

# Introductory Statistics

Sungchan Yi

January 2019

## Contents

1	자료의 생성	2
2	대표값과 산포도	5
3	순열과 조합	8
4	확률의 뜻과 활용	11
5	조건부확률	14
6	이산확률변수	19
7	이산확률분포	24
8	연속확률변수	28
9	표본분포	33
10	통계적 추론	40

## 1 자료의 생성

**통계학**(statistics)이란, 주어진 문제에 대하여 합리적인 답을 줄 수 있도록 숫자로 표시되는 정보를 수집하고 정리하며, 이를 해석하고 **신뢰성 있는 결론**을 이끌어 내는 방법을 연구하는 과학의 한 분야이다.

그러면 두 가지 질문이 생긴다.

1. **수집**: 어떻게 수집해야 전체를 잘 대표할 수 있는가?
2. **신뢰성 있는 결론**: 어떻게 신뢰성을 측정하여 결론을 내릴 것인가?

### 정의 1.1.

- **추출단위**(sampling unit): 전체를 구성하는 각 개체
- **특성값**(characteristic): 각 추출단위의 특성을 나타내는 값
- **모집단**(population): 관심의 대상이 되는 모든 추출단위의 특성값을 모아 놓은 것  
추출단위의 개수가 유한하면 **유한모집단**, 무한하면 **무한모집단**이라 한다.
- **표본**(sample): 실제로 관측한 추출단위의 특성값의 모임

### 정의 1.2. (자료의 종류)

1. **범주형 자료**(categorical data), **질적 자료**(qualitative data)는 관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료이다.
  - **명목자료**(nominal data): 순위의 개념이 없다. 예) 혈액형, 성별
  - **순서자료**(ordinal data): 순위의 개념을 갖는다. 예) A ~ F 학점, 9등급제
2. **수치형 자료**(numerical data), **양적 자료**(quantitative data)는 자료 자체가 숫자로 표현되며 숫자 자체가 자료의 속성을 반영한다.
  - **이산형 자료**(discrete data) 예) 교통사고 건수
  - **연속형 자료**(continuous data) 예) 키, 몸무게

### 정의 1.3. (통계학의 분류)

- **기술통계학**(descriptive statistics)은 표나 그림 또는 대표값 등을 통하여 수집된 자료의 특성을 쉽게 파악할 수 있도록 자료를 정리·요약 하는 방법을 다루는 분야이다.
- **추측통계학**(inferential statistics)은 표본에 내포된 정보를 분석하여 모집단의 여러가지 특성에 대하여 과학적으로 추론하는 방법을 다루는 분야이다.

**정의 1.4.**  $N$ 개의 추출단위로 구성된 유한모집단에서  $n$ 개의 추출단위를 비복원추출할 때,  ${}_NC_n$ 개의 모든 가능한 표본들이 동일한 확률로 추출되는 방법을 **단순랜덤추출법**(simple random sampling)이라 하고, 이 방법을 위해서는 난수표(random number table)나 난수생성기(random number generator) 등을 이용한다. 그리고 단순랜덤추출로 얻은 표본을 **단순랜덤표본**(simple random sample)이라 한다.

### 정의 1.5. (통계적 실험)

- 실험이 행해지는 개체를 **실험단위**(experimental unit/subject)라 하고, 각각의 실험단위에 특정한 실험환경 또는 실험조건을 가하는 것을 **처리**(treatment)라 한다.
- 처리를 받는 집단을 **처리집단**(treatment group), 처리를 받지 않은 집단을 **대조집단**(control group)이라 한다.
- 실험환경이나 실험조건을 나타내는 변수를 **인자**(factor)라 하고, 인자가 취하는 값을 그 인자의 **수준**(level)이라 한다.
- 인자에 대한 반응을 나타내는 변수를 **반응변수**(response variable)라 한다.
- 실험단위가 처리집단이나 대조집단에 들어갈 기회를 동등하게 부여하는 방법을 **랜덤화**(randomization)라 한다.
- 랜덤화에 의해 모든 실험단위를 각 처리에 배정하는 실험계획을 **완전 랜덤화 계획**(completely randomized design)이라 한다.
- 실험 이전에 동일 처리에 대한 반응이 유사할 것으로 예상되는 실험단위들끼리 모은 것을 **블록**(block)이라 하고, 랜덤화에 의해 모든 블록을 각 처리에 배정하는 실험계획을 **블록화**(randomized block design)라 한다.

### 정의 1.6. (통계적 실험계획의 원칙)

1. **대조(control)**: 관심 인자 이외의 다른 외부 인자의 효과를 극소화하고, 처리에 대한 대조집단을 통해 비교 실험을 한다.
2. **랜덤화(randomization)**: 완전랜덤화계획
3. **반복 시행(replication)**: 처리효과의 탐지를 용이하게 하기 위해 반복 시행한다.

## 2 대표값과 산포도

주어진 자료의 변량  $x_1, \dots, x_n$ 에 대하여

**정의 2.1.** (산술)평균(mean  $\mu$ )은 변량의 총합을 변량의 개수로 나눈 값이다.

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

**정의 2.2.** 변량  $x_i$ 의 편차는  $x_i - \mu$ 로 정의한다. [편차의 제곱]의 평균을 분산(variance  $\sigma^2$ )으로, 분산의 양의 제곱근을 표준편차(standard deviation  $\sigma$ )로 정의한다.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

**정리 2.3.**  $\sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \mu^2$  (분산은 [변량<sup>2</sup>의 평균] - [평균의 제곱])

**정의 2.4.** 모집단의 평균을 모평균(population mean), 모집단의 분산을 모분산(population variance), 모집단의 표준편차를 모표준편차(population standard deviation)라고 한다.

**정의 2.5.** 특성값을 작은 것부터 순서대로 나열했을 때  $p\%$ 의 특성값이 그 값보다 작거나 같고,  $(100-p)\%$ 의 특성값이 그 값보다 크거나 같게 되는 값을 제  $p$  백분위수( $p$ -th percentile)라 한다.

**정의 2.6.** (사분위수 - quartile)

- 제 1 사분위수(first quartile): 제 25 백분위수이며,  $Q_1$ 으로 표기한다.
- 제 2 사분위수(second quartile) 또는 중앙값(median): 제 50 백분위수이며,  $Q_2$ 으로 표기한다.
- 제 3 사분위수(third quartile): 제 75 백분위수이며,  $Q_3$ 으로 표기한다.

**정의 2.7.** 자료의 값들 중 가장 자주 등장하는 값을 **최빈값(mode)**라고 한다. 최빈값은 유일하지 않을 수도 있다.

**정의 2.8.** 변량과 중앙값 사이의 거리에 대한 평균을 **평균절대편차(mean absolute deviance MAD)**라 한다.

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - Q_2|$$

**정의 2.9.** 변량  $x_i$ 가 **최댓값(maximum)**이면 모든  $j$ 에 대해  $x_i \geq x_j$  이고,  $x_i$ 가 **최솟값(minimum)**이면 모든  $j$ 에 대해  $x_i \leq x_j$  이다. 최댓값에서 최솟값을 뺀 값을 **범위(range  $R$ )**라 한다.

