 25% 절사평균 $\bar{X}_{0.25}$는 $8\times(0.25/2)=1$개의 가장 큰 값과 가장 작은 값을 절사한 나머지 6개의 관측값에 대한 평균이다. 즉

$$\bar{x}_{0.25(남)}=\frac{1}{6}(43+52+60+78+82+92)=67.8$$

이다.

체조 경기에서는 오래 전부터 절사평균을 사용해 왔다. 체조 경기의 채점은 5명의 심판이 하는데 우수한 선수라 하더라도 한 명의 심판이 심한 주관적 편견에 의하여 0점을 준다면 그 선수는 입상권에서 제외될 수밖에 없다. 이와 같이 한 두 명의 심판 편견을 배제하기 위하여 가장 높은 성적과 가장 낮은 성적을 제외한 나머지 3명의 심판 성적의 평균으로 순위를 결정하는 방법을 사용해 왔는데 이런 방법이 대표값 중에서 40% 절사평균을 이용한 대표적 예이다.

최빈값

최빈값은 표본자료 중에서 가장 높은 빈도수를 가진 값(또는 계급 또는 항목)으로 정의된다. 빈도수가 높다는 의미는 자주 발생한다는 뜻을 갖고 있기 때문에 대표값으로 사용되고 있다. 예를 들어 어떤 제조업체의 급여 자료에 대하여 살펴보면 임원은 소수이며 고액을 받고 단순 근로자는 많은 인원을 차지하면서 낮은 임금을 받는 업체가 많다. 대부분의 급여(임금) 자료와 같이 오른쪽으로 기

울어진 자료에서는 다음과 같은 관계를 갖고 있다.

최빈값 < 중앙값 < 평균

이런 경우에 노사협상에서의 사용자는 회사원의 급여에 대한 대표값으로 평균을 사용하고자 할 것이며, 반면에 노동자는 중앙값을 대표값으로 사용하는 것이 더욱 적절하겠다. 이렇게 급여 분포가 좌우 대칭형태가 아니며 한쪽으로 기울어진 형태의 경우에는 자료의 중심을 나타내는 평균은 대표값으로 적절한 척도가 되지 못한다. 이러한 경우에는 평균보다는 중앙값과 최빈값 등을 모두 계산하여 자료와 더불어 상세히 살펴보면서 어디에 집중되어 있고 모여 있는지를 분석한 다음 어떤 대표값이 자료를 잘 대표하는가를 결정하는 것이 바람직하다.

표본자료가 [표 2-1]과 같은 자료에서는 같은 값을 갖는 관측값이 발생하지 않으므로 최빈값의 의미가 없다. 이런 경우에는 최빈값을 구하기 위해서는 자료를 구간별로 나누어 계급별 빈도수를 구하여 빈도수가 가장 많은 계급값을 최빈값으로 정하는 것이 바람직하다.

나. 산포도

교과부는 국가수준 학업성취도(이하 일제고사)를 치루고 나면 각 지역별, 고교별, 학년별 평균, 기초학력 미달 비율과 보통학력 이상 비율 등 평가결과를 발표한다. 기초학력 미달 비율이 감소하거나 보통학력 이상 비율이 증가 했을 때 학력이 향상되었다고 보도한다. 이때 평균에 의해 전년대비 오르거나 내림만으로 학력 향상을 정확히 판단하기는 어렵다. 따라서 성적의 양극단인 기초학력 미달과 보통학력 이상 비율을 통해 보다 정확한 학력 상승에 대한 정보를 얻어낼 수 있다. 이러한 정보가 자료의 산포도를 의미하는 통계량이다. 이와 같이 자료의 중심을 의미하는 대표값도 중요하지만 그 다음으로 중요한 것으로는 관측값이 대표값 주위에 흩어져 있는(또는 모여 있는) 정도를 측정하는 산포도이다. 산포도를 측정하는 방법도 여러 가지가 있는데 본장에서는 표본분산(sample variance), 표본표준편차(sample standard deviation) 그리고 범위(range)와 사분위간 범위(inter-quartile range)에 대해 서술한다.

표본분산과 표준편차

표본분산은 식 (1-3)과 같이 정의되며, S^2이라고 표시한다. 이 식에서 가장 중요한 핵심은 각 관측값 x_i와 표본평균 $\overline{X}$와의 차이, 즉 $(x_i - \overline{X})$이다. 이 차이를 편차(deviation)라고 하는데 표본평균을 중심으로 얼마나 떨어져 있는지를 측정해 준다.

그러나 이 편차의 합은 항상 0이므로 $(\sum_{i=1}^{n}(x_i - \overline{X}) = 0)$산포도의 척도로는 편차의 제곱합을 고려한다. 표본분산은 편차의 제곱합을 $n-1$로 나누어 준 것이다. 여기서 표본분산 식의 분모가 n이 아니라 $n-1$인 것은 편차의 합이 0이기 때문이다. 다시 말하여 n개의 관측값이 있는 경우에 $n-1$개의 편차만 주어지면 나머지 한 개의 편차는 저절로 정해지기 때문이다. 이를 통계학에서는 자유도(degree of freedom)가 $n-1$이라고 한다.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \qquad (1\text{-}3)$$

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} \qquad (1\text{-}4)$$

식 (1-4)와 같이 표본분산의 제곱근을 표본표준편차 (sample standard deviation)라고 정의하며 S로 표시한다.

예를 들어 [표 2-1]의 자료에서 남학생들의 체험활동 만족도의 표본분산과 여학생들의 표본분산은 다음과 같이 구한다.

$$S_{남}^2 = \frac{1}{7}[(52-68.1)^2+(99-68.1)^2+\cdots+(78.1-68.1)^2] = \frac{1}{7}(3618.87) = 516.98$$

$$S_{여}^2 = \frac{1}{6}[(192-76.7)^2+(62-76.7)^2+\cdots+(85-76.7)^2] = \frac{1}{6}(1379.43) = 229.90$$

또한 각각의 표본표준편차는 다음과 같다.

$$S_{남} = \sqrt{516.98} = 22.74$$

$$S_{여} = \sqrt{229.90} = 15.16$$

체험활동 만족도에 대한 남학생의 평균은 여학생의 평균보다 낮았지만 남학생의 분산과 표준편차는 여학생보다 큰 값을 갖고 있음을 알 수 있다. 이는 여학생들은 남학생보다 체험활동에 대한 만족감은 더욱 느끼고 있으며 그 만족감의 폭도 남학생보다 좁다는 것으로 판단된다. 남학생들은 체험활동에 대하여 매우 만족하는 사람도 있는 반면에 매우 만족하지 못하는 사람들도 있고 전반적으로 여자들보다는 만족감이 낮음을 알 수 있다.

일반적으로 산포도의 척도로서 분산보다는 표준편차를 많이 사용한다. 왜냐하면 분산은 편차의 제곱합이므로 분산의 단위는 사용되는 측정 단위의 제곱형태로 나타나는 반면에, 표준편차는 편차의 제곱합의 제곱근으로 표준편차의 단위는 사용되는 측정 단위와 동일하기 때문이다.

범위와 사분위간 범위

범위는 관측자료 중에서 최대값과 최소값의 차이로 정의된다. n개의 순서통계량 $\{x_{(1)},\ x_{(2)},\ \ldots,\ x_{(n)}\}$에 대하여 범위는

$$R = x_{(n)} - x_{(1)} \qquad (1\text{-}5)$$

으로 정의되며 가장 간단한 산포도 척도이다. [표 2-1] 자료를 크기 순서대로 정렬한 [표 2-2]의 자료에서 체험활동 만족도에 대한 남학생과 여학생들의 범위는 다음과 같다.

$$R_{남} = 99-39 = 60$$

$$R_{여} = 95-54 = 41$$

표본자료 전체를 크기 순서대로 나열한 다음 크게 4등분하여 보자. 크게 2등분했을 때 가운데 값이 중앙값이므로 2등분된 각각을 다시 2등분하면, 표본전체를 4등분할 수 있다. 이렇게 4등분된 자료에서 최소값부터 시작하여 처음 등분된(25%) 위치에 있는 관측값을 하사분위수(lower quartile)라고 부르며, 두번째 등분된(50%) 위치에 있는 관측값은 중앙값이며, 세번째 등분된(75%) 위치에 있는 관측값을 상사분위수(upper quartile)라고 한다. 여기서 하사분위수, 중앙값, 상사분위수를 사분위수라고 한다.

범위 이외의 산포도를 측정하는 척도로 사분위간 범위(inter-quartile range ; IQR)는

IQR = (상사분위수) - (하사분위수) (1-6)

로 정의되며, 중앙값을 기준으로 양쪽에 각각 25% 씩 해당되는 관측값을 포함하는 구간의 길이로서 정의된다. 다시 말하면 중앙값을 기준으로 전체표본자료의 가운데 50%에 해당하는 자료에 대한 범위를 사분위간 범위라고 한다.

3. 정규분포

관측된 표본의 자료를 이용하여 모집단에 대한 추론을 하기 위해 확률이 사용된다. 일반적으로 실험의 결과를 나타내는 사상은 말이나 문자로 나타내지만 실제로 많은 실험의 결과의 형태는 숫자이다. 일정한 확률을 가지고 발생하는 사건에 수를 부여하는 것을 확률변수라고 한다. 확률변수에는 이산형 확률변수와 연속형 확률변수가 있고 각각의 정의는 다음과 같다.

이산형 확률변수

확률변수의 가능한 값들이 셀 수 있는(countable) 경우 **이산형 확률변수**(discrete random variable)이라 한다.

연속형 확률변수

확률변수의 값이 하나 혹은 그 이상의 구간(interval) 내의 임의의 값을 가질때 **연속형 확률변수** (continuous random variable)라고 한다.

또한 확률분포란 확률변수가 가질 수 있는 값과 그와 관련된 확률을 나타내는 것이다. 확률변수의 확률분포는 표, 그래프 또는 공식으로 타나낼 수 있다. 이산형 확률변수의 확률분포 중에서 가장 널리 사용되는 확률분포에는 **이항분포**가 있고, 연속형 확률변수의 확률분포의 대표적인 것으로 정규분포가 있다. 본장에서는 연속형 확률변수의 분포인 정규분포에 대하여 그 특성을 살펴보고자 한다.

가. 연속형 확률변수

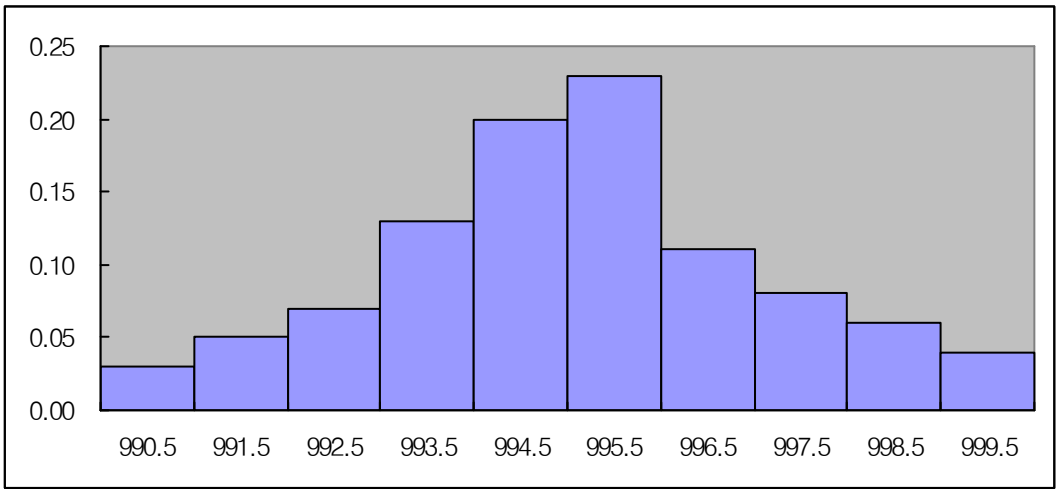
연속형 확률변수는 하나 또는 하나 이상의 구간 내에 있는 임의의 값을 가질 수 있으므로 연속형 확률변수가 가질 수 있는 가능한 값은 무한개(infinite)이거나 셀 수 없다(uncountable). 연속형 확률변수의 확률분포를 설명하기 위해 다음의 상대도수 분포의 예를 들어보자.

어느 한 대학의 체육학과는 대학입학전형요소 중 실기시험에서 만점을 1000점으로 하고 기본점수를 900점으로 부여한다고 하자. 측정된 실기시험성적 X는 1000에 가까운 값을 갖는 임의의 실수가 되며 가능한 값은 무한개이므로 X는 연속형 확률변수이다. [표 3-1]은 100명의 체육학과 지원자에 대한 실기시험성적의 상대도수 분포표이며, [그림 3-1]은 [표 3-1]의 상대도수 분포에 대한 히스토그램이다. 또한 [그림 3-2]는 상대도수의 분포를 곡선으로 연결시켜 그린 것이다.