**정의 2.10.**  $Q_3$ 에서  $Q_1$ 을 뺀 값을 **사분위수범위(interquartile range  $IQR$ )**로 정의한다.

$$IQR = Q_3 - Q_1$$

위 대표값들을 표본에 대해서도 정의할 수 있다. 모집단으로부터 표본  $x_1, \dots, x_n$ 를 얻었다고 하고, 이를 오름차순으로 나열한 것을  $x_{(1)}, \dots, x_{(n)}$ 이라 하자.

**정의 2.11.**

- **표본평균(sample mean):**  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- **표본분산(sample variance):**  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- **표본표준편차(sample standard deviation):**  $s = \sqrt{s^2}$

**정의 2.12.** 표본을 크기 순서로 나열했을 때  $p\%$ 가 그 값보다 작고,  $(100-p)\%$ 가 그 값보다 크게 되는 값을 **표본의 제  $p$  백분위수**라 하고, 다음과 같이 계산한다.

$$\begin{cases} \frac{x_{(k)} + x_{(k+1)}}{2} & \text{if } n \cdot \frac{p}{100} = k \\ x_{(k+1)} & \text{if } k < n \cdot \frac{p}{100} < k+1 \end{cases}$$

정의 2.13. 표본의 제  $i$  사분위수  $\widehat{Q}_i$  는 표본의 제  $25i$  백분위수로 정의한다. (단,  $i = 1, 2, 3$ )

정의 2.14.

- 표본의 평균절대편차:  $\widehat{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \widehat{Q}_2|$
- 표본의 범위:  $\widehat{R} = x_{(n)} - x_{(1)}$
- 표본의 사분위수범위:  $\widehat{IQR} = \widehat{Q}_3 - \widehat{Q}_1$

### 3 순열과 조합

**정의 3.1.**  $0! = 1$ ,  $n! = \prod_{i=1}^n i = n \cdot (n-1) \cdots \cdots 2 \cdot 1$  ( $n \geq 1$ ) 로 정의하고, ! 는 **팩토리얼**(factorial) 이라 읽는다.

**정의 3.2.** 서로 다른  $n$ 개의 원소에서 서로 다른  $r$ 개를 택하여 일렬로 배열하는 것을  $n$ 개에서  $r$ 개를 택하는 **순열**(permutation)이라 하고, 기호로  ${}_nP_r$  와 같이 나타낸다.

**정리 3.3.**  ${}_nP_r = n(n-1) \cdots (n-r+1) = \frac{n!}{(n-r)!}$  (단,  $0 \leq r \leq n$ )

**정의 3.4.** 서로 다른  $n$ 개의 원소에서 순서를 생각하지 않고  $r$ 개를 택하는 것을  $n$ 개에서  $r$ 개를 택하는 **조합**(combination)이라 하고, 기호로  ${}_nC_r$  또는  $\binom{n}{r}$  과 같이 나타낸다.

**정리 3.5.**  $\binom{n}{r} = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!}$  (단,  $0 \leq r \leq n$ )

**정리 3.6.** (조합의 성질)

$$(1) \quad \binom{n}{r} = \binom{n}{n-r} \quad (\text{단, } 0 \leq r \leq n) \quad (\text{대칭성})$$

$$(2) \quad \binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1} \quad (\text{단, } 1 \leq r \leq n-1) \quad (\text{파스칼 법칙})$$

**정의 3.7.** 서로 다른  $n$ 개의 원소에서 중복을 허락하여  $r$ 개를 택하는 순열을  $n$ 개에서  $r$ 개를 택하는 **중복순열**이라 하고, 기호로  ${}_n\Pi_r$  과 같이 나타낸다.

**정의 3.8.** 서로 다른  $n$ 개의 원소에서 중복을 허락하여  $r$ 개를 택하는 조합을  $n$ 개에서  $r$ 개를 택하는 **중복조합**이라 하고, 기호로  ${}_nH_r$  과 같이 나타낸다.



**정리 3.9.**  ${}_n\Pi_r = n^r$ ,  ${}_nH_r = \binom{n+r-1}{r}$ .

**정리 3.10.**  $n \in \mathbb{N}$  에 대하여,

$$(x+y)^n = \sum_{r=0}^n \binom{n}{r} x^{n-r} y^r = \binom{n}{0} x^n y^0 + \binom{n}{1} x^{n-1} y^1 + \cdots + \binom{n}{n} x^0 y^n$$

이다. 이를  $(a+b)^n$ 에 대한 **이항정리**(binomial theorem)라 하고,  $\binom{n}{r} x^{n-r} y^r$ 을 전개식의 **일반항**, 전개식의 각 항의 계수  $\binom{n}{r}$ 들을 **이항계수**라 한다.

**정리 3.11.** (이항계수의 성질)

$$(1) \quad (1+x)^n = \sum_{r=0}^n \binom{n}{r} x^r = \binom{n}{0} + \binom{n}{1} x + \cdots + \binom{n}{n} x^n \quad (\text{for all } x \in \mathbb{C})$$

$$(2) \quad \sum_{r=0}^n \binom{n}{r} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$$

$$(3) \quad \sum_{r=0}^n (-1)^r \binom{n}{r} = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n \binom{n}{n} = 0$$

$$(4) \quad \sum_{r=0}^n r \binom{n}{r} = \binom{n}{1} + 2 \cdot \binom{n}{2} + \cdots + n \cdot \binom{n}{n} = n \cdot 2^{n-1}$$

$$(5) \quad \sum_{r=0}^n r^2 \binom{n}{r} = 2^2 \cdot \binom{n}{2} + 3^2 \cdot \binom{n}{3} + \cdots + n^2 \cdot \binom{n}{n} = n(n+1) \cdot 2^{n-2}$$

$$(6) \quad \sum_{r=0}^n \frac{1}{r+1} \binom{n}{r} = \frac{1}{1} \binom{n}{0} + \frac{1}{2} \binom{n}{1} + \cdots + \frac{1}{n+1} \binom{n}{n} = \frac{1}{n+1} (2^{n+1} - 1)$$

**정리 3.12.**  $n \in \mathbb{N}$  에 대하여,

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{r_1+r_2+\cdots+r_m=n} \binom{n}{r_1, r_2, \dots, r_m} x_1^{r_1} x_2^{r_2} \cdots x_m^{r_m}$$

이고, 이를  $(x_1 + x_2 + \cdots + x_m)^n$ 에 대한 **다항정리**(multinomial theorem)라 한다. 이 때  $\binom{n}{r_1, r_2, \dots, r_m}$ 를 **다항계수**라 하고, 다음과 같이 정의한다.

$$\binom{n}{r_1, r_2, \dots, r_m} = \frac{n!}{r_1! \cdot r_2! \cdot \cdots \cdot r_m!}$$

**정의 3.13.** 서로 다른  $n$ 개의 원소를 원형으로 배열하는 순열을 **원순열**이라 하고, 그 경우의 수는  $(n-1)!$  이다.

**정리 3.14.** (원순열의 일반공식)  $n$ 개 중에서 서로 같은 것이  $p_1, p_2, \dots, p_k$ 개씩 있을 때, 이  $n (= p_1 + \dots + p_k)$ 개를 원형으로 배열하는 방법(원순열)의 수는 다음과 같다.

$$\frac{1}{n} \sum_{d|g} \left\{ \phi(d) \binom{\frac{n}{d}}{\frac{p_1}{n}, \frac{p_2}{n}, \dots, \frac{p_k}{n}} \right\}$$

단,  $g = \gcd(p_1, \dots, p_k)$ ,  $d > 0$  이고  $\phi(d)$ 는  $d$  이하의 자연수 중에서  $d$  와 서로소인 자연수의 개수로 정의된다.

## 4 확률의 뜻과 활용

확률은 모집단에서 표본을 추출할 때, 특정 성질을 만족하는 표본이 관측될 가능성에 대한 측도로, 표본을 바탕으로 **모집단에 대한 결론을 이끌어낼 때 논리적 근거**가 된다.

**정의 4.1.** 같은 조건 아래에서 반복할 수 있고, 그 결과가 우연에 의하여 결정되는 실험이나 관찰을 **시행**이라고 한다. 어떤 시행에서 일어날 수 있는 모든 가능한 결과 전체의 집합을 **표본공간**(sample space  $S$ )이라 하고, 표본공간의 부분집합을 **사건**(event)이라고 한다.

**정의 4.2.** 표본공간의 부분집합 중에서 원소의 개수가 한 개인 집합을 **근원사건**이라 하고, 반드시 일어나는 사건은 **전사건**, 절대로 일어나지 않는 사건은 **공사건**( $\emptyset$ )이라 한다.

**정의 4.3.** 두 사건  $A, B$ 에 대하여,  $A$  또는  $B$ 가 일어나는 사건을  $A$ 와  $B$ 의 **합사건**이라 하고,  $A \cup B$ 로 나타낸다. 그리고  $A$ 와  $B$ 가 동시에 일어나는 사건을  $A$ 와  $B$ 의 **곱사건**이라 하고,  $A \cap B$ 로 나타낸다.

**정의 4.4.** 표본공간  $S$ 의 부분집합인 두 사건  $A, B$ 에 대하여  $A \cap B = \emptyset$ 이면  $A$ 와  $B$ 는 서로 **배반사건**(disjoint)이라 한다. 또, 사건  $A$ 가 일어나지 않는 사건을 사건  $A$ 의 **여사건**이라 하고,  $A^C$ 로 나타낸다.

**정의 4.5.** 표본공간  $S$ 의 공사건이 아닌 사건  $A_1, \dots, A_n$ 이 다음 조건을 만족하면,

$$(1) \bigcup_{i=1}^n A_i = S$$

$$(2) A_i \cap A_j = \emptyset \quad (\text{for all } 1 \leq i \neq j \leq n) \quad (\text{pairwise disjoint})$$

사건  $A_1, \dots, A_n$ 을  $S$ 의 **분할**(partition)이라 한다.

**정의 4.6.** 어떤 시행에서 사건  $A$ 가 일어날 가능성을 수로 나타낸 것을 사건  $A$ 가 일어날 **확률**이라 하고, 기호로  $P(A)$ 와 같이 나타낸다.

**정의 4.7.** (수학적 확률) 어떤 시행의 표본공간  $S$ 가  $m$ 개의 근원사건으로 이루어져 있고, 각 근원사건이 일어날 가능성이 모두 같은 정도로 기대될 때, 사건  $A$ 가  $r$ 개의 근원사건으로 이루어져 있으면 사건  $A$ 가 일어날 확률은 다음과 같다.

$$P(A) = \frac{(\text{사건 } A \text{가 일어나는 경우의 수})}{(\text{모든 경우의 수})} = \frac{n(A)}{n(S)} = \frac{r}{m}$$

**정의 4.8.** (통계적 확률) 같은 시행을  $n$ 번 반복하여 사건  $A$ 가 일어난 횟수를  $r_n$ 이라고 하자. 이 때, 시행 횟수  $n$ 이 한없이 커짐에 따라 그 상대도수  $r_n/n$ 은  $P(A)$ 에 가까워진다.

$$P(A) = \lim_{n \rightarrow \infty} \frac{r_n}{n}$$

**정의 4.9.** (기하학적 확률) 연속적인 변량을 크기로 갖는 표본공간의 영역  $S$  안에서 각각의 점을 잡을 가능성이 같은 정도로 기대될 때, 영역  $S$ 에 포함되어 있는 영역  $A$ 에 대하여 영역  $S$ 에서 임의로 잡은 점이 영역  $A$ 에 속할 확률은 다음과 같다.

$$P(A) = \frac{(\text{영역 } A \text{의 크기})}{(\text{영역 } S \text{의 크기})}$$

**정의 4.10.** (확률의 공리 - Axioms of Probability) 표본공간  $S$ 와 사건  $A$ 에 대하여,

$$(1) 0 \leq P(A) \leq 1$$

$$(2) P(S) = 1$$

$$(3) \text{서로 배반인 사건열 } A_1, A_2, \dots \text{에 대해 } P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

**정리 4.11.** (확률의 기본 성질) 사건  $A, B$ 에 대하여 다음이 성립한다.

$$(1) P(\emptyset) = 0$$

$$(2) P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{확률의 덧셈정리})$$

$$(3) P(A^C) = 1 - P(A) \quad (\text{여사건의 확률})$$

**정리 4.12.** 사건  $A, B, C$ 에 대하여 다음이 성립한다.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$$

**정리 4.13. (포함 배제 원리)** 사건  $A_1, \dots, A_n$ 에 대하여 다음이 성립한다.

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k+1} \left( \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) \right)$$

## 5 조건부확률

**정의 5.1.** 확률이 0이 아닌 두 사건  $A, B$ 에 대하여 사건  $A$ 가 일어났을 때, 사건  $B$ 가 일어날 확률을 사건  $A$ 가 일어났을 때의 사건  $B$ 의 **조건부확률**(conditional probability)이라 하고, 기호로  $P(B|A)$  와 같이 나타낸다. 이는 다음과 같이 계산한다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (P(A) > 0)$$

**연습문제 5.2.** 주사위 한 개를 던지는 시행에서 소수의 눈이 나오는 사건을  $A$ , 홀수의 눈이 나오는 사건을  $B$ 라 할 때, 다음을 구하여라.

(1)  $P(A \cap B)$

(2)  $P(B|A)$

(3)  $P(A^C|B)$

**정리 5.3. (확률의 곱셈정리)** 공사건이 아닌 두 사건  $A, B$ 에 대하여 다음이 성립한다.

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$$

**연습문제 5.4.** 장난감 100개 중 20개가 불량품이다. 이 중 2개를 임의로 추출할 때, 2개 모두 불량품일 확률을 구하여라.

**정의 5.5.** 두 사건  $A, B$ 에 대하여 사건  $A$ 가 일어났을 때의 사건  $B$ 의 조건부확률이 사건  $B$ 가 일어날 확률과 같을 때, 즉

$$P(B|A) = P(B|A^C) = P(B)$$

이면, 두 사건  $A, B$ 는 서로 **독립**(independent)이라 하고, 기호로  $A \perp B$  와 같이 나타낸다. 두 사건이 독립이 아닐 때는 **종속**이라 한다.