[표 3-1] 체육학과 지원자 실기시험성적의 도수와 상대도수

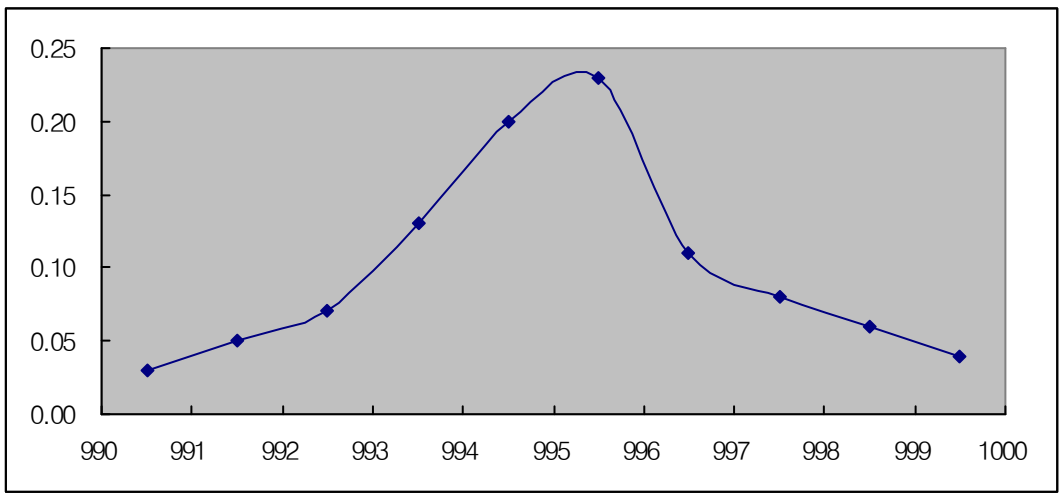
X	f	상 대 도 수
990≤X<991	3	0.03
991≤X<992	5	0.05
992≤X<993	7	0.07
993≤X<994	13	0.13
994≤X<995	20	0.20
995≤X<996	23	0.23
996≤X<997	11	0.11
997≤X<998	8	0.08
998≤X<999	6	0.06
999≤X<1000	4	0.04

EXCEL의 [차트마법사]-[세로막대형]을 선택하고, 옵션을 수정하면 다음과 같은 히스토그램이 완성된다.



[그림 3-1] 실기시험성적에 대한 히스토그램

또한, [차트마법사]-[분산형]을 선택하면 아래와 같이 곡선으로 연결된 그래프를 얻게 된다.

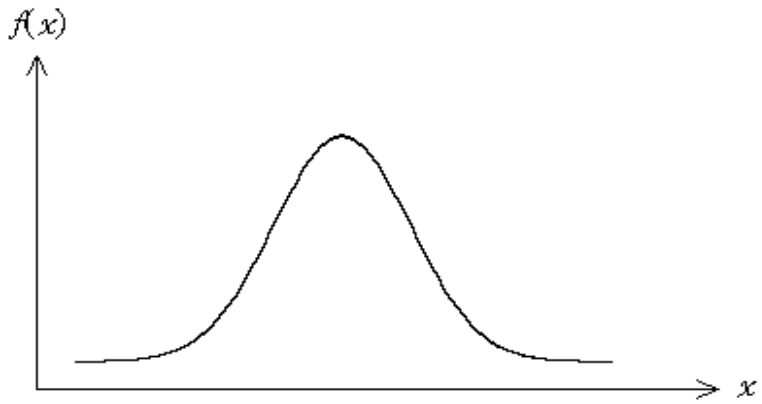


[그림 3-2] 실기시험성적에 대한 확률분포 곡선

[그림 3-2]는 연속형 확률변수 X의 확률분포를 근사 시킨 **분포곡선**이다. 히스토그램의 X의 간격이 1이므로 히스토그램의 각 막대 면적의 합은 1이다. 만약 상대도수 분포를 그릴 때 X의 간격이 1이 아니면 상대도수를 간격너비(width)로 나눈 **상대도수 밀도**(relative frequency

density)를 각 구간의 높이로 표시한 것이 **히스토그램**이다. 따라서 상대도수 밀도를 그린 경우 모든 막대 면적의 합은 1 이 된다.

이제 실험을 무한히 많이 하고 구간의 너비를 가능한 작게 하여, 연속형 확률변수 X의 확률분포함수를 그리면 [그림 3-3]처럼 된다. 연속형 확률변수의 확률분포곡선을 **확률밀도함수** (probability density function)라 하며 $f(x)$로 나타낸다. 연속형 확률변수의 확률밀도함수는 히스토그램의 극한곡선으로 설명할 수 있다.

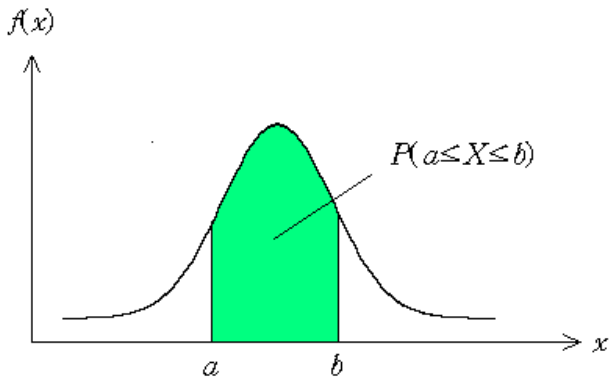


[그림 3-3] 연속형 확률변수 X의 확률 밀도 함수

연속형 확률변수 X의 확률밀도함수는 다음의 특성을 가진다.

1. 확률밀도 함수하의 임의의 구간의 면적은 그 구간의 확률이 된다. 예를 들어 [그림 3-4]에서처럼 X가 a 로부터 b 까지의 면적은 그 구간의 확률 $P(a \leqq X \leqq b)$ 이 되며 이 확률은 0 과 1사이의 값을 갖는다.

$P(a \leqq X \leqq b)$ ⇒ a 로부터 b 까지의 면적



[그림 3-4] 확률밀도함수 $f(x)$ 의 면적

2. 확률밀도 함수하의 구간의 면적은 확률을 나타내므로 확률밀도 함수하의 전체 면적의 합은 전체 확률의 합인 1이 된다.

$$P(-\infty \leqq X \leqq \infty) = 1$$

X의 한 점에서는 면적이 0이므로 임의의 X의 한 점에서의 확률은 0이다. 즉, $P(X = a) = 0$

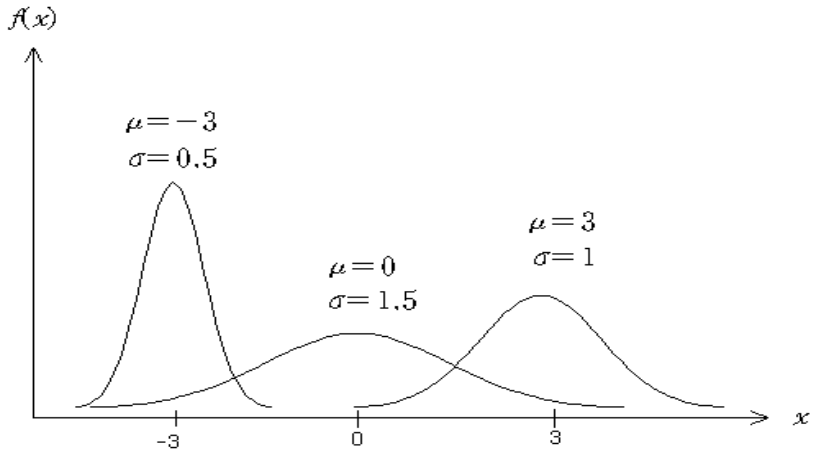
이다. 따라서 연속확률변수 X에 대해서 구간의 양 끝점 a 와 b 가 구간에 포함되던 포함되지 않던 X가 그 구간에 속할 확률은 같다.

$$P(a \leqq \mathrm{X} \leqq b) = P(a \leqq \mathrm{X} < b) = P(a < \mathrm{X} \leqq b) = P(a < \mathrm{X} < b)$$

나. 정규분포

연속형 확률분포 중에서 가장 많이 사용되는 것 중의 하나가 정규분포(normal distribution)이다. 실제로 많은 현상들은 정규분포를 따르거나 정규분포에 의해 근사 될 수 있다. 예를 들어 사람의 몸무게, 키 또는 전국 수능 응시자의 성적 등은 정규분포에 대해 잘 근사된다. 또한 기상실험, 강우량 조사 또는 부품의 측정 등과 같은 물리적 실험은 정규분포에 적합하다고 알려져 있다.

정규분포의 그래프는 [그림 3-5]에서처럼 종 모양의 곡선으로 평균 μ에 대해 좌우대칭이며 퍼져있는 정도는 표준편차 σ에 의해 결정된다.



[그림 3-5] μ와 σ의 변화에 따른 정규분포

정규분포

평균 μ와 표준편차 σ를 가지는 **정규확률변수** X의 확률분포는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

여기서 $\pi = 3.14\ldots$, $e = 2.71..$이다.

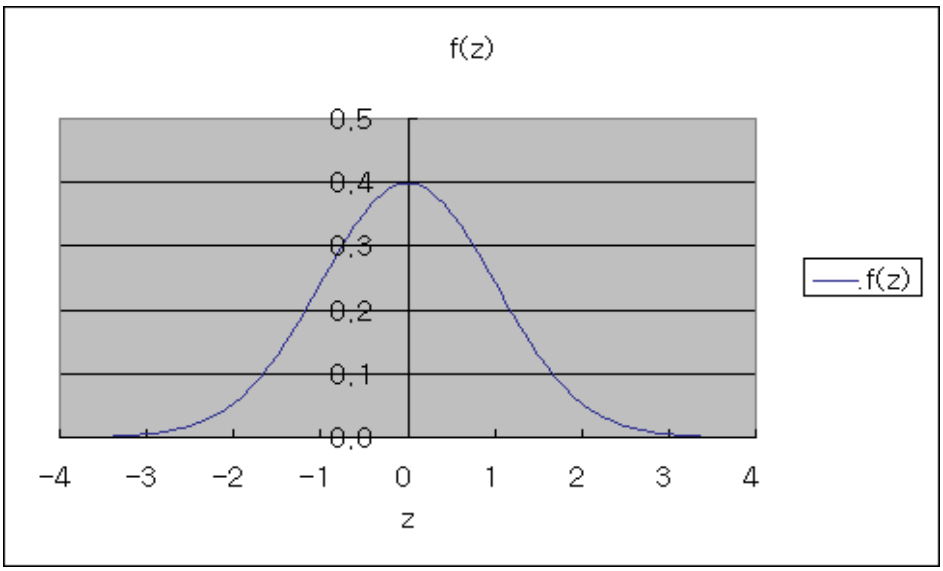
정규분포는 μ와 σ에 따라 무수히 많은 정규곡선을 갖는다. 또한 확률을 계산하기 위해 확률밀도 함수를 이용해야 하나 무수히 많은 μ와 σ에 대해 확률밀도 함수를 이용하여 확률을 구하는 것은 매우 복잡하며 쉽지 않다. 따라서 확률계산이 용이하게 평균을 0 그리고 표준편차를 1로 변환시킨다.

표준정규분포

평균 μ=0, σ=1인 정규확률변수의 분포를 **표준정규분포**라 한다. 표준정규 확률변수 Z 의 확률분포는 다음과 같다.

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}, \quad -\infty<z<\infty$$

평균 $\mu=0$, 표준편차 $\sigma=1$인 정규분포를 **표준정규분포**(standard normal distribution)라 하며 $N(0,1)$으로 나타낸다. 표준정규분포를 갖는 확률변수를 **표준정규 확률변수**라 하며 Z 로 나타낸다. 표준정규 확률변수 Z 가 $N(0,1)$을 따른다는 것을 $Z\sim N(0,1)$이라 표시하며 일반적으로 확률변수 X가 평균 μ, 표준편차 σ 인 정규분포를 따를 때 우리는 $X\sim N(\mu,\sigma^2)$ 으로 표시한다. [그림 3-6]은 표준정규분포의 그래프를 나타낸다.



[그림 3-6] 표준정규분포

위의 수평축에서 z값은 평균과 그 지점까지의 거리를 표준편차로 나타낸다. 예를 들어 $z=1.5$인 점은 평균의 오른쪽으로 1.5배 표준편차만큼 떨어진 점이다. 또한 정규곡선하의 총면적은 1이므로 정규곡선하의 오른쪽 반과 왼쪽 반의 면적은 $\frac{1}{2}$이다.

[그림 3-7]은 표준정규분포에서 가운데 구간의 면적이 0.9, 0.95, 0.99인 Z 값을 나타낸다. 이 값들은 정규확률 계산시 자주 사용되는 Z 값이다.

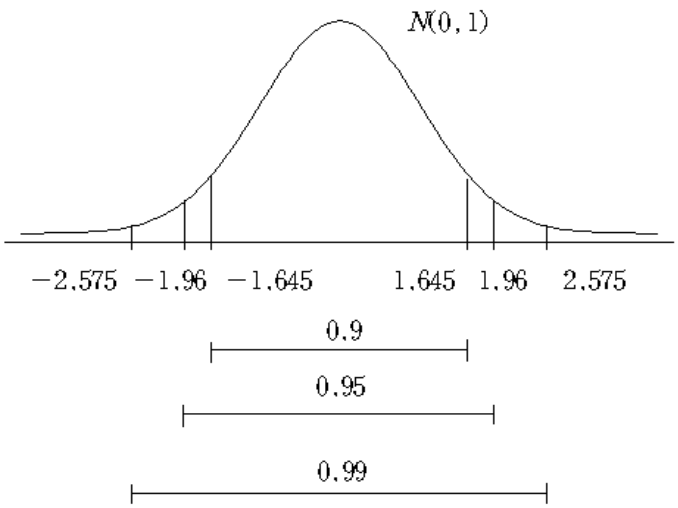
$$P(-1.645<Z<1.645)=0.90$$

$$P(-1.96<Z<1.96)=0.95$$

$$P(-2.575<Z<2.575)=0.99$$

$P(|Z|>1.96)=0.05$이다. Z = 1.96의 의미는 X의 값이 평균보다 표준편차의 1.96배만큼 떨

어져 있음을 의미한다. 따라서 확률변수가 평균으로부터 표준편차의 1.96배 이상 떨어져 있을 확률은 0.05이다.