**연습문제 5.6.** 주사위 2개를 던질 때, 다음 두 사건이 독립인지 판정하여라.

- (1)  $A$ : 두 주사위 눈의 합이 6인 사건,  $B$ : 첫 번째 주사위 눈이 4인 사건
- (2)  $A$ : 두 주사위 눈의 합이 7인 사건,  $B$ : 첫 번째 주사위 눈이 4인 사건

**정리 5.7.** 공사건이 아닌 두 사건  $A, B$ 에 대하여 다음 조건은 서로 동치이다.

- (1)  $A, B$ 가 서로 독립이다.
- (2)  $P(A \cap B) = P(A)P(B)$

**정리 5.8.** 공사건이 아닌 두 사건  $A, B$ 에 대하여 다음이 성립한다.

$$[A, B \text{가 독립}] \iff [A^C, B \text{가 독립}] \iff [A, B^C \text{가 독립}] \iff [A^C, B^C \text{가 독립}]$$

**연습문제 5.9.** 다음을 증명하여라.

- (1) 정리 5.8.
- (2) 공사건이 아닌 두 사건  $A, B$ 가 서로 배반사건이면, 사건  $A, B$ 는 종속이다.

**연습문제 5.10.** 두 사건  $A, B$ 가 서로 독립이고  $P(A) = 0.25, P(B) = 0.4$  일 때, 다음을 구하여라.

- (1)  $P(A \cap B)$
- (2)  $P(A^C \cap B)$
- (3)  $P(A | B^C)$
- (4)  $P(B^C | A^C)$

**연습문제 5.11.** 두 사건  $A, B$ 가 서로 독립이고,  $P(A \cup B) = 0.8, P(A \cap B) = 0.3$  일 때,  $P(A), P(B)$ 를 각각 구하여라. (단,  $P(A) > P(B)$ )

**정의 5.12.** 공사건이 아닌 사건  $A_1, \dots, A_n$ 에 대하여,

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \text{for all } 1 \leq i \neq j \leq n$$

이 성립하면 사건  $A_1, \dots, A_n$ 이 **쌍마다 독립**(pairwise independent)이라고 한다.

**정의 5.13.** 공사건이 아닌 사건  $A_1, \dots, A_n$ 가 있다. 임의의  $J \subset \{1, 2, \dots, n\}$ 에 대해

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$$

이 성립하면 사건  $A_1, \dots, A_n$ 이 **상호 독립**(mutually independent)이라고 한다.

**연습문제 5.14.** 사건  $A, B, C$ 에 대하여  $P(A \cap B \cap C) = P(A)P(B)P(C)$  이지만  $A, B, C$ 가 상호 독립은 아닌  $A, B, C$ 의 예시를 찾아라.

**정의 5.15.** 동일한 시행을 반복할 때, 각 시행에서 일어나는 사건이 서로 독립이면 이러한 시행을 **독립시행**이라고 한다.

**정리 5.16. (독립시행의 확률)** 1회의 시행에서 사건  $A$ 가 일어날 확률을  $p$  라 할 때, 이 시행을  $n$ 회 반복하는 독립시행에서 사건  $A$ 가  $r$ 번 일어날 확률은 다음과 같다.

$$\binom{n}{r} p^r (1-p)^{n-r} \quad (r = 0, 1, \dots, n)$$

**연습문제 5.17.** 주사위를 한 번 던지는 시행에서 3의 배수의 눈이 나오는 사건을  $A$ 라 할 때, 다음 물음에 답하여라.

(1)  $P(A)$ 를 구하여라.

(2) 주사위를 20번 던지는 시행에서 사건  $A$ 가 12번 일어날 확률을 구하여라.



(3) 주사위를 100번 던지는 시행에서 사건  $A$ 가  $r$ 번 일어날 확률을 구하여라.

**정리 5.18. (전확률공식 - Law of Total Probability)** 표본공간  $S$ 의 분할인 사건  $A_1, \dots, A_n$ 에 대하여 다음이 성립한다.

$$P(B) = \sum_{i=1}^n P(B | A_i) P(A_i)$$

**증명.**  $P(B) = P\left(\bigcup_{i=1}^n (B \cap A_i)\right) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B | A_i) P(A_i)$  □

**연습문제 5.19.** H 대학의 통계학과 학생의 30%는 1학년이고, 25%는 2학년이고, 25%는 3학년이고, 20%는 4학년이라고 하자. 그런데 1학년의 50%, 2학년의 30%, 3학년의 10%, 4학년의 2%가 수학 과목의 수강생이라 한다. 통계학과 학생 중 한 학생을 임의로 선택할 때 그 학생이 수학 과목의 수강생일 확률을 구하여라.

**정리 5.20. (베이즈 정리 - Bayes' Theorem)** 사건  $A_1, \dots, A_n$ 이 표본공간  $S$ 의 분할이고  $P(B) > 0$  이면 다음이 성립한다.

$$P(A_k | B) = \frac{P(B | A_k) P(A_k)}{\sum_{i=1}^n P(B | A_i) P(A_i)} \quad (k = 1, \dots, n)$$

**증명.** 조건부확률의 정의와 전확률공식으로부터 자명하다. □

**연습문제 5.21.** 베이즈 정리를 증명하여라.

**연습문제 5.22.** 주머니 A에는 흰 공 3개, 검은 공 2개가 들어 있고, 주머니 B에는 흰 공 2개, 검은 공 3개가 들어 있다. 임의로 주머니 하나를 택하여 2개의 공을 동시에 꺼냈더니 모두 흰 공이었을 때, 그것이 주머니 B에서 나왔을 확률을 구하여라.

**연습문제 5.23.** A 주머니에 흰 공 2개, 검은 공 5개, B 주머니에 흰 공 3개, 검은 공 4개가 들어있다. A 주머니에서 한 개의 공을 임의로 꺼내어 B 주머니에 넣은 다음 다시 B 주머니에서 하나의 공을 꺼내기로 한다. B에서 꺼낸 공이 흰 공일 때, A에서 B로 옮겨진 공이 흰 공이었을 확률을 구하여라.

**연습문제 5.24.** 어떤 지역의 결핵환자의 비율이 0.1%로 알려져 있다. 결핵에 걸려있는지 알아보는 검사에서 결핵에 걸렸을 때 양성 반응이 나타날 확률은 95%이고 그렇지 않을 때 양성 반응이 나타날 확률은 1.1%라고 한다. 양성 반응이 나타났을 때 결핵에 걸렸을 확률을 구하여라.

## 6 이산확률변수

**정의 6.1.** 표본공간 위에 정의된 실수값 함수를 **확률변수**(random variable)라 하고, 확률변수  $X$ 의 값에 따라 확률이 어떻게 흩어져 있는지를 합이 1인 양수으로써 나타낸 것을  $X$ 의 **확률분포**(probability distribution)라고 한다.

**정의 6.2.** 확률변수  $X$ 가 어떤 값  $x$ 를 취할 확률은 기호로  $P(X = x)$ ,  $a$  이상  $b$  이하의 값을 취할 확률은 기호로  $P(a \leq X \leq b)$  와 같이 나타낸다.

**정의 6.3.** 유한 개이거나, 자연수와 같이 셀 수 있는 값을 취하는 확률변수를 **이산확률변수**(discrete random variable)라 한다.

**연습문제 6.4.** 서로 다른 3개의 동전을 던지는 시행에서 뒷면이 나오는 동전의 개수를  $X$ 라 할 때, 확률변수  $X$ 의 확률분포를 표로 나타내어라.

$X$	합계
$P(X = x)$	1

**연습문제 6.5.** 검은 공 2개와 흰 공 4개가 들어 있는 주머니에서 동시에 3개의 공을 꺼낼 때 나오는 검은 공의 개수를  $X$ 라 하자. 확률변수  $X$ 의 확률분포를 표로 나타내어라.

**정의 6.6.** 이산확률변수  $X$ 가 취할 수 있는 값이  $x_1, \dots, x_n$  일 때,  $X$ 의 각 값에  $[X$ 가 이 값을 취할 확률  $p_1, \dots, p_n]$  을 대응시키는 함수

$$P(X = x_i) = p_i \quad (i = 1, \dots, n)$$

를 이산확률변수  $X$ 의 **확률질량함수**(probability mass function)라 하며, 확률질량함수는 다음 조건을 만족해야 한다.

$$(1) \ 0 \leq P(X = x_i) \leq 1$$

$$(2) \sum_{i=1}^n P(X = x_i) = 1$$

$$(3) P(a \leq X \leq b) = \sum_{x=a}^b P(X = x)$$

**연습문제 6.7.** 5개의 제비 중에 3개의 당첨 제비가 있다. 임의로 뽑은 2개의 제비 중에 있는 당첨 제비의 개수를  $X$ 라고 할 때, 확률변수  $X$ 의 확률질량함수를 구하여라.

**연습문제 6.8.** 주어진 이산확률변수  $X$ 에 대해 다음 값을 구하여라.

$X$	0	1	2	3	4
$P(X = x)$	$a$	$\frac{1}{8}$	$b$	$\frac{1}{4}$	$\frac{1}{8}$

$$(1) a + b$$

$$(2) P(X = 1 \cup X = 3)$$

$$(3) P(0 \leq X \leq 2)$$

**정의 6.9.** 이산확률변수  $X$ 가 취하는 값이  $x_1, \dots, x_n$ 일 때,

- **평균(mean), 기댓값(expectation):**  $\mu = E(X) = \sum_{i=1}^n x_i \cdot P(X = x_i)$
- **분산(variance):**  $V(X) = E((X - \mu)^2) = \sum_{i=1}^n (x_i - \mu)^2 \cdot P(X = x_i)$
- **표준편차(standard deviation):**  $\sigma(X) = \sqrt{V(X)}$

**연습문제 6.10.** 빨간 공 4개, 흰 공 6개가 들어 있는 주머니에서 3개의 공을 꺼낼 때, 빨간 공이 나오는 개수를  $X$ 라 한다.  $E(X)$ ,  $V(X)$ ,  $\sigma(X)$  를 구하여라.

**정의 6.11.** 이산확률변수  $X$ 와 함수  $g(x)$ 에 대하여, 다음이 성립한다.

$$E(g(X)) = \sum_{x \in X} g(x) \cdot P(X = x)$$

**연습문제 6.12.** 확률변수  $X$ 에 대하여 다음을 보여라. (단,  $a, b$ 는 상수)

$$(1) \mathbf{E}(aX + b) = a\mathbf{E}(X) + b$$

$$(2) \mathbf{V}(aX + b) = a^2\mathbf{V}(X)$$

$$(3) \sigma(aX + b) = |a| \cdot \sigma(X)$$

$$(4) \mathbf{V}(X) = \mathbf{E}(X^2) - \{\mathbf{E}(X)\}^2$$

**연습문제 6.13.** 연습문제 6.10 의 확률변수  $X$ 에 대하여  $\mathbf{E}(5X + 4), \mathbf{V}(5X - 1), \sigma(5X + 3)$ 을 각각 구하여라.

**정의 6.14.** 두 확률변수  $X, Y$ 에 대하여 이들이 취할 수 있는 값들의 모든 순서쌍에 확률이 할어진 정도를 합이 1인 양수로 나타낸 것을  $X, Y$ 의 **결합분포**(joint probability distribution)라 한다.

**정의 6.15.** 이산확률변수  $X, Y$ 에 대하여  $X$ 가 취할 수 있는 값을  $x_1, \dots, x_n$ ,  $Y$ 가 취할 수 있는 값을  $y_1, \dots, y_m$  이라 하자. 순서쌍  $(x_i, y_j)$ 에 대하여 [결합분포가 이 값을 취할 확률  $p_{ij}$ ]을 대응시키는 함수

$$P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j) = p_{ij}$$

를 이산확률변수  $X, Y$ 의 **결합확률밀도함수**라 하며, 이 함수는 다음 조건을 만족해야 한다.

$$(1) 0 \leq P(X = x_i, Y = y_j) \leq 1$$

$$(2) \sum_{i=1}^n \sum_{j=1}^m P(X = x_i, Y = y_j) = 1$$

$$(3) P(a \leq X \leq b, c \leq Y \leq d) = \sum_{x=a}^b \sum_{y=c}^d P(X = x, Y = y)$$

**정의 6.16.** 이산확률변수  $X, Y$ 의 결합확률밀도함수  $P(X = x_i, Y = y_j)$ 에서 확률변수  $X$ 와  $Y$ 의 확률분포를 얻을 수 있다.

$$P(X = x_i) = \sum_{y \in Y} P(X = x_i, Y = y), \quad P(Y = y_j) = \sum_{x \in X} P(X = x, Y = y_j)$$

이를 각각  $X, Y$ 의 **주변확률밀도함수**라 한다.

**연습문제 6.17.** 하나의 주사위를 던져서 나오는 눈의 종류에 따라 상금이 걸린 게임이 있다. A, B 두 사람이 다음과 같은 게임을 한다.

**A:** 1 또는 2 가 나오면 100원, 3 또는 4가 나오면 200원, 5 또는 6이 나오면 300원

**B:** 짝수가 나오면 100원, 홀수가 나오면 (눈의 수 $\times$ 100)원

이 때, A의 수입을  $X$ , B의 수입을  $Y$ 라 하자.  $X, Y$ 의 결합확률분포, 주변분포를 구하고, 확률변수  $Z = X + Y$ 로 정의할 때,  $Z$ 의 확률분포와  $E(Z)$ 를 구하여라.

$x \setminus y$				행의 합
열의 합				1

$z$						합계
$P(Z = z)$						1

**정의 6.18.** 이산확률변수  $X, Y$ 와 함수  $g(x, y)$ 에 대하여, 다음이 성립한다.

$$E(g(X, Y)) = \sum_{x \in X} \sum_{y \in Y} g(x, y) \cdot P(X = x, Y = y)$$

**연습문제 6.19.** 확률변수  $X, Y$ 에 대하여  $E(aX + bY) = aE(X) + bE(Y)$  임을 보여라. (단,  $a, b$ 는 상수)

**정의 6.20.** 이산확률변수  $X, Y$ 가 주어져 있다. 임의의  $x_i \in X, y_j \in Y$ 에 대해 다음이 성립하면 확률변수  $X, Y$ 는 서로 독립이라 한다.

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

**연습문제 6.21.** 확률변수  $X, Y$ 가 서로 독립일 때, 다음이 성립함을 보여라.

(1)  $E(XY) = E(X)E(Y)$

(2)  $V(X \pm Y) = V(X) + V(Y)$

**연습문제 6.22.** 위 연습문제의 역은 성립하지 않음을 보여라.<sup>1</sup>

**정리 6.23.** (큰 수의 법칙) 어떤 시행에서 사건  $A$ 가 일어날 확률이  $p$ 일 때,  $n$ 번의 독립시행에서 사건  $A$ 가 일어나는 횟수를  $X$ 라 하면, 임의의 양수  $h$ 에 대하여 다음이 성립한다.

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X}{n} - p\right| < h\right) = 1$$

---

<sup>1</sup>Hint:  $-1, 1$ 을 각각  $1/2$ 의 확률로 취하는 확률변수  $X$ 에 대해  $X$ 와  $X^2$ 을 고려한다.