[그림 3-7] 표준정규분포에서 가운데 구간의 면적이 0.9, 0.95, 0.99 인 Z의 값

실제 자료에서는 평균과 표준편차가 각각 0과 1이 아닌 임의의 값들을 갖는 정규분포를 따르므로 확률계산 시 표준정규분포표는 이용할 수 없다. 따라서 정규분포를 표준정규분포로 변환시키는데 이를 정규분포의 **표준화**(standardization)라고 한다.

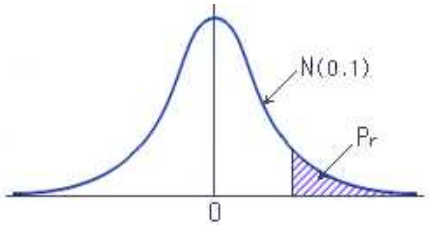
정규분포의 표준화

X가 평균이 μ, 표준편차 σ인 정규분포를 따를 때, 변환된

$$Z=\frac{X-\mu}{\sigma}$$

는 평균이 0, 표준편차가 1인 **표준정규분포**를 따른다.

※ 표준정규분포표



z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1778	0.1762	0.1736	0.1711	0.1658	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

4. 표본 추출

* 통계학에서 자료를 획득하는 것은 분석을 위하여 중요한 부분이다.
* 모집단에 대한 어떤 정보나 특성을 알고자 할 때 모집단 전체를 조사하지 않고 모집단에서 추출된 표본을 조사하여 정보를 얻는 경우가 많다.
  ⇒ 이 때 추출된 표본은 모집단의 일부이므로 반드시 오차가 발생한다.
  ⇒ 이러한 오차를 줄이기 위해서는 모집단을 잘 대표할 수 있는 표본을 추출

가. 무작위 표본

1) 모집단(population)

* 모집단 : 연구 대상이 되는 물체나 사람들의 총체 ⇒ 여기에서 표본 추출
  ⇒ 유한집단 / 무한집단
  ⇒ 모집단은 일정한 확률분포를 가지고 있고, 모집단 분포의 수량적인 특성을 모수(patameter)라 함
  ⇒ 표본에 의해 얻은 정보를 이용하여 모집단의 특성을 일반화한다는 것은 모집단 분포의 모든 특성에 관한 것이 아니라 모집단의 평균이나 분산 등과 같은 모수에 대해 추정하는 것을 말함.
* 무작위표본(확률표본, random sample)에 속한 각 관측값의 확률분포 $p(x)$ 는 모집단의 확률분포와 동일하다.

2) 표본추출방식

* 전수조사(complete survey) : 모집단에 관한 어떤 지식이나 특성을 얻기 위하여 모집단 전체를 조사하는 통계조사
* 표본조사(sample survey) : 모집단으로부터 추출한 표본을 조사하여 그 결과를 바탕으로 모집단에 대한 특징을 추정하는 통계조사
* 표본추출방법 : 유의추출법(purporsive selection)과 임의추출법(random sampling)
  ① 유의추출법 : 표본을 추출할 때 주관적인 판단, 즉 조사자의 지식 또는 경험을 토대로 하여 적당하다고 생각되는 표본을 추출하는 방법
② 임의추출법 : 원소가 표본으로 뽑힐 확률을 같도록 하여 객관적으로 추출하는 방법
* 임의추출법의 종류 : 단순임의추출법, 층화임의추출법, 집락추출법
  ① 단순임의추출법(simple random sampling, SRS)
    ⇒ 모집단에서 관측값을 택할 때마다 모집단에 속해있는 각 관측값들이 채택될 확률이 동일한 표본
    ⇒ 카드뽑기, 난수 주사위, 난수표 이용
    ⇒ 표본평균 $\overline{X}$ 는 개개의 관측값들에 비해 모집단의 중심에 더욱 가까이 위치함
  ② 층화임의추출법(stratified random sampling)
    ⇒ 모집단을 동질적인 성질을 가진 것끼리 몇 개의 집단 또는 층으로 분류한 다음 각 층에서 어떤 규칙에 따라 표본을 단순임의추출하는 방법
    ⇒ 동질적인 층으로 나눌 때, 추정량의 분산을 줄일 수 있다.
    ⇒ 이럴 때 사용 : 부분집단별로 정확도가 요구 될 때
    : 간편한 조사가 요구될 때
    : 모집단의 분포가 한쪽으로 치우쳐 있다고 생각될 때

⇒ 층간의 이질성은 크게, 층내의 이질성은 작게

③ 집락추출법(cluster sampling)

⇒ 모집단을 집락(cluster)이라 부르는 몇 개의 부분으로 분류하고 그 집락을 임의추출하여 추출된 집락을 전수조사하거나 그 집락의 표본을 추출하는 방법

⇒ 집락 사이의 이질성이 작아야 한다.

3) 복원추출과 비복원추출

* 복원추출(sampling with replacement) : n 개의 관측값이 서로 독립

⇒ 단순임의추출시 다음 성질 만족

> SRS에 속한 n 개의 관측값 $X_1, \cdots, X_n$ 은 서로 독립적이며 각 관측값의 분포는 모집단의 분포 $p(x)$ 와 동일하다. 즉, 다음이 성립한다.
>
> $$p(x_1) = \cdots = p(x_n) = p(x)$$
>
> 따라서 각 관측값의 평균과 표준편차는 모집단의 평균 μ 와 표준편차 σ 와 같다.

* 비복원추출(sampling without replacement) : 각 관측값은 독립이 아니다. 그러나 모집단이 큰 경우는 독립이라고 가정해도 좋다.

나. 표본의 성격

* 표본평균

$$\overline{X} = \frac{1}{n}[X_1 + \cdots + X_n]$$

⇒ X_i는 독립, 평균이 μ

$$E(\overline{X}) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) = \mu$$

* 표본분산

$$Var(\overline{X}) = \frac{1}{n^2}[Var(X_1) + \cdots + Var(X_n)] = \frac{1}{\sigma^2}$$

* 표준오차

$$SE = \sigma / \sqrt{n}$$

⇒ 우리가 알고자 하는 목표 μ 에서 $\overline{X}$ 가 떨어져 있는 정도, 즉 추정오차라고도 함

다. 표본평균의 분포

1) 정규모집단인 경우

* 정규근사규칙

SRS로 n 개의 관측값을 택했을 때 표본평균 $\overline{X}$ 는 모평균 μ 를 중심으로 변동하며 표준오차는 $\sigma / \sqrt{n}$ 이다(σ 는 모표준오차). 즉, n 이 커감에 따라 $\overline{X}$ 의 표본분포는 목표값 μ 에 더욱 더 밀집되어가며, 분포는 정규형태에 더욱 닮아간다.

2) 비정규모집단인 경우

* 중심극한 정리(Central Limit Theorem)

X 가 임의의 모집단으로부터 추출된 확률표본이라 하더라도, n 이 충분히 크다면 모집단의 분포에 관계없이 표본평균 $\overline{X}$ 는 근사적으로 정규분포 $N\left(\mu, \frac{\sigma^2}{n}\right)$ 을 따른다.

또한, $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ 의 분포는 표준정규분포 $N(0, 1)$ 을 따른다.

라. 비 율

1) 비율도 평균과 같이 정규 정규근사규칙이 적용된다.

표본비율 $\hat{p}$ 는 모집단 비율의 추정량으로서 $\overline{X}$ 와 마찬가지로 표본마다 값이 변하고 이 변동은 표본분포에 의해 설명된다.

> **비율에 관한 정규근사규칙** : 표본크기 n 인 단순 무작위 표본(SRS)에서 표본비율 $\hat{p}$ 는 모집단 비율 p 를 중심으로 변동하며 표준오차는 $\sqrt{p(1-p)/n}$ 이다. 따라서 $\hat{p}$ 의 표본분포는 N 이 증가함에 따라 모집단의 비율 p 에 집중되며 정규분포에 더욱 접근한다.

(이 규칙은 평균에 관한 정규근사규칙과 유사하다. 사실 표본비율 $\hat{p}$ 도 역시 표본평균의 일종이다. 이 내용에 대해서는 3항에서 더 자세한 설명을 하게 된다.)

정당의 지지율에 관한 예를 생각해 보자. 어떤 지역의 투표 인구 중에서 A당과 B당을 지지하는 사람이 각각 60%와 40%라고 한다. 이 지역으로부터 100명을 무작위로 추출하고 A당의 지지율 $\hat{p}$ 를 계산하였다. $\hat{p}$ 가 택할 수 있는 모든 가능한 값으로 이루어진 상대도수는 표본분포가 되며, 정규근사규칙으로부터 이 분포의 평균은 목표값 p=0.6이고 표준오차는 SE=$\sqrt{p(1-p)/n}$=$\sqrt{(0.60)(0.40)/100}$=0.05다. 이 결과로부터 우리는 $\hat{p}$ 에 관한 한 어떠한 확률도 계산할 수 있다.

2) 연속성 수정

비율에 관한 확률의 계산에서 정규근사는 문자 그대로 근사값을 제공해주기 때문에 오차가 있게 마련이다.

이제 15명 중에서 남아가 10명보다 많다는 것을 15명 중에서 남아가 10명 이상이라고 고칠 때 정규근사에 의한 확률을 계산해보자. 15명 중에서 남아가 10명 이상은 남아의 비율이 10/15 이상이고, 따라서 구하는 확률은 $P(\hat{p} \geq 10/15)$ 이므로 다음이 성립한다.

$$P(\hat{p} \geq 10/15) = P(Z \geq 1.29) = P(Z > 1.29)$$

$$= 0.099 \simeq 10\%$$

이 결과에 의하면 정규근사규칙을 사용할 경우 “10명보다 많다”는 확률과 ‘10명 이상’일 확률은 같다. 그러나 ‘10명 이상’과 “10명보다 많다”는 분명히 차이가 있다. 정수를 실수로 근사시키는 경우 ‘10명 이상’은 ‘9.5명 이상’으로 하고, “10명보다 많다”는 ‘10.5이상’이라고 하는 것

이 타당하다. 이 문제의 경우 “10명 보다 많다”이므로 정규근사를 위한 표준화 식을 다음과 같이 고쳐 쓸 수 있다.

$$Z = (10.5/15-0.5)/0.129 = 1.55$$

따라서 구하는 확률 $P(\hat{p} > 10/15)$의 정규근사값을 다음과 같이 구할 수 있다.

$$P\left(\hat{p} > \frac{10}{15}\right) \simeq P\left(\hat{p} \geq \frac{10.5}{15}\right) = P(Z \geq 1.55) = 0.0611 \simeq 6\%$$

이 값은 정확한 값과 소수점 이하 한자리에서 같은 것을 알 수 있다. 왜 이와 같은 결과가 나타나겠는가? 먼저 계산한 부정확한 근사값은 이산형분포(이항분포)를 연속형분포(정규분포)로 근사시켰기 때문에 발생한 것이다. 그러나 앞에서 계산한 것과 같이 연속성수정(continuity correction)을 적용시키면 참값에 매우 근접한 값을 얻을 수 있다.

3) 비율과 표본평균

변수의 값이 0-1 두 가지 중에서 하나의 값을 택하는 집계변수(counting variable) 또는 이진변수(binary variable) 또는 가변수(dummy variable)인 경우 다음이 성립한다.

> n개의 0-1로 이루어진 관측값들을 택했을 때 표본평균 $\overline{X}$는 1들을 모두 합하여 n으로 나눈 것이고 이것은 표본비율 $\hat{p}$과 같다. 따라서 $\overline{X}$의 기대값과 표준편차는 각각 μ, $\frac{\sigma}{\sqrt{n}}$ 이다. 여기서 μ =(모비율 p)이고 $\sigma = \sqrt{p(1-p)}$이다.

마. 모집단의 크기가 작은 경우

이미 설명한 바와 같이 모집단의 크기 N이 작고 관측값을 택한 후 이를 다시 넣지 않는다면(비복원 추출) 이렇게 택한 표본은 단순무작위 표본이 아니다. 그러나 한 번 택한 구슬을 그릇에 다시 넣지 않기 때문에 똑같은 구슬을 다시 택할 위험성이 없어진다.

예를 들어 어떤 학과의 학생들 중에서 10명을 무작위로 택하여 키를 측정한다고 하자. 우연히 첫 번째 택한 학생이 이 과에서 제일 키가 큰 농구선수로서 190cm였다고 하면 이 경우 한 학생으로 인하여 표본평균이 매우 크게 나타날 우려가 있다. 이 학생을 다시 넣고 9명을 택한다고 하면 190cm가 다시 나타날 수 있으며 표본평균이 비정상적으로 크게 나타날 수 있다. 그러나 비복원추출을 시행한다면 190cm의 특이값에 대해 우려할 필요가 없을 것이다. 따라서 비복원추출에 의해 계산된 표본평균은 복원추출에 의한 경우보다 변동이 작아진다.

일반적으로 N이 모집단의 크기이고, n이 표본크기일 때 비복원추출된 표본비율 $\hat{p}$의 분산은 다음과 같이 요약된다.

> (비복원 추출에 의한 표본비율의 분산) =
>
> $$\text{(복원추출에 의한 표본비율의 분산)} \times \frac{N-n}{N-1}$$

다시 말하면 비복원추출에 의한 $\hat{p}$의 표준오차는 복원추출에 의한 표준오차 $\sigma/\sqrt{n}$으로부터 다음과 같이 변한다.

$$\hat{p} \text{ 의 표준오차} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

그러나 $\sigma = \sqrt{p(1-p)}$ 이므로 다음이 성립한다.

비복원추출에 의한 표본비율의 표준오차

$$\hat{p} \text{ 의 표준오차} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

5. 추정

관심의 대상인 모집단으로부터 표본을 추출하여 모수에 대응하는 것들을 연구하고 모수에 대한 정보를 얻는다. 그리하여 좀 더 근거있는 여러 상황을 고려하고 가장 타당한 과학적 근거에 따라 결론을 얻어야 할 것이다. 이때 사용되는 과정이 통계적 추론(statistical inference)이다. 다시 말해서 통계적 추론이란 모집단으로부터 표본을 추출하여 자료를 얻고 이 자료를 이용하여 모집단의 미지의 값을 추측하거나 결정하는 과정을 말한다. 통계적 추론은 크게 추정(estimation)과 가설검정(hypothesis testing)의 두 분야로 분류된다. 추정은 모집단에서 추출한 표본의 데이터를 이용하여 모집단의 어떤 미지의 값을 추측하는 과정을 말하고 가설검정이란 모집단에서 추출한 표본을 이용하여 모집단에 대한 가설을 세우고 이 가설의 채택이나 기각을 결정하는 과정을 말한다.

모집단의 모수에 대한 추정은 항상 표본 관측치의 함수인 표본통계량에 의해 행한다. 이 때 모수를 추정하는 공식을 나타내는 표본통계량을 추정량(estimator)이라 하고, 이 함수에 실제의 관측치를 넣어 계산한 표본통계량의 값을 추정치(estimate)라고 한다. 추정량은 확률 변수로써 표본분포에 따라 변하고 추정치는 특정 표본에 의해 정해지는 상수이다.

가. 점추정

1) 점추정과 추정량의 성질

미지의 모수를 추정하는 방법에는 두 가지가 있다. 하나는 추출된 표본을 이용하여 모수를 하나의 수치로 추정하는 점추정(point estimation)이고 , 다른 하나는 모수가 포함되리라고 기대하는 범위를 구하는 구간추정(interval estimation)이다.

점추정을 하는 목적은 추출된 표본으로부터 모수에 가장 가까운 하나의 수치를 추정하는 것이다. 모수를 추정하는 통계량, 즉 추정량은 여러 개 있을 수 있다. 예를 들어 모평균의 추정량으로 표본평균이나 표본중앙치 또는 최빈치를 사용할 수 있다. 그러나 그 중 어느 것을 선택할 것인가가 문제이다.

일반적으로 모수를 추정하기 위한 통계량의 선택에서 좋은 추정량의 조건으로 다음 네 가지를 들 수 있다.

① 불편성(Unbiased)
② 일치성(Consistency)
③ 유효성(Efficiency)
④ 충분성(Sufficiency)

가) 불편성

모집단의 모수에 대한 추정량의 기대치가 모집단의 모수와 일치할 때 이 추정량을 그 모수의 불편추정량이라 한다. 모수 Θ의 추정량 $\hat{\Theta}$ 가 다음 조건을 만족시킬 때, $\hat{\Theta}$ 를 모수 Θ의 불편추정량이라 한다.

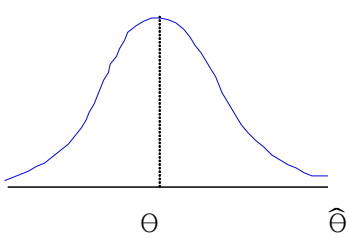
$$E(\hat{\theta}) = \theta$$

추정량의 기대치와 모수의 차이, 즉 편의가 0인 추정량을 불편추정량이라 하고, 편의가 0 이 아닌 추정량을 편의 추정량이라 한다.

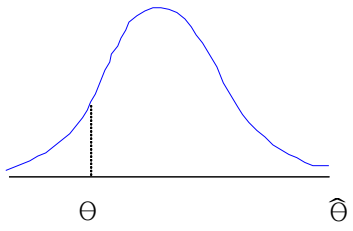
불편추정량의 예로써, 다음과 같은 예가 있다.

$$E(\overline{X}) = \mu$$

$$E(S^2) = \frac{1}{n-1} E[\sum (X_i - \overline{X})^2] = \sigma^2$$



불편추정량



편의추정량

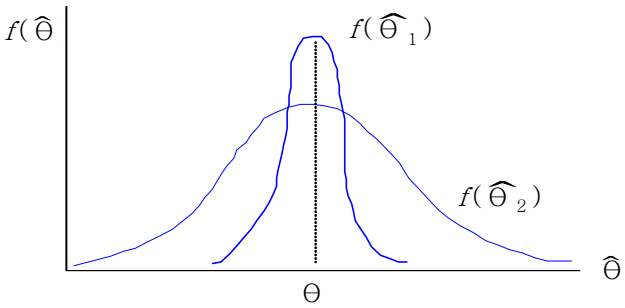
나) 유효성

불편추정량은 추정량이 따르는 분포의 중심에 대하여 요구되는 성질을 갖춘 추정량이지만, 그 분포에 대한 산포의 정도를 나타내지는 못한다.

좋은 추정량이 갖추어야 할 기본적인 조건으로 표본오차가 작아야 된다는 것이다. 이 조건은 추정량의 편의와 분산이 다 같이 작을 때 충족된다. 편의가 없더라도 분산이 너무 크거나, 분산이 작더라도 편의가 큰 추정량은 큰 표본오차를 수반하기 때문이다. 이런 의미에서 좋은 추정량이란 평균평방오차(mean squared error) $MSE = E(\hat{\Theta} - \Theta)^2$ 이 최소가 되는 추정량을 칭하게 된다.

$$MSE(\hat{\Theta}) = E[\hat{\Theta} - E(\hat{\Theta})]^2 + [E(\hat{\Theta}) - \Theta]^2 = Var(\hat{\Theta}) + (\text{편의})^2$$

불편추정량 중에서 분산이 최소인 추정량을 최소분산불편추정량 또는 유효추정량, 또는 최량추정량이라 한다.



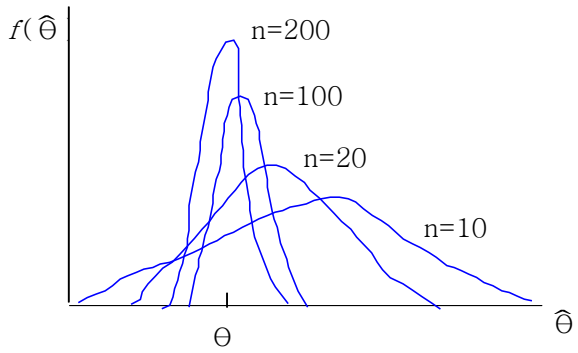
다)일치성

좋은 추정량이란 또한 표본크기가 커질수록 표본오차가 작아지는 성질을 가져야 된다. 일치추정

량이란 표본크기 n이 무한히 증가할 때 표본에서 구한 추정량 $\hat{\theta}$ 가 모수 θ 와 일치하는 추정량을 말한다. θ_n 을 크기 n 인 표본에서 계산된 추정량이라 할 때,

$$\lim_{n\to\infty} E(\hat{\theta}_n - \theta)^2 = 0, \qquad \lim_{n\to\infty} \hat{\theta}_n = \theta$$

의 조건이 성립하면 $\hat{\theta}_n$ 를 θ 의 일치 추정량이라 한다.



라) 충분성

동일한 표본으로부터 얻은 추정량 $\hat{\theta}$ 가 모집단의 모수 θ 에 관한 모든 정보와 지식을 포함하고 있을 때 이러한 추정량을 충분추정량이라 한다.

$$\hat{\theta} \rightarrow \theta$$

지금까지 좋은 추정량이 되기 위한 4가지 조건을 살펴보았다. 점추정치를 실제로 계산하는 방법은 이러한 조건을 만족하는 추정량 즉, 최우추정량을 찾아 특정표본에서 이들 추정량의 값을 계산하고 모집단 모수의 추정치로 사용하는 것이다.

2) 모평균의 추정

모평균이 μ 이고 모표준편차가 σ 인 임의의 모집단으로부터 크기 n인 임의표본을 $X_1, X_2, \cdots, X_n$이라 할 때,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

을 각각 표본평균, 표본분산이라 하고, 표본분산의 양의 제곱근 s를 표본표준편차라고 한다. 표본평균 $\overline{X}$에 대하여 다음이 성립한다.

$$E(\overline{X}) = \mu, \quad \sigma_{\overline{X}} = \sigma/\sqrt{n}$$

위의 첫 번째 결과는 모평균 μ 의 추정량인 $\overline{X}$의 기대값은 다시 μ 가 된다는 사실을 보여준다. 따라서 표본평균 $\overline{X}$는 모평균 μ 의 불편추정량이고 표준오차가 $\sigma/\sqrt{n}$ 임을 보여준다. 모표준편차 σ 를 모를 경우에는 표준오차의 공식 $\sigma/\sqrt{n}$ 을 구하는 어려움이 있으므로 모표준편차 s를 사용한다. 따라서 표준오차의 추정량은 $s/\sqrt{n}$ 가 된다.

모평균 μ의 점추정

추 정 량 : 표 본 평 균 $\hat{\mu} = \overline{X}$

표 준 오 차 : $\sigma_{\overline{X}} = \sigma/\sqrt{n}$

표 준 오 차 의 추 정 량 : $S_{\overline{X}} = S/\sqrt{n}$

예제) 어느 건전지회사에서 생산된 건전지의 평균수명을 측정하기 위해 6개의 건전지를 임의로 선택해서 수명시간을 측정한 결과는 다음과 같다.

80, 100, 90, 120, 70, 110

이 자료로부터 건전지의 평균수명 μ의 추정치와 표준오차의 추정치를 구하라.

$$\hat{\mu} = \overline{X} = (80+100+90+120+70+110)/6 = 95$$

$$S = \sqrt{\sum (X_i - \overline{X})^2/(n-1)} = \sqrt{\sum (X_i - 95)^2/5}$$

$$= \sqrt{1750/5} \fallingdotseq 18.71$$

$\overline{X}$의 표준오차의 추정치 $S_{\overline{X}} = s/\sqrt{n} = 18.71/\sqrt{6} = 7.6383$

3) 모분산의 추정

모분산은 모집단의 분포가 흩어져 있는 정도를 나타내는 양으로 표본분산, 평균편차, 범위 등의 통계량을 모분산에 대한 정보를 나타내는데 이용할 수 있다.

이 중 가장 많이 사용되는 표본분산을 다음과 같이 정의한다.

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

이 사실로부터 다음 식이 성립함을 알 수 있다.

$$E(s^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \overline{X})^2\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \mu) - (\overline{X} - \mu)^2\right]$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2\right]$$

$$= \frac{1}{n-1} \left[n\sigma^2 - n \cdot \frac{\sigma^2}{n}\right]$$

$$= \frac{1}{n-1} (n-1)\sigma^2 = \sigma^2$$

위의 식이 성립하므로 표본분산 s^2은 모분산 σ^2의 불편추정량이다.

모분산의 점추정

모분산의 추정량은 다음과 같다.

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

4) 모비율의 추정

불량률, 실업률, 지지율과 같이 모집단에서 어떤 특정한 속성을 가진 것의 비율p를 모비율(population proportion) 이라 한다.

모집단에서 크기가 n인 표본을 임의로 추출할 때 이 표본에서 특정한 속성을 가진 것의 개수를 X라고 하자. 표본비율은 $\hat{p} = \frac{X}{n}$ 이 고 $\hat{p}$ 를 모 비 율 p의 추정량으로 생각 할 수 있다.

실제로 X는 이항분포 $b(n,p)$를 따르므로 $E(X) = np$, $Var(X) = npq$ 가 성립한다. 여기서 $q = 1 - p$ 이다.

따라서 표본비율 $\hat{p} = \frac{X}{n}$ 에 대하여 다음이 성립한다.

$$E(\hat{p}) = \frac{1}{n} E(X) = \frac{np}{n} = p$$

$$Var(\hat{p}) = \frac{1}{n^2} Var(X) = \frac{npq}{n^2} = \frac{pq}{n}$$

위의 사실로부터 $\hat{p}$는 p의 불편추정량임을 알 수 있다.

비율의 점추정

점 추 정 량 : 표 본 비 율 $\hat{p} = X/n$

표 준 오 차 : $\sigma_{\hat{p}} = \sqrt{pq/n}$

표 준 오 차 의 추 정 량 : $S_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$

나. 구간추정

1) 구간추정과 신뢰구간의 개념

표본의 특성치를 가지고 모집단의 특성치를 유추하는 것을 통계치로부터 모수를 추정한다고 말한다. 이때의 표본은 모집단으로부터 무작위로 추출함으로써 모집단을 대표할 수 있는 것이어야 한다. 그러나 무작위로 뽑혀진 표본도 표본오차가 있으므로 정확하게 모집단을 대표하지는 못한다. 최선의 방법은 모수를 추정하면서 얼마만큼의 표본오차가 있는가를 밝히는 것이다. 이 방법에 도움을 주는 것이 표본분포이론이다.