## 7 이산확률분포

**정의 7.1.** 시행의 결과가 오직 성공(success,  $s$ ) 또는 실패(failure,  $f$ )뿐이며, 각 시행이 독립이고, 성공의 확률이  $p$ 로 항상 일정한 시행을 **베르누이 시행**(Bernoulli trial)이라 한다. 성공하면 1, 실패하면 0을 값으로 갖는 확률변수를 **베르누이 확률변수**(Bernoulli random variable)라 한다.

**정의 7.2.** 베르누이 확률변수의 확률분포를 **베르누이 분포**(Bernoulli distribution)라 하고,  $X$ 가 성공 확률이  $p$ 인 베르누이 분포를 따를 때,  $X \sim \text{Berr}(p)$ 와 같이 나타낸다.<sup>2</sup>

**연습문제 7.3.**  $X \sim \text{Berr}(p)$ 일 때,  $X$ 의 확률분포표를 구하고,  $\mathbf{E}(X)$ 와  $\mathbf{V}(X)$ 를 구하여라.

**정의 7.4.** 한 번의 시행에서 사건  $A$ 가 일어날 확률이  $p$ 로 일정할 때,  $n$ 번의 독립시행에서 사건  $A$ 가 일어나는 횟수를  $X$ 라고 하면 확률변수  $X$ 의 확률분포를 **이항분포**(binomial distribution)라 하고 기호로  $B(n, p)$ 와 같이 나타낸다.

**정의 7.5.** 성공 확률이  $p$ 인 베르누이 시행을  $n$ 번 독립적으로 반복 시행할 때, 성공 횟수의 분포를 **이항분포**라 한다. 즉,  $i = 1, \dots, n$ 에 대하여  $X_i \sim_{i.i.d} \text{Berr}(p)$ 일 때,<sup>3</sup> 이항분포는  $n$ 개의 베르누이 확률변수의 합으로 정의된다.<sup>4</sup>

$$\sum_{i=1}^n X_i = X \sim B(n, p)$$

---

<sup>2</sup> $X$ 가 분포  $\mathcal{A}$ 를 따를 때,  $X \sim \mathcal{A}$ 와 같이 표기한다.

<sup>3</sup>*i.i.d.*: independent and identically distributed.

<sup>4</sup>따라서  $\text{Berr}(p) = B(1, p)$ 이다.



**정의 7.6.**  $X \sim B(n, p)$ 일 때,  $X$ 의 확률질량함수는 다음과 같다.

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (x = 0, \dots, n)$$

**연습문제 7.7.** 다음 확률변수  $X$ 가 이항분포를 따르는지 조사하시오.

- (1) 10개의 동전을 동시에 던질 때 뒷면이 나오는 동전의 개수  $X$
- (2) 검정 구슬 4개와 흰 구슬 2개 중에서 차례로 2개의 구슬을 꺼낼 때 나오는 흰 구슬의 개수  $X$
- (3) 4지선다형 문제 12개에 임의로 답할 때 정답의 개수  $X$

**연습문제 7.8.** 타율이 0.2인 야구 선수가 10번의 타석에서 안타를 친 횟수를  $X$ 라 하자.  $P(X \leq 9)$ 의 값을 구하여라.

**연습문제 7.9.**  $X \sim B(n, p)$ 이면,  $E(X) = np$ ,  $V(X) = np(1 - p)$ 임을 보여라.

**연습문제 7.10.** 두 사람 A, B가 게임을 한다. 매 회 동전을 던져 앞면이 나오면 A가 이기고, 뒷면이 나오면 B가 이긴다. 10회 게임을 할 때, A가 이긴 횟수를  $X$ , B가 이긴 횟수를  $Y$ 라 하자.  $E(X)$ ,  $E(Y)$ 를 구하여라.

**연습문제 7.11.**  $X \sim B(100, p)$ 라 하자.  $X$ 의 분산이 최대일 때,  $E(X)$ 의 값을 구하여라.

**정의 7.12.** 특성값 1의 개수가  $D$ , 0의 개수가  $N - D$  인 크기  $N$ 의 유한 모집단에서 크기  $n$ 인 랜덤 표본을 뽑을 때, 표본에서 1의 개수를  $X$ 라 하자. 이 때, 확률변수  $X$ 가 따르는 분포를 **초기하분포**(hypergeometric distribution)라 하고, 기호로는  $X \sim H(N, D, n)$ 으로 나타낸다. 초기하분포의 확률질량함수는 다음과 같이 주어진다.

$$P(X = x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}} \quad (\text{단, } \max\{0, n - N + D\} \leq x \leq \min\{n, D\})$$

**정리 7.13.**  $X \sim H(N, D, n)$  일 때, 다음이 성립한다.

$$E(X) = np, \quad V(X) = np(1-p) \frac{N-n}{N-1} \quad \left(p = \frac{D}{N}\right)$$

**정리 7.14.**  $X \sim H(N, D, n)$  일 때,  $N \gg n$  이면  $X$ 는 근사적으로  $B(n, D/N)$  를 따른다.

**정의 7.15.** 성공 확률이  $p$ 인 베르누이 시행을 반복하여 최초로 성공할 때 까지의 시행 횟수를  $X$ 라 하자. 이 때, 확률변수  $X$ 는 **기하분포**(geometric distribution)을 따른다. 기하분포의 확률질량함수는 다음과 같이 주어진다.

$$P(X = k) = p(1-p)^{k-1} \quad (k = 0, 1, \dots)$$

**연습문제 7.16.** 성공 확률이  $p$ 인 기하분포를 따르는 확률변수  $X$ 에 대하여 다음이 성립함을 보여라.

$$E(X) = \frac{1}{p}, \quad V(X) = \frac{1-p}{p^2}$$

**정리 7.17.** 실수열  $\{p_n\}$  ( $0 \leq p_i \leq 1$  for all  $i$ ) 에 대하여  $\lim_{n \rightarrow \infty} np_n = \lambda$  라 하면 다음이 성립한다.

$$\lim_{n \rightarrow \infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}$$

**정의 7.18.** 정해진 시간 안에 어떤 사건이 일어날 횟수에 대한 기댓값을  $\lambda$ 라 할 때, 그 사건이 일어난 횟수를  $X$ 라 하자. 이 때, 확률변수  $X$ 는 **포아송 분포**(Poisson distribution)를 따르며, 기호로는  $X \sim \text{Poi}(\lambda)$  로 나타낸다. 포아송 분포의 확률질량함수는 다음과 같이 주어진다.

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k = 0, 1, \dots)$$

**정리 7.19.**  $X \sim B(n, p)$  일 때,  $n$ 이 충분히 크면  $X \sim \text{Poi}(np)$  이다.<sup>5</sup>

**연습문제 7.20.**  $X \sim \text{Poi}(\lambda)$  일 때,  $E(X) = \lambda$ ,  $V(X) = \lambda$  임을 보여라.

**연습문제 7.21.** 오후 3시부터 4시에 어느 병원에 도착하는 손님이 평균적으로 6.5명이라 하자. 오늘 오후 3시부터 4시 사이에 도착하는 손님 수에 대한 확률질량함수를 구하고, 도착한 손님이 4명일 확률, 최대 2명일 확률을 각각 구하여라.