미지의 모수에 대한 추정으로써 하나의 점만을 사용하는 점추정의 경우에는 그 추정량의 산포도에 관한 정보가 없기 때문에 어느 정도 정확성이 있는지를 판단하기 어렵다. 일반적으로 점추정치가 좋은 추정량의 조건을 갖추었다고 하더라도 표준오차 때문에 이 점추정치가 반드시 모수와

일치한다고 할 수 없다. 점추정은 추정의 정확도를 보여 줄 수 없기 때문에 점추정치를 중심으로 구간을 설정해 줄 필요가 있다.

구간추정은 모수가 포함되리라고 생각되는 범위, 즉 구간을 설정하여 모수를 추정하는 방법이다. 이 때 표본으로부터 계산된 통계치를 중심으로 하여 모수가 일정한 확률을 갖고서 포함되리라고 기대되는 구간을 신뢰구간(confidence interval)이라 한다. 신뢰구간의 상한과 하한의 값을 신뢰한계(confidence limits)라고 한다. 그리고 이 구간내에 모수의 실제 값이 존재할 확률을 신뢰수준(confidence level) 또는 신뢰계수라고 한다.

신뢰구간이 길어지면 신뢰수준이 높아지고 신뢰구간이 짧아지면 신뢰수준이 낮아진다. 신뢰구간이 길어지면 신뢰수준은 높아지나 정보로서의 가치는 떨어지므로 정보로서의 가치와 신뢰수준을 잘 고려하여 신뢰구간의 길이를 결정해야 한다.

모수 Θ 에 대한 신뢰구간은 일반적으로 두 값 L과 U에 의해

$$P(L \leq \Theta \leq U) = 1 - \alpha$$

와 같은 형식으로 나타낸다. 여기사 L과 U를 신뢰한계라 하고, 1-α를 신뢰수준이라 한다. 신뢰수준은 모수 Θ가 L과 U사이에 있을 확률이 100(1-α)%라는 것을 의미한다. 신뢰구간 (L, U)를 모수 Θ의 100(1-α)% 확률구간이라고도 한다. 그리고 α는 주어진 구간내에 모수를 포함하지 못할 확률을 나타내며 이것을 오차율(probability of error)이라 한다.

2) 평균의 구간추정

모집단의 평균에 관한 추정에서는 표본평균 $\overline{X}$가 추정량으로 이용된다. 앞에서 배운 표본평균의 분포에 대한 내용을 정리하면 다음과 같다.

(1) 확률변수 X가 모집단 평균 μ, 분산 σ^2인 정규분포를 이루면, 크기 n인 표본의 표본평균 $\overline{X}$는 평균이 μ이고, 분산이 σ^2/n인 정규분포를 이룬다.

$$X \sim N(\mu, \sigma^2) \Rightarrow \overline{X} \sim N(\mu, \sigma^2/n)$$

(2) 위의 항에서 모집단의 분포가 정규분포가 아니더라도 표본의 크기 n이 크다면, 표본평균 $\overline{X}$의 분포는 정규분포에 근사하다고 할 수 있다.

(3) 모집단의 분포가 정규분포를 이루고 크기가 n인 표본의 분산을 S^2이라고 한다면, $\sqrt{n}(\overline{X} - \mu)/S$는 자유도 (n-1)의 t 분포를 이룬다.

표본평균 $\overline{X}$의 표본분포는 모분산이 알려져 있는지에 따라, 표본의 크기에 따라 달라진다. 모집단의 평균에 대한 구간추정은 상황별로 구분해서 다루어져야 한다.

가) 모분산 σ^2을 알고 있을 경우

표준정규분포에서 오른쪽 끝 면적이 α/2 되는 Z값을 $Z_{\alpha/2}$, 그 반대쪽의 Z값을 $-Z_{\alpha/2}$로 한다면,

$$P(-Z_{\alpha/2} \leq (X - \mu)/\sigma \leq Z_{\alpha/2}) = 1 - \alpha$$

그리고 표본정규변수 $Z = (X - \mu)/\sigma$ 이므로

$$P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1 - \alpha$$

모집단이 정규분포이거나 정규분포가 아니더라도 표본크기 $n \geq 30$인 경우에는 $\overline{X}$도 정규분포를 따르므로,

$$P(-Z_{\alpha/2} \le (\overline{X}-\mu)/\sigma_{\overline{X}} \le Z_{\alpha}) = 1-\alpha$$

$$P(\overline{X}-Z_{\alpha/2}\cdot\sigma_{\overline{X}} \le \mu \le \overline{X}+Z_{\alpha}\cdot\sigma_{\overline{X}}) = 1-\alpha$$

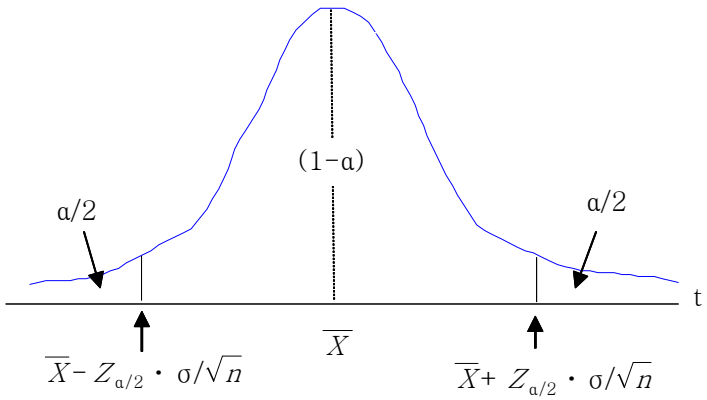
여기서 $\sigma_{\overline{X}} = \sigma/\sqrt{n}$ 이므로

$$P(\overline{X}-Z_{\alpha/2}\cdot\sigma/\sqrt{n} \le \mu \le \overline{X}+Z_{\alpha}\cdot\sigma/\sqrt{n}) = 1-\alpha$$

따라서 모평균 μ 의 신뢰수준 100(1-α)%로 포함될 신뢰구간은

$$(\overline{X}-Z_{\alpha/2}\cdot\sigma/\sqrt{n}\ ,\ \overline{X}+Z_{\alpha}\cdot\sigma/\sqrt{n})$$

이다.



예제) 다음 자료는 어느 해 중학교 3학년생의 임의로 40명을 뽑아 모의고사 성적을 조사하였다. 이 자료를 토대로 95% 신뢰구간을 구하여 보자. 이때 모의고사 성적의 표준편차는 4이다.

119	125	126	128	132	135	135	135	136	138
138	140	140	142	142	144	144	145	145	146
146	147	147	148	149	150	150	152	153	154
156	157	158	161	163	164	165	168	173	176

풀이) 주어진 자료로부터 $\overline{X}$ = 139.75, σ = 4이고, 95%의 신뢰구간을 구하고자 할 때 임계값은 $Z_{\alpha/2} = Z_{0.025} = 1.96$ 이므로

$$\left(\overline{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(139.75-1.96\frac{4}{\sqrt{40}},\ 139.75+1.96\frac{4}{\sqrt{40}}\right)$$

$$= (138.5296,\ 140.9896)$$

나) 모분산 σ^2 을 모르고 소표본 ($n \le 30$)인 경우

모집단의 분포가 정규분포를 이루고 크기가 n인 표본의 분산을 S^2이라고 한다면, $\sqrt{n}(\overline{X}-\mu)/S$는 자유도 (n-1)의 t 분포를 갖는다는 것을 알고 있다. $n \le 30$인 경우에는 자유도 (n-1)인 t 분포를 이용한다.

$$t = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}\text{분 포}$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$$

따라서

$$P(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha$$

$$P(-t_{\alpha/2} \le \sqrt{n}(\overline{X} - \mu)/S \le t_{\alpha/2}) = 1 - \alpha$$

$$P(\overline{X} - t_{\alpha/2} \cdot s/\sqrt{n} \le \mu \le \overline{X} + t_{\alpha/2} \cdot s/\sqrt{n}) = 1 - \alpha$$

그러므로 모평균 μ의 신뢰수준 100(1-α)%로 포함될 신뢰구간은

$$(\overline{X} - t_{\alpha/2} \cdot s/\sqrt{n} \ , \ \overline{X} + t_{\alpha/2} \cdot s/\sqrt{n})$$

그리고 모집단이 정규분포를 따르지 않는 경우에도 표본의 크기가 큰 경우 ($n \ge 30$)에는

$$(\overline{X} - t_{\alpha/2} \cdot s/\sqrt{n} \ , \ \overline{X} + t_{\alpha/2} \cdot s/\sqrt{n})$$

을 모평균 μ의 100(1-α)%로 포함될 신뢰구간으로 사용할 수 있다.

예제) 어느 건전지회사에서 생산된 건전지의 평균수명을 측정하기 위해 6개의 건전지를 임의로 선택해서 수명시간을 측정한 결과는 다음과 같다.

80, 100, 90, 120, 70, 110

이 자료로부터 모평균의 95% 신뢰구간을 구하여라.

풀이) 위 자료로 부터 $\overline{x} = 95$, $s = 18.71$ 이므로 $t_{\alpha/2}(n-1) = t_{0.025}(5) = 2.571$

신뢰구간은

$$\left(\overline{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$$

$$= \left(95 - 2.571\frac{18.71}{\sqrt{6}}, \ 95 + 2.571\frac{18.71}{\sqrt{6}}\right)$$

$$= (75.365, \ 114.364)$$

다) 모분산 σ^2을 모르고 대표본 ($n \ge 30$)인 경우

모집단의 분포가 정규분포가 아니더라도 표본의 크기 n이 크다면, $\sqrt{n}(\overline{X} - \mu)/S$의 표본분포는 t 분포에 근사하다는 것을 알고 있다. 이 때 모분산을 모르면 추정량 S^2으로 대체하게 되는데, 표본의 크기가 크면 $t = \sqrt{n}(\overline{X} - \mu)/S$ 의 표본 분포는 표본정규분포에 접근하게 된다. 따라서

$$P(-Z_{\alpha/2} \le \frac{X - \mu}{S/\sqrt{n}} \le Z_{\alpha}) = 1 - \alpha$$

$$P(\overline{X} - Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \le \mu \le \overline{X} + Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}) = 1 - \alpha$$

모평균 μ의 신뢰도 100(1-α)%로 포함될 신뢰구간은

$$(\overline{X} - Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \ , \ \overline{X} + Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}})$$

예제) 어느 학급의 학생들의 성적은 평균 μ, 분산 σ^2인 정규분포를 이룬다고 한다. 크기 $n=30$인 임의표본을 추출하여 측정한 결과 $\overline{X}=78$, $S=10$을 얻었다. 모평균의 90%신뢰구간을 구하여라.

풀이) $n \geq 30$이므로 $t=\sqrt{n}(\overline{X}-\mu)/S$은 정규분포를 따른다.

그러므로 신뢰구간은 1

$$(\overline{X}-Z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X}+Z_{\alpha/2}\frac{S}{\sqrt{n}})$$

$$=(78-1.645\frac{10}{\sqrt{30}},\ 78+1.645\frac{10}{\sqrt{30}}) = (75,\ 81)$$

이다.

3) 분산의 구간추정

모분산의 구간추정은 χ^2분포를 이용해서 추정하게 된다. 모분산의 추정량으로는 표본분산 S^2을 이용한다.

확률표본 n개의 변수 $X_1, X_2, \cdots, X_n$이 정규모집단 $N(\mu, \sigma)^2$으로 부터의 확률분포이라고 할 때, 표본분산 $S^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$에 대해

$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i-\overline{X})^2}{\sigma^2}$ 은 자유도 $(n-1)$인 χ^2분포를 따른다.

따라서

$$p[\chi^2_{(n-1,1-\alpha/2)} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{(n-1,\alpha/2)}] = 1-\alpha$$

$$p\left[\frac{(n-1)s^2}{\chi^2_{(n-1,\alpha/2)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}}\right] = 1-\alpha$$

따라서 , 모분산 σ^2의 신뢰도 100(1-α)% 신뢰구간은

$$\left[\frac{(n-1)s^2}{\chi^2_{(n-1,\alpha/2)}},\ \frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}}\right]$$

이다.

예제) 모집단에서 임의로 추출한 크기 15인 표본의 표본분산 $S^2=5$, 표본평균 $\overline{X}$=11이었다. 모집단이 정규분포를 따른다면 모분산에 대한 99% 신뢰구간을 구하여라.

풀이) $n=15, S^2=5$, $\overline{X}=11$, $\alpha=0.01$ 이고

$\chi^2_{(n-1,\alpha/2)}=\chi^2_{(14,0.005)}=31.3193$

$\chi^2_{(n-1,1-\alpha/2)}=\chi^2_{(14,0.995)}=4.07468$ 이다.

신뢰구간은

$$\left[\frac{(n-1)s^2}{\chi^2_{(n-1,\alpha/2)}},\ \frac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}}\right] = \left[\frac{14*5}{31.3193},\ \frac{14*5}{4.07468}\right]$$

$$= (2.2350,\ 17.1793)$$

이다.

4) 비율의 구간추정

확률표본 $X_1, X_2, \cdots, X_n$이 성공의 확률 p인 베르누이분포를 따를 때, $Y = \sum X_i$는 이항분포 $B(n,p)$를 따르고 모비율의 점추정으로 $\hat{p} = \overline{X} = \frac{Y}{n}$가 사용되는 것을 알고 있다.

표본크기 n이 크고 p가 0.5에 가까울 때는 ($np \geq 5$), 표본비율 $\hat{p} = \overline{X} = \frac{Y}{n}$는 평균 $E(\hat{p}) = p$, 표준편차 $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$인 정규 분포를 따른다는 사실을 알고 있다. 이를 이용하여 모비율 p에 대한 구간추정문제를 생각해 보자.