---

<sup>5</sup>대략  $n \geq 100, np \leq 10$  이면 근사할 수 있다.

## 8 연속확률변수

**정의 8.1.** 어떤 구간에 속하는 모든 실수 값을 취할 수 있는 확률변수를 **연속확률변수**(continuous random variable)라 한다. 연속확률변수  $X$ 가 구간  $[\alpha, \beta]$ 에 속하는 모든 실수 값을 취하고,

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad (\alpha \leq a \leq x \leq b \leq \beta)$$

와 같이 나타낼 수 있을 때, 함수  $f(x)$ 를  $X$ 의 **확률밀도함수**(probability density function)라 하며, 확률밀도함수는 다음 조건을 만족해야 한다.

(1)  $f(x) \geq 0$

(2)  $\int_{\alpha}^{\beta} f(x)dx = 1$

(3)  $\alpha \leq a \leq x \leq b \leq \beta$  일 때,

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) = \int_{\alpha}^b f(x)dx - \int_{\alpha}^a f(x)dx = \int_a^b f(x)dx$$

**정의 8.2.** 연속확률변수  $X$ 가 구간  $[\alpha, \beta]$ 에 속하는 모든 실수 값을 취할 때,

- **평균**(mean), **기댓값**(expectation):  $\mu = \mathbf{E}(X) = \int_{\alpha}^{\beta} xf(x)dx$
- **분산**(variance):  $\mathbf{V}(X) = \mathbf{E}((X - \mu)^2) = \int_{\alpha}^{\beta} (x - \mu)^2 f(x)dx$
- **표준편차**(standard deviation):  $\sigma(X) = \sqrt{\mathbf{V}(X)}$

**연습문제 8.3.** 연속확률변수  $X$ 의 확률밀도함수가  $f(x) = ax$  ( $0 \leq x \leq 2$ ) 일 때, 상수  $a$ 의 값과  $P(0.5 \leq X \leq 1)$ 의 값을 구하고,  $\mathbf{E}(X)$ 와  $\mathbf{V}(X)$ 의 값을 구하여라.

**정의 8.4.** 연속확률변수  $X$ 가 모든 실수 값을 취하고, 확률밀도함수  $f(x)$ 가 다음과 같이 주어질 때,

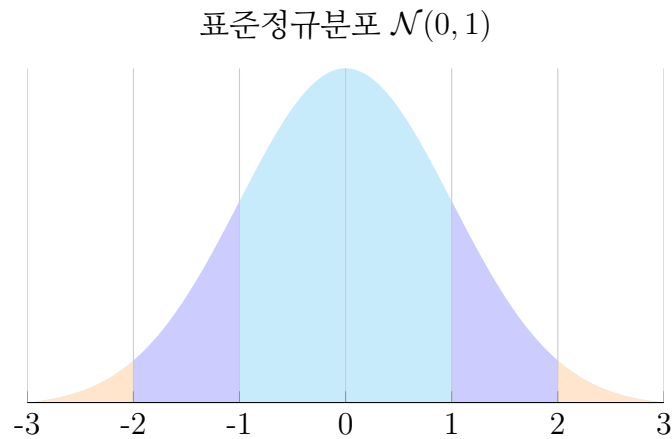
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < \infty)$$

$X$ 의 확률분포를 **정규분포**(normal distribution)라 하고, 평균이  $\mu$ , 분산이  $\sigma^2$ 인 정규분포를 기호로  $\mathcal{N}(\mu, \sigma^2)$ 와 같이 나타낸다.

**정리 8.5.** 서로 독립인 확률변수  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , 상수  $c_i$  ( $i = 1, \dots, k$ ) 에 대하여

$$X = \sum_{i=1}^k c_i X_i \sim \mathcal{N}\left(\sum_{i=1}^k c_i \mu_i, \sum_{i=1}^k c_i^2 \sigma_i^2\right)$$

이 성립한다.



**정의 8.6.**  $\mathcal{N}(0, 1)$  을 **표준정규분포**(standard normal distribution)라 하고, 확률밀도함수  $f(z)$ 는 다음과 같다.

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

**정리 8.7.** (정규분포의 표준화)  $X \sim \mathcal{N}(\mu, \sigma^2)$  일 때,

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

가 성립하고, 이를 **정규분포의 표준화**(standardization)라 한다. 이렇게 표준화된 값을 **z-점수**(z-score)라 하고, 다음이 성립한다.

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)$$

**정리 8.8.** 표준정규분포  $\mathcal{N}(0, 1)$  의 확률밀도함수  $f(z)$ 는 다음과 같은 성질을 갖는다.

- 곡선과  $x$ 축 사이의 넓이는 1이다.
- 직선  $x = 0$  에 대하여 대칭이다.

이로부터 다음이 성립함을 알 수 있다. 임의의 실수  $a, b$  에 대하여

- (1)  $P(Z \geq a) = P(Z \leq -a)$ .
- (2)  $P(Z \geq a) = 1 - P(Z < a)$
- (3)  $P(a \leq Z \leq b) = P(Z \leq b) - P(Z < a)$

**연습문제 8.9.** 표준정규분포표가 주어져 있다.  $X \sim \mathcal{N}(27, 4^2)$  일 때, 다음을 구하여라.

$z$	$P(0 \leq Z \leq z)$
0.5	0.1915
1.0	0.3413
1.5	0.4332
2.0	0.4772

- (1)  $P(X \leq 21)$
- (2)  $P(29 \leq X \leq 35)$
- (3)  $P(X \geq 25)$

**정의 8.10.**  $\mathcal{N}(0, 1)$  의  $100(1 - \alpha)$  백분위수를  $z_\alpha$  로 나타낸다. 즉  $P(Z \geq z_\alpha) = \alpha$  이다.<sup>6</sup>

**연습문제 8.11.**  $P(Z < 1.96) = 0.975$ ,  $P(Z < 2.58) = 0.995$  일 때,  $z_{0.025}$ ,  $z_{0.005}$  의 값을 구하여라.

---

<sup>6</sup>상방백분위수라고도 한다.

**정리 8.12.** (68.26-95.44-99.74 Rule)  $X \sim \mathcal{N}(\mu, \sigma^2)$  일 때,

- 68.26% 의 관측값들이  $[\mu - \sigma, \mu + \sigma]$  에 있다.  $P(|X - \mu| < \sigma) = 0.6826$ .
- 95.44% 의 관측값들이  $[\mu - 2\sigma, \mu + 2\sigma]$  에 있다.  $P(|X - \mu| < 2\sigma) = 0.9544$ .
- 99.74% 의 관측값들이  $[\mu - 3\sigma, \mu + 3\sigma]$  에 있다.  $P(|X - \mu| < 3\sigma) = 0.9974$ .

**연습문제 8.13.** 전세계 사람들의 IQ는 평균이 100, 표준편차가 16인 정규분포를 따른다고 한다. IQ가 116, 132, 148인 사람은 각각 IQ 상위 몇 % 인지 구하여라.

**연습문제 8.14.** 대수능 모의평가에서 어느 고등학교 3학년 학생 500명의 수학 성적이 평균 70점, 표준편차 20점인 정규분포를 따른다고 한다. 300등을 한 학생의 점수를 구하여라.

$z$	$P(0 \leq Z \leq z)$
0.25	0.1
0.52	0.2
1.28	0.4

**연습문제 8.15.** 다음은 9등급제의 산출 방식 중 일부이다.

등급	$z$	$P(Z \geq z)$
1	1.75	0.04
2	1.25	0.11
3	0.75	0.23
4	0.25	0.40

어떤 시험의 평균이 50점이고 표준편차가 18점일 때, 각 등급컷 점수를 구하여라.