중심극한정리에 의해 아래 값은 표준정규분포를 따른다.

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

분모의 p 대신에 $\hat{p}$를 대입하면 근사적으로 정규분포를 따른다.

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

그러므로,

$$p\left[-Z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq Z_{\alpha/2}\right] \fallingdotseq 1-\alpha$$

$$p\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \fallingdotseq 1-\alpha$$

모비율 p의 신뢰수준 100(1-α)% 신뢰구간은

$$\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

이다. 그러나 표본크기 n이 작을 때에는 이항분포는 정규분포에 근사하지 않으므로 신뢰구간의 값은 이항분포식에 따라 계산되어야 한다.

예제) 어느 도시의 갑 시의원에 대한 지지율을 알아보고자 한다. 이 지지율을 알고자 1000명을 조사한 결과 382명의 지지를 받고 있다고 한다. 모비율 p에 대한 95% 신뢰구간을 구하라.

풀이) 표본비율 $\hat{p} = \frac{382}{1000} = 0.382$가 되고, 표준오차는

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.382*(1-0.382)}{1000}} = 0.015$$

이다.

따라서 95% 신뢰구간은

$$\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$= (0.382 - 1.96 * 0.015, 0.382 + 1.96 * 0.015)$$

$$= (0.353,\ 0.411)$$

이다.

6. 여러 가지 통계 기법

가. 분산분석의 소개

일상생활 주변에서는 두 개 이상의 모집단의 평균을 동시에 비교해야 하는 경우가 흔히 발생하며, 이를 위해 사용되는 통계적 기법을 **분산분석(analysis of variance : ANOVA)**이라고 한다.

우리나라의 대학들은 각각의 인재상에 부합하는 학생을 선발하기 위하여 전형 시기별로 다양한 전형방법들을 개발하고 있다. 또한 전형시기별 전형요소의 강조점이 구분되어 있는 것이 보통이며 시기별 모집인원 비율도 고려해야 한다. 이 때 전형시기에 따라 입학한 학생들의 학점을 비교함으로써 전형시기별 선발방법이 대학에서의 학업성취도에 영향을 미치는가를 분석할 수 있다. 예를 들어 전형시기를 수시1차, 수시2차, 정시로 구분한다고 하자. 여기에서 전형시기는 독립변수가 되고 학점이 종속변수가 되어 결국 분산분석이란 독립변수와 종속변수의 관계를 분석하는 기법인 것이다. 또한 독립변수를 인자(factor)라고도 하며 앞의 예에서는 세 가지 상태 즉, 수시1차, 수시2차, 정시의 인자수준(factor level)이 있다고 할 수 있다.

고등학교에서는 반별 학업성취도의 차이, 심화수업을 받은 것이 학업성취도에 영향을 미치는지, 임원이나 동아리와 같은 비교과적인 활동을 많이 한 집단과 그렇지 않은 집단간의 성적차이가 발생하는지, 자율학습에 참여한 학생과 그렇지 않은 학생간의 성적차이는 있는지 등을 분석하기 위한 통계적 기법이다.

분산분석에는 단 하나의 인자를 분석대상으로 하는 일원분산분석(one-way ANOVA)과 두 개의 인자를 분석대상으로 하는 이원분산분석(two-way ANOVA)이 있다. 두 개 이상의 인자를 분석대상으로 하는 경우를 통틀어 다원분산분석(multi-way ANOVA)이라고 하기도 한다.

나. 상관분석의 소개

이제까지 주어진 자료를 설명하는 통계적인 기법들과 자료가 어떻게 분포되어 있는지도 살펴볼 수 있었고, 여러 가지의 기술통계량을 통하여 자료의 통계적인 특성들도 파악할 수 있었다. 이들은 모두 하나의 조사대상 즉 객체에 대하여 하나의 측정값만을 조사한 것이다. 그런데 현실세계에서 발생하는 자료들을 살펴보면 대부분의 경우에 둘 이상의 자료들이 서로 연관을 가지고 변화하는 것을 알 수 있다. 예를 들어 한 고등학교 3학년 학생들의 내신성적과 수능성적을 생각해 보자. 우리는 일차적으로 학생들의 평균내신성적과 평균수능성적에 관심이 있어 이를 계산할 것이다. 그 다음은 학생들의 내신성적과 수능성적이 어떠한 관계를 지니고 있는가에 관심을 가질 것이다. 즉, 내신성적과 수능성적의 관계가 비례하는지 반비례하는지 아니면 아무런 관계가 없는지를 알고자 할 것이다. 이러한 관계를 알아보기 위하여 사용되는 통계적 기법이 **상관분석(correlation analysis)**이다.

두 변수 사이의 관계에는 여러 가지가 있을 수 있다. 예를 들어 선형관계가 있을 수 있고, 2차

함수관계가 있을 수 있다. 상관분석은 두 변수의 여러 가지 관계 중 선형관계만을 분석한다. 즉 내신성적을 X축, 수능성적을 Y축에 표시할 때, 이들이 선형관계를 이루는가를 분석한다. 두 변수 X, Y가 일직선상에 가깝게 놓이게 되면 상관관계가 높다고 하고, 일직선과 멀리 떨어져 있으면 상관관계가 낮다고 한다. 예를 들어 어떤 두 변수 X, Y를 그래프상에 그려보니 정확히 2차함수관계를 이룬다는 사실을 확인했다고 하자. 그러나 두 변수 사이의 상관분석에서는 두 변수 간에 아무런 관계가 없다는 결론이 도출될 것이다. 왜냐하면 상관분석은 두 변수 사이에 선형관계가 존재하는지 존재하지 않는지를 측정하기 때문이다.

상관분석을 할 경우에는 하나의 조사대상에 대하여 반드시 2개의 변수를 측정하여야 한다. 하나의 조사대상 즉 실험단위가 지니고 있는 2개의 변수들에 대한 상관관계를 측정하는 것이다. 내신성적과 수능성적의 관계를 상관분석 하는데 있어서 조사대상을 달리하여 조사하면 이는 상관분석이 아니다. 예를 들어 인문계 학생에 대하여 내신성적을 조사하고 자연계 학생에 대하여 수능성적을 조사한다면 여기서 조사된 내신성적과 수능성적은 상관분석과 아무런 관계가 없는 것이다. 두 변수를 산포도로 나타낼 때, 그래프에 나타난 좌표 하나하나가 조사대상 즉 객체가 되어야 한다.

다. 회귀분석의 소개

연관된 자료를 분석하는 방법으로는

1. 두 변량 사이의 관계를 계수를 통하여 알아보는 상관분석과
2. 하나의 변량과 이에 영향을 미치는 여러 변량에 대한 관계를 모형을 통하여 알아보는 회귀분석이 있다.

연관관계를 갖는 두 변수 또는 변량이 서로 원인과 결과, 즉 인과관계를 갖는 경우에 대하여 생각해보자. 인과관계란 하나의 변량에 대한 값이 다른 변량의 값에 따라 영향을 받는 경우를 의미한다. 하나의 변량이 다른 하나 이상의 변량에 영향을 받는 경우 이러한 인과관계를 적당한 함수로 표현하고 이를 통하여 자료를 설명하고 예측하는 분석방법을 **회귀분석(回歸分析: Regression Analysis)**이라고 한다. 예를 들어 내신성적(x)에 따른 수능성적(y)의 변화 정도를 설명하는 것이 회귀분석의 주요 내용이다. 이때 내신성적을 독립변수라 하고, 수능성적을 종속변수 또는 반응변수라 한다. 독립변수가 하나이고 독립변수와 종속변수와의 관계가 선형인 경우를 단순선형회귀분석이라고 한다.

라. 표본조사에서 항목 무응답 대체 방법

각종 통계 자료는 경제·사회현상을 파악하는데 중요한 근거로 이용되고 있다. 이러한 통계자료를 생산하기 위한 표본조사에는 표본오차와 비표본오차가 발생하는데, 비표본오차는 표본조사의 조사 기획 단계부터 실제 자료 분석 단계 등 전체 표본조사과정에서 다양한 형태의 실수 또는 결함에 의하여 발생하는 오차이다. 특히 무응답은 비표본오차를 발생시키는 중요한 요인 중 하나이다.

무응답에 의한 오차는 두 가지로 나뉠 수 있는데 하나는 단위 무응답(unit nonresponse)으로 조사단위로부터 얻어진 정보가 전혀 없는 경우를 의미하며 다른 하나는 항목 무응답(item nonresponse)으로 조사단위가 표본조사에 참가는 했지만 질문 중 몇 가지 항목에 대한 대답을 얻지 못한 경우를 말한다. 이와 같은 무응답의 종류에 따라 적절한 처리방법을 고려해야 하는데 본 강좌에서는 항목 무응답에 대한 대체 방법에 대하여 몇 가지 소개하고자 한다.

(1) 연역적 대체 방법

연역적 대체 방법(detective method)이란 결측값에 대해서 현재 자료로부터 확실시되거나, 확신

을 갖고 결측값을 유추할 수 있는 경우에 사용될 수 있는 방법이다. 예를 들어 가족들 중 가장의 인종에 대한 응답이 없는 경우에 다른 가족들이 모두 백인이라면 이 사람도 백인이라고 생각하여 백인으로 이 결측값을 대신하는 방법이다.

그러므로 이 방법은 보조 변수가 결측값을 결정하는데 있어서 거의 오차를 가지고 있지 않다고 볼 수 있는 경우에 사용이 가능한 방법으로 만약 이 오차가 너무나 커서 무시할 수 없다면 이 방법 이외에 다른 방법이 적용되어야 한다.

(2) 정확한 대응 대체

정확한 대응 대체(exact match imputation)는 경우에 따라 결측된 정보를 다른 조사자료로부터 얻을 수 있는 경우 결측값과 동일한 조사단위에 해당하는 다른 외부 자료의 값으로 대체하는 방법을 의미한다. 이 방법은 대체값이 다른 외부 조사결과로부터 얻어지는 값이므로 그 외부 자료가 신뢰할 수 있어야 하고, 조사단위를 대응시키는데 발생하는 비용이 많이 들고 사용하기에 복잡하다는 제약이 있다. 아울러 일부 결측값에 대하여는 대응된 조사단위를 찾을 수 없는 경우가 빈번히 발생하는 문제점이 있다.

(3) 핫덱

지역, 직업, 성별 등 적절한 기준에 따라 조사된 자료를 순서대로 입력한 자료화일을 작성하는데, 이 과정에서 만약 대체를 고려하는 변수가 무응답인 경우 자료화일상의 입력 순서에 입각하여 바로 앞에 입력되는 응답자에 대한 관측값으로 무응답을 축차적으로 대체한다. 자료화일의 순서에 따른 이런 전통적인 대체방법을 일반적으로 축차 핫덱(sequential hot-deck)이라 부른다.

이 방법은 컴퓨터기술이 낙후된 상황에서 자료처리상의 편리성과 비용이 중요하게 고려된 방법이다. 만약 자료의 순서와 조사 항목의 값이 우연히 양의 상관관계이면 매우 효과적인 방법이나, 몇 가지의 심각한 단점을 가지고 있다. 첫째, 결측값을 할당하는데 있어서 확률구조가 아닌 자료화일 상의 순서에 의존한다는 점과 둘째, 동일한 제공값을 여러 번 사용하게 될 수 있다는 것이다.

(4) 평균 대체

평균 대체(mean-value imputation) 방법은 표본을 대체층으로 나눈 후에 각 층내에서 응답자들의 평균을 구하여 그 층의 모든 무응답에 이 평균을 삽입하는 것이다.

이 방법은 동일한 대체층 내에서 결측값이 모두 한 개의 값 즉, 평균으로 대체됨으로 인해 관심변수의 경험적 분포가 상당히 왜곡된다는 큰 단점이 있지만 사용하시 쉽고 평균이나 총합들과 같은 일변량 모수에 대한 추정에 있어 무응답 편의를 감소시키는 데는 상당히 효과적이어서 간단한 점추정이나 예산상의 제약이 있을 때 주로 사용되는 방법이다.

이 방법을 가장 간단하게 적용한 것으로 대체층을 고려하지 않고 모든 결측값 대신에 전체응답자의 평균을 대체하는 방법을 고려할 수 있다. 이 방법은 상당한 문제점을 내포하고 있지만 실제적으로 많은 사회과학조사에서 대체 방안으로 사용되고 있는 것이 현실이다.

(5) 랜덤 대체

랜덤 대체(random imputation) 방법은 대체층 내에서 제공값을 확률추출(probability sampling)에 의해서 선택하여 그 값으로 결측값을 대체하는 것이다. 이 방법의 최대 장점은 평균 대체 방법을 사용하는 경우 분포를 왜곡시킨다는 문제점을 어느 정도 해결할 수 있다는 것이다.

(6) 회귀대체

회귀대체(regression imputation)는 무응답이 있는 항목을 종속변수로 응답된 보조변수들을 독립

변수로 하는 회귀모형을 적용하는 방법으로 독립변수들은 질적 변수이거나 양적 변수이어도 상관이 없다. 관심 변수가 질적 변수인 경우에는 대수 선형 또는 로지스틱 모형을 적용할 수 있다. 이 방법은 현 자료의 값을 그대로 대체값으로 사용하는 것이 아니라 회귀모형을 통해 얻게 되는 예측값을 사용한다는 측면에서 다른 대체 방법들과 큰 차이가 있다.

마. 평가결과 조정방법

1) 평가결과 조정의 필요성

(1) 부정확한 정보, 편견, 주관성 등으로 인한 판단의 오류가 평가 과정에 수반되므로 이를 조정하여 평가의 정확성을 높일 필요가 있음

(2) 관대화 경향 또는 엄격화 경향을 보이는 평가자에게 평가를 받은 피평가자는 상대적으로 이득 또는 손해를 볼 수 있으므로 평가의 공정성을 위해 평가결과를 조정하여야 함

(3) 평가의 신뢰성을 확보하기 위해 평가 기준 이상으로 지나치게 **호의적이거나 악의적으로 평가하는 경우에 대한 제재 장치**를 사전에 명시함

<평가결과 예시>

		평 가 자										평균
		A	B	C	D	E	F	G	H	I	J	
피평가자	가	90	54	84	90	80	92	86	100	80	78	83.40
	나	88	54	90	88	82	90	88	96	78	76	83.00
	다	68	64	78	74	78	72	70	94	72	74	74.40
	라	72	56	82	70	76	74	70	94	60	74	72.80
	마	66	54	78	70	76	72	68	92	70	74	72.00
평균		76.80	56.40	82.40	78.40	78.40	80.00	76.40	95.20	72.00	75.20	77.12
표준편차		11.37	4.34	4.98	9.84	2.61	10.10	9.74	3.03	7.87	1.79	

※ 1. ○○대학 ☆☆학과에서 실시한 면접고사 결과라고 가정

2. 평가자 B는 엄격(평균 56.4)하게 평가한 반면 평가자 H는 관대(평균 95.2)하게 평가

3. 평가자 A는 평균값에 근접하게 평가(평균 76.8)하였으나 점수편차가 가장 큼(표준편차 11.37)

2) 평가점수 조정 방법

가) 평가자별 평균·표준편차 일치방법

(1) 계산방법 : (예) 평균 '70점', 표준편차 '10'으로 조정하는 경우

$$70 + [(\text{원점수} - \text{평가자 평균점수})/\text{평가자 표준편차}] \times 10$$

(2) 특 징

- 평가자간 평균과 표준편차를 표준정규분포를 이용하여 일정하게 조정하는 방법
- 평균과 표준편차가 매번 동일하므로 장기적으로 비교·활용 가능

<조정 예시 가>

		평 가 자										조정 평균
		A	B	C	D	E	F	G	H	I	J	
피평가자	가	81.61	64.46	73.21	81.79	76.14	81.88	79.86	85.83	80.16	85.65	79.06
	나	79.85	64.46	85.26	79.76	83.81	79.90	81.91	72.64	77.62	74.47	77.97
	다	62.26	87.53	61.16	65.53	68.47	62.08	63.43	66.04	70.00	63.29	66.98
	라	65.78	69.08	69.20	61.46	60.80	64.06	63.43	66.04	54.76	63.29	63.79
	마	60.50	64.46	61.16	61.46	60.80	62.08	61.37	59.45	67.46	63.29	62.20
평균		70.00	70.00	70.00	70.00	70.00	70.00	70.00	70.00	70.00	70.00	70.00
표준편차		10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	

※ 1. 21쪽의 평가결과 예시를 기준으로 조정한 결과임
2. 평가자 'A~J'의 점수가 평균 70점, 표준편차 10으로 조정
3. A의 '가'에 대한 조정 평균 81.61=70+[(90-76.80)/11.37)]×10

나) 평가자별 평균 일치 방법

(1) 계산방법 : 평가점수 × (피평가자그룹 전체평균/ 평가자 평균)

(2) 특 징

- 평가자의 평균점수를 피평가자 전체 평균점수와 동일하게 조정
- 조정 후의 표준편차는 원래의 표준편차와 일치하지 않음

<조정 예시 나>

		평 가 자										조정 평균
		A	B	C	D	E	F	G	H	I	J	
피평가자	가	90.38	73.84	78.62	88.53	78.69	88.69	86.81	81.01	85.69	79.99	83.22
	나	88.37	73.84	84.23	86.56	80.66	86.76	88.83	77.77	83.55	77.94	82.85
	다	68.28	87.51	73.00	72.79	76.73	69.41	70.66	76.15	77.12	75.89	74.75
	라	72.30	76.57	76.75	68.86	74.76	71.34	70.66	76.15	64.27	75.89	72.75
	마	66.28	73.84	73.00	68.86	74.76	69.41	68.64	74.53	74.98	75.89	72.02
평균		77.12	77.12	77.12	77.12	77.12	77.12	77.12	77.12	77.12	77.12	77.12
표준편차		11.41	5.93	4.66	9.68	2.57	9.74	9.83	2.46	8.43	1.83	

※ 1. 평가자 'A~J'의 평균점수가 그룹 전체평균점수인 77.12점으로 조정
2. A의 '가'에 대한 조정 평균 90.38=90×(77.12/76.80)

다) 특이점수를 제외하는 방법

(1) 계산방법 : 피평가자가 받은 점수 중 최상·최하점수를 제외한 후 점수를 기준으로 계산

<조정 예시 다>

		평 가 자										조정 평균
		A	B	C	D	E	F	G	H	I	J	
피평가자	가	90	**54**	84	90	80	92	86	**100**	80	78	85.00
	나	88	**54**	90	88	82	90	88	**96**	78	76	85.00
	다	68	**64**	78	74	78	72	70	**94**	72	74	73.25
	라	72	**56**	82	70	76	74	70	**94**	60	74	72.25
	마	66	**54**	78	70	76	72	68	**92**	70	74	71.75

※ 1. 피평가자 '가~마'가 받은 점수 중 최상·최하 점수를 제외
2. 피평가자 '가'의 경우 받은 점수 "90, 54, 84, ~ 100, 80, 90" 중 최저점 54, 최고점 100을 제외하고 조정 평균점 산출

7. 결과자료 통계분석과 활용

가. 학교생활기록부 내신성적 산출(수시 예시)

(1) 석차백분위 또는 표준점수(Z)를 반영하는 경우

(예시) 홍길동이 A전형으로 사회과학계열에 지원자할 경우
⇨ 학생부 평가척도 : 원점수, 과목평균, 표준편차를 이용한 표준점수
⇨ 학생부 반영교과 : 국어/영어/수학/사회 (3학년 1학기까지 이수한 전과목)
⇨ 학년별 반영비율 : 1학년 20%, 2학년 40%, 3학년 40%

□ 홍길동의 교과학습발당사항(예시)

학년	교과	과목	1 학 기			2 학 기			비고
			단위수	원점수/과목평균 (표준편차)	석차등급/ (이수자수)	단위수	원점수/과목평균 (표준편차)	석차등급/ (이수자수)	
1	국어	국어	4	89/78.1(10.2)	3(528)	4	78/70.7(13.7)	4(530)	
	사회	국사	2	97/81.3(14.1)	2(528)	2	88/71.8(16)	3(530)	
2	수학	수학 I	4	78/63.6(17.6)	4(354)	4	78/57.7(21.1)	3(356)	
	외국어	영어 I	5	79/64.9(16.5)	3(354)	5	74/62.1(18.9)	4(356)	
3	사회	정치	4	96/75.1(14.6)	1(354)				
	수학	수학 I	4	85/62.1(18.9)	3(354)				

□ 계산과정

○ 기호 정의: 학년 학기별 과목의 원점수 X, 평균 μ 그리고 표준편차 σ

(1단계) 표준화점수 구하기 : $Z = \dfrac{X - \mu}{\sigma}$

(2단계) 표준화점수에 대한 누적확률 구하기 : $P(Z) = \int_{-\infty}^{Z} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$

(표준정규분포 이용)

(3단계) 과목별 상대적 위치 찾기 : $f(Z) = (1 - P(Z)) \times 100$ => **표준점수**로 정의

(4단계) 학년별 표준점수 산출 : {∑(과목별 표준점수)×(이수단위)}/{∑(이수단위)}

(5단계) 해당대학 A전형 학생부 평가기준표에 따라 학년별 환산점수 산출

(6단계) 학년별 반영 비율에 따라 가중평균한 최종 학생부성적 산출

학년	학기	교과	과목	1단계 Z	2단계 $P(Z)$	3단계 (A)	단위수 (B)	(C) = (A)×(B)	4단계 (학년별)	5단계 (환산점수)	6단계 (학생부성적)
1	1	국어	국어	1.07	0.8574	14.26	4	57.04	233.56/12 = 19.46	794	794×0.2 + 792×0.4 + 797×0.4 = 794.40
		사회	국사	1.11	0.8672	13.28	2	26.56			
	2	국어	국어	0.53	0.7029	29.71	4	118.84			
		사회	국사	1.01	0.8444	15.56	2	31.12			
	합계						**12**	**233.56**			
2	1	수학	수학 I	0.82	0.7934	20.66	4	82.64	380.29/18 = 21.13	792	
		외국어	영어 I	0.85	0.8036	19.64	5	98.20			
	2	수학	수학 I	0.96	0.8320	16.80	4	67.20			
		외국어	영어 I	0.63	0.7355	26.45	5	132.25			
	합계						**18**	**380.29**			
3	1	사회	정치	1.43	0.9239	7.61	2	15.22	91.74/6 = 15.29	797	
		수학	수학 I	0.87	0.8087	19.13	4	76.52			
	합계						**6**	**91.74**			

(2) 석차등급을 반영하는 경우

(예시) 홍길동이 B전형으로 공학계열에 지원자할 경우

⇨ 학생부 평가척도 : 석차등급

⇨ 학생부 반영교과 : 국어/영어/수학/과학 (3학년 1학기까지 이수한 전과목)

⇨ 학년별 반영비율 : 1학년 20%, 2학년 40%, 3학년 40%

□ 계산과정

(1단계) 과목별 이수자수가 12명 이하인 경우, 대학에서 제공하는 등급별 조정점수표에 의해 과목석차등급 보정

(2단계) 학년별 석차등급평균 산출 : {∑(과목별 석차등급)×(이수단위)}/{∑(이수단위)}

(3단계) 해당대학 B전형 학생부 평가기준표에 따라 학년별 환산점수 산출

(4단계) 학년별 반영 비율에 따라 가중평균한 최종 학생부성적 산출

※ 조기졸업자인 경우 2학년 1학기까지 성적 반영(학년별 반영비율 : 1학년 40%, 2학년 60%)

※ 석차등급을 적용하는 전형은 과목별 이수자수를 고려하여 등급 조정

(과목별 조정등급) = (석차등급) - (조정점수)

[표] 이수자 수에 따른 등급별 조정점수

등급 / 이수자 수	1	2	3	4	5	6	7	8	9
1명					4				
2명				3		3			
3 ~ 4명			2	2	2	2	2		
5 ~ 12명		1	1	1	1	1	1	1	

(3) 학생부 평가표 예시

□ 표준점수 반영 시 학생부 평가 기준표

표준점수 평균	등급	환산점수	급간점수차	표준점수 평균	등급	환산점수	급간점수차
7.00 이하	1	800.0	-	30.00 이하	9	787.0	3
9.00 이하	2	799.0	1	34.00 이하	10	784.0	3
11.00 이하	3	798.0	1	41.00 이하	11	779.0	5
13.00 이하	4	797.0	1	52.00 이하	12	774.0	5
16.00 이하	5	796.0	1	68.00 이하	13	767.0	7
19.00 이하	6	794.0	2	88.00 이하	14	760.0	7
22.00 이하	7	792.0	2	88.00 초과	15	750.0	10
26.00 이하	8	790.0	2				

□ 석차등급 반영 시 학생부 평가 기준표

석차등급 평균	등급	환산점수	급간점수차	석차등급 평균	등급	환산점수	급간점수차
2.00 이하	1	400	-	6.50 이하	6	393	2
2.50 이하	2	399	1	7.50 이하	7	391	2
3.50 이하	3	398	1	8.50 이하	8	388	3
4.30 이하	4	397	1	8.50 초과	9	380	8
5.00 이하	5	395	2				

나. 학교생활기록부 내신성적 산출(정시 예시)

(1) 학교생활기록부 반영요소 및 반영방법

가) 학교생활기록부 반영방법

계열	학생부 반영총점	교과				비교과	
		반영점수	반영교과	반영지표	학년별 반영비율	반영점수	반영지표
인문계	500	470	국어, 영어, 수학, 사회	석차등급	1학년20% 2학년30% 3학년50%	30	출석, 봉사시간
자연계			국어, 영어, 수학, 과학				
예체능계	400	370	국어, 영어, 수학, 사회				

나) 교과성적 반영방법

1) 학생이 이수한 계열별 반영교과 전 과목에 대하여, 학년별 최상위 석차등급에 해당하는 4개 과목(총 12과목)만을 선택하여 반영

2) 교과성적 산출방법

(1단계) 학년별 과목에 상관없이 최상위 석차등급 4개 과목 선택

(2단계) 학년별 4개 과목 석차등급평균 산출

(3단계) 해당대학 자체기준에 따라 학년별 석차등급평균에 해당하는 등급점수 부여

(4단계) 학년별 등급점수를 반영비율에 따라 가중평균한 최종 학생부 교과성적 산출

다) 비교과 평가기준 및 반영방법

① 비교과 평가기준표

점수	15	14.5	14	13	12	10	8	4	0
출석	5일 이하	6~8일	9~11일	12~14일	15~17일	18~20일	21~23일	24~26일	27일이상
봉사시간	20시간이상	18~19시간	16~17시간	14~15시간	12~13시간	10~11시간	8~9시간	6~7시간	5시간이하

② 반영방법 : 출석과 봉사시간에 대하여 각각 15점 만점으로 점수를 구한 후 합산하여 산출

(2) 학생부 누락자 및 수능에 의한 비교내신자 교과성적 처리기준

가) 학생부 누락자 교과성적 처리기준

① 2008년 2월 졸업예정자 또는 조기졸업(예정)자 (교과목별 석차등급이 표기되어 있지 않은 자 대상)

② 1개 학년의 성적 중 1개 학기의 성적이 없는 경우 나머지 1개 학기의 성적을 해당 학년의 성적으로 처리

③『1개 학년의 성적이 모두 누락한 경우』와『1개 학년(또는 1개 학기)의 성적만 있는 경우』에는 다음의 반영기준에 따라 처리한다.

구분	누락여부			반영기준		
	1학년	2학년	3학년	1학년	2학년	3학년
1개 학년 성적 누락	○	○	×	40%	60%	-
	○	×	○	40%	-	60%
	×	○	○	-	40%	60%
1개 학(기)년 성적만 있을 경우	○	×	×	100%	-	-
	×	○	×	-	100%	-
	×	×	○	-	-	100%

나) 수능에 의한 비교내신 반영 대상

① 2007년 2월 이전 졸업자

② 국내 고등학교 졸업(예정)자 중 학생부에 교과목 석차등급이 1개 학기도 기록되어 있지 않은 자

③ 학생부 성적이 없는 자(해외 고교 출신자, 검정고시 출신자 등)

④ 공업계 2+1, 일반고교 직업과정위탁생 졸업(예정)자

⑤ 교과교육 소년원 고교과정 이수자

▶ 비교내신 대상자의 학생부 교과 및 비교과 성적은 본교 지원자의 수능성적에 의해 성적 산출

다. 수학능력시험 활용

(1) 영역별, 등급별 빈도 분석(예시)

[표 7-1] 언/수/외 등급 조합 분포표

언어	외국어	수리 '가'형 (자연계)				수리 '나'형 (인문계)				합계
		1	2	3	소계	1	2	3	소계	
1	1	1,775	1,734	1,371	4,880	2,911	1,407	625	4,943	9,823
	2	535	893	1048	2,476	1,435	1,549	1,074	4,058	6,534
	3	112	342	556	1,010	641	1,067	1,121	2,829	3,839
소계		2,422	2,969	2,975	8,366	4,987	4,023	2,820	11,830	20,196
2	1	843	1,054	961	2,858	1,374	989	575	2,938	5,796
	2	435	953	1,277	2,665	1,373	1,745	1,391	4,509	7,174
	3	174	462	799	1,435	810	1,731	1,936	4,477	5,912
소계		1,452	2,469	3,037	6,958	3,557	4,465	3,902	11,924	18,882
3	1	410	610	671	1,691	584	655	425	1,664	3,355
	2	337	795	1,225	2,357	949	1,565	1,509	4,023	6,380
	3	174	606	1,236	2,016	916	2,262	2,759	5,937	7,953
소계		921	2,011	3,132	6,064	2,449	4,482	4,693	11,624	17,688
합계		4,795	7,449	9,144	21,388	10,993	12,970	11,415	35,378	56,766

[표 7-2] 언/수/외 등급 조합 분포 비율표

언어	외국어	수리 '가'형 (자연계)				수리 '나'형 (인문계)				합계
		1	2	3	소계	1	2	3	소계	
1	1	1.42%	1.39%	1.10%	3.90%	0.86%	0.42%	0.18%	1.46%	1.79%
	2	0.43%	0.71%	0.84%	1.98%	0.42%	0.46%	0.32%	1.20%	1.19%
	3	0.09%	0.27%	0.44%	0.81%	0.19%	0.31%	0.33%	0.83%	0.70%
소계		1.94%	2.37%	2.38%	6.69%	1.47%	1.19%	0.83%	3.49%	3.66%
2	1	0.67%	0.84%	0.77%	2.29%	0.41%	0.29%	0.17%	0.87%	1.05%
	2	0.35%	0.76%	1.02%	2.13%	0.41%	0.51%	0.41%	1.33%	1.31%
	3	0.14%	0.37%	0.64%	1.15%	0.24%	0.51%	0.57%	1.32%	1.08%
소계		1.16%	1.97%	2.43%	5.56%	1.05%	1.32%	1.15%	3.52%	3.42%
3	1	0.33%	0.49%	0.54%	1.35%	0.17%	0.19%	0.13%	0.49%	0.61%
	2	0.27%	0.64%	0.98%	1.88%	0.28%	0.46%	0.45%	1.19%	1.16%
	3	0.14%	0.48%	0.99%	1.61%	0.27%	0.67%	0.81%	1.75%	1.45%
소계		0.74%	1.61%	2.50%	4.85%	0.72%	1.32%	1.38%	3.43%	3.21%
합계		3.83%	5.96%	7.31%	17.10%	3.24%	3.83%	3.37%	10.44%	10.24%

[표 7-3] 언/수/외 3개영역 등급 조합 분포 및 비율

등급조합(예)	언수외탐	사탐		과탐	
		누적도수	누적비율	누적도수	누적비율
(1, 1, 1)	3	928	0.53%	2,031	0.47%
(1, 1, 2)	4	3,290	1.88%	6,913	1.6%
(1, 2, 2)	5	6,650	3.8%	13,394	3.1%
(2, 2, 2)	6	11,200	6.4%	21,603	5.0%
(2, 2, 3)	7	16,800	9.6%	31,972	7.4%
(2, 3, 3)	8	22,750	13.0%	43,205	10.0%
(3, 3, 3)	9	31,501	18.0%	77,769	18.0%
(3, 3, 4)	10	40,251	23.0%	95,051	22.0%
(3, 4, 4)	11	50,751	29.0%	116,654	27.0%
(4, 4, 4)	12	61,251	35.0%	142,577	33.0%
(4, 4, 5)	13	73,501	42.0%	168,500	39.0%
(4, 5, 5)	14	87,502	50.0%	194,423	45.0%
(5, 5, 5)	15	99,752	57.0%	224,667	52.0%

[표 7-4] 언/수/외/탐 4개영역 등급 조합 분포 및 비율

등급조합(예)	언수외탐	사탐		과탐	
		누적도수	누적비율	누적도수	누적비율
(1, 1, 1, 1)	4	765	0.23%	484	0.24%
(1, 1, 1, 2)	5	3,458	1.0%	2,298	1.14%
(1, 1, 2, 2)	6	7,314	2.2%	4,838	2.4%
(1, 2, 2, 2)	7	11,637	3.5%	8,063	4.0%
(2, 2, 2, 2)	8	17,289	5.2%	11,893	5.9%
(2, 2, 2, 3)	9	23,606	7.1%	16,730	8.3%
(2, 2, 3, 3)	10	31,252	9.4%	22,173	11.0%
(2, 3, 3, 3)	11	39,897	12.0%	28,220	14.0%
(3, 3, 3, 3)	12	49,871	15.0%	36,283	18.0%
(3, 3, 3, 4)	13	63,170	19.0%	44,346	22.0%
(3, 3, 4, 4)	14	73,144	22.0%	52,409	26.0%
(3, 4, 4, 4)	15	86,443	26.0%	62,487	31.0%
(4, 4, 4, 4)	16	103,067	31.0%	72,566	36.0%

[표 7-5] 언/수/외/탐 4개영역 1등급 분포 및 비율

언수외	탐구 1등급 수	탐구등급 조합(예)	사탐		과탐	
			누적도수	누적비율	누적도수	누적비율
(1, 1, 1)	4	(1, 1, 1, 1)	466	0.15%	369	0.20%
	3	(1, 1, 1, 2)	880	0.28%	529	0.28%
	2	(1, 1, 2, 2)	1,187	0.38%	491	0.26%
	1	(1, 2, 2, 2)	1,048	0.34%	419	0.22%
합 계			3,581	1.16%	1,808	0.97%

[표 7-6] 고교별 수능영역별 등급 비율

등급	1	2	3	4	5	6	7	8	9
목표비율	4	7	12	17	20	17	12	7	4
언어	4.27	7.67	11.6	18.35	19.2	16.73	11.57	7.32	3.28
수리 '가'	4.69	7.17	11.89	16.44	20.82	17.32	13.13	4.72	3.99
외국어	5.41	5.73	12.49	16.74	20.76	17.07	11.31	7.39	3.1

(2) 대학별, 수능 영역 반영 비율 적용(예시)

- 수능 영역별 반영 비율에 따라 학생별 환산점수 차이 발생
- 대학별 계산식에 의해 유불리 확인

[표 7-7] 인문계열 대학별, 모집단위별 수능 반영 비율

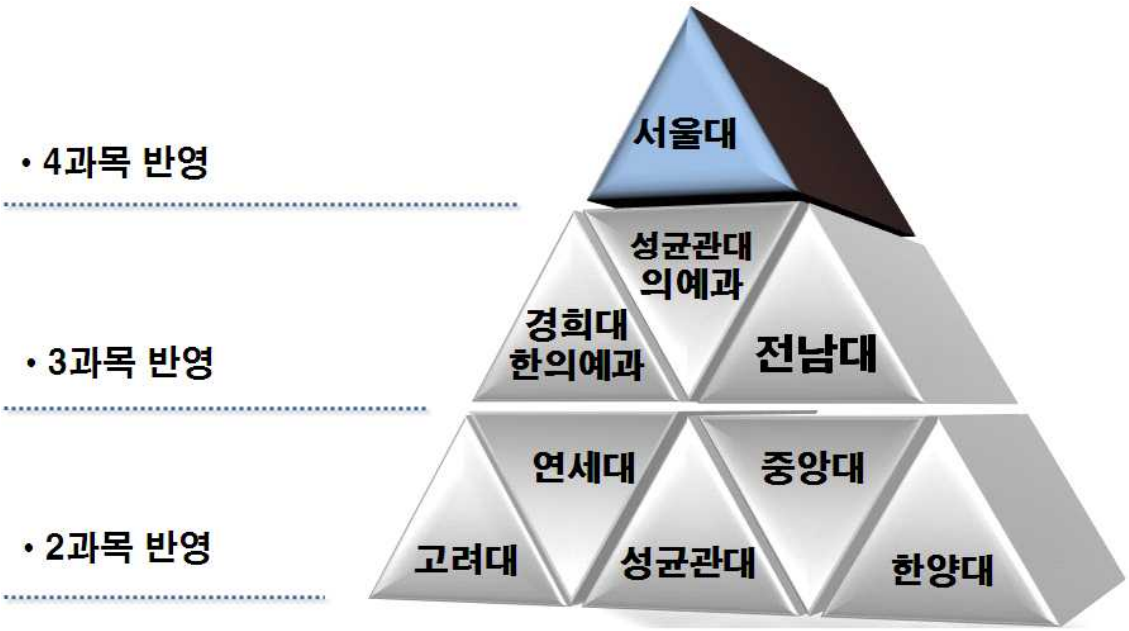
대학명	모집군	모집단위	활용 지표	수능 영역별 반영비율(100%)							
				언어	수리			외국어	탐구		
					가	나	가/나		사회	과학	사/과
A	나	경영/경제	표	22.5		30		30	17.5		
		전체	표	25		27.5		30	17.5		
B	나	인문사회1	표+백	22			28	22			22
		인문사회2	표+백	22			27.8	22			22.2
C	가,나	전체	표+백	25			30	30	15		
D	가,나	전체	표	20			30	30			20

[표 7-8] 인문계열 대학별, 모집단위별 수능 반영 비율

대학명	모집군	모집단위	활용 지표	수능 영역별 반영비율(100%)							
				언어	수리			외국어	탐구		
					가	나	가/나		사회	과학	사/과
A	나	자연과학부	표	17.5	30			27.5		25	
B	나	전체(간호, 사범제외)	표+백	23.5	29.5			23.5		23.5	
		사범	표+백	29			29	29			14
		간호대학	표+백	23.5			29.5	23.5			23.5
C	가,나	전체	표+백	25	30			30		15	
D	가,나	전체	표	20	30			20		30	

(3) 탐구 과목수에 따른 점수 변동(예시)

- 탐구 영역 반영 과목수의 축소에 따라 동점자 수 증가
- 상위권 대학 및 학과 진학 시 탐구에서의 변별력 약화, 언수외 영향력 증가
- 탐구 과목 중 하나로 제2외국어 포함 여부에 따라 총점 순위 변동 심화



[그림 7-1] 주요대학 탐구 반영 과목 수(2011학년도)

[표 7-9] 탐구 과목수에 따른 동점자 수 증가(인문)

표준점수(800)	탐구2개 빈도	탐구 3개 빈도	빈도 차이
560	36	17	**19**
559	66	17	**49**
558	86	23	**63**
557	97	38	**59**
556	125	50	**75**
555	140	91	**49**
554	156	92	**64**
553	196	139	**57**
552	222	136	**86**
551	243	174	**69**

[표 7-10] 탐구 과목 축수에 따른 점수변동(인문)

탐구2개 반영	빈도	3개반영 대비 점수변동	탐구2개 반영	빈도	3개반영 대비 점수변동
567	8	5.67	557	97	9.67
566	13	9.33	556	125	10.67
565	11	7.33	555	140	9.00
564	22	9.67	554	156	7.33
563	22	8.00	553	196	10.67
562	25	7.33	552	222	9.33
561	40	7.33	551	243	12.67
560	36	11.00	550	274	12.00
559	66	9.33	549	302	11.67
558	86	8.00	548	359	17.00