**정리 8.16.** (드 무아브르-라플라스의 정리)  $X \sim B(n, p)$  일 때,  $n$ 이 충분히 크면<sup>7</sup>  $X$ 는 근사적으로  $\mathcal{N}(np, np(1-p))$ 를 따른다.

**정리 8.17.** (연속성 수정 - continuity correction) 연속확률분포를 이용하여 이산확률분포의 확률을 근사시킬 때, 근사의 정밀도를 높이는데 사용한다.  $X \sim B(n, p)$  일 때,

$$P(a \leq X \leq b) \approx P\left(\frac{a - np - 0.5}{\sqrt{np(1-p)}} \leq Z \leq \frac{a - np + 0.5}{\sqrt{np(1-p)}}\right)$$

**연습문제 8.18.** 현재 20살인 사람이 45년 후 살아있을 확률이 0.8 이라고 한다. 20살인 사람 500명을 임의로 추출했을 때, 다음 값을 식으로 표현하여라.

- (1) 45년 후, 정확히 390명이 살아있을 확률
- (2) 45년 후, 375명 이상 425명 이하의 사람들이 살아있을 확률

---

<sup>7</sup>일반적으로,  $np \geq 5, n(1-p) \geq 5$  이면 근사한다.



## 9 표본분포

모집단 전체를 조사하는 것은 비용이 많이 들고, 또 현실적으로 어렵다. 따라서 통계적 추정을 할 때에는 표본을 뽑아 조사하는 것이 경제적이다.

**정의 9.1.** 모집단에서 임의추출한 표본으로부터 얻은 통계량은 확률변수이므로 분포를 가지게 된다. 이를 **표본분포**(sampling distribution)라 한다.

**정의 9.2.** 모집단에서 임의추출한 크기  $n$ 인 표본을  $X_1, \dots, X_n$  이라 할 때,

- **표본평균**(sample mean):  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- **표본분산**:  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

표본을 뽑을 때마다 표본평균, 표본분산은 달라질 수 있으므로  $\bar{X}, S^2$ 은 확률변수가 된다. 따라서, 기댓값, 분산, 표준편차도 계산할 수 있다.

**연습문제 9.3.** 모집단  $\{1, 3, 5, 7\}$  에서 크기가 2인 표본을 복원추출 할 때, 표본평균  $\bar{X}$ 의 확률분포가 다음과 같다.

$\bar{X}$	1	2	3	4	5	6	7	합계
$P(\bar{X} = \bar{x})$	$\frac{1}{16}$	$\frac{1}{8}$	$a$	$b$	$\frac{3}{16}$	$c$	$\frac{1}{16}$	1

이 때,  $a, b, c$  의 값을 구하고, 확률변수  $\bar{X}$ 의 기댓값과 분산을 구하여라.

**정의 9.4.**  $X_i$  를  $i$ -번째로 뽑힌 추출단위의 특성값을 나타내는 확률변수라 하자. 다음 조건을 만족하는  $X_1, \dots, X_n$  을 **랜덤표본**(random sample)이라 한다.

- (1) (유한모집단) 단순랜덤 비복원추출로 뽑은 표본
- (2) (무한모집단)  $X_i$  들은 서로 독립이고 각 분포가 모집단 분포와 동일

참고: 유한모집단에서 모집단의 크기가 큰 경우에는 흔히 무한모집단에서의 랜덤표본으로 간주하여 표본분포를 구한 다음 이를 실제표본분포의 근사분포로 사용한다.

**정리 9.5.** 모평균이  $\mu$ 이고 모표준편차가  $\sigma$ 인 모집단에서<sup>8</sup> 복원추출하여 뽑은 크기가  $n$ 인 표본의 표본평균  $\bar{X}$ 에 대하여 다음이 성립한다.

$$\mathbf{E}(\bar{X}) = \mu, \mathbf{V}(\bar{X}) = \frac{\sigma^2}{n}, \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

**증명.**  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  를 이용한다.

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

$$\mathbf{V}(\bar{X}) = \frac{1}{n^2} \mathbf{V}(X_1 + \cdots + X_n) = \frac{1}{n^2} \sum_{i=1}^n \mathbf{V}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

□

참고: 모평균이  $\mu$ , 모분산이  $\sigma^2$ , 크기가  $N$ 인 유한모집단에서 크기  $n$ 인 표본을 비복원추출하는 경우에는 다음이 성립한다.

$$\mathbf{E}(\bar{X}) = \mu, \mathbf{V}(\bar{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

여기서  $\sqrt{\frac{N-n}{N-1}}$  은 finite-population correction factor (FPC) 로 불린다.

**연습문제 9.6.** 연습문제 9.3 에서 구한  $\bar{X}$  의 기댓값과 분산이 위 정리를 사용하여 구한 것과 일치함을 확인하여라.

**정리 9.7.** 표본의 크기가 클수록 표본평균과 모평균의 오차가 줄어든다.

**정리 9.8.** 모집단의 분포가  $\mathcal{N}(\mu, \sigma^2)$  일 때, 표본평균  $\bar{X}$ 는  $\mathcal{N}(\mu, \sigma^2/n)$  을 따른다.

---

<sup>8</sup>무한모집단의 경우 복원추출이나 비복원추출이나 큰 차이가 없다.

**연습문제 9.9.**  $\mathcal{N}(100, 2^2)$  을 따르는 모집단에서 크기가 4인 표본을 임의추출할 때, 표본평균  $\bar{X}$ 가 따르는 분포를 구하여라.

**연습문제 9.10.** 어느 전기 회사에서 생산하는 전구의 수명을 나타내는 확률변수  $X$ 에 대하여,  $X \sim \mathcal{N}(2000, 200^2)$  이라고 한다. 이 회사가 생산한 전구 중 임의추출한  $n$ 개 전구의 평균 수명을  $\bar{X}$ 라 할 때,  $P(1900 \leq \bar{X} \leq 2100) \geq 0.9$  가 성립하기 위한  $n$ 의 최솟값을 구하여라. 단,  $z_{0.05} = 1.65$  이다.

**정리 9.11.** (중심극한정리 - Central Limit Theorem) 평균이  $\mu$ 이고 분산이  $\sigma^2$ 인 임의의 무한모집단에서 표본의 크기  $n$ 이 충분히 크면, 랜덤탐본의 표본평균  $\bar{X}$ 는 근사적으로 정규 분포  $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$  을 따른다.<sup>9</sup>

**연습문제 9.12.** 대학 신입생 신장의 평균이 168cm이고 표준편차가 6cm임이 알려져 있다. 100명의 신입생을 단순랜덤추출하는 경우 표본평균이 167cm 이상 169cm 이하일 확률을 구하여라. 단,  $z_{0.0475} = 1.67$  이다.

**정의 9.13.** 확률변수  $Z_1, \dots, Z_k$  이  $\mathcal{N}(0, 1)$ 의 랜덤탐본일 때,  $V = Z_1^2 + \dots + Z_k^2$  의 분포를 자유도(degrees of freedom)  $k$ 인 카이제곱분포( $\chi^2$ -distribution)라고 한다. 기호로는 다음과 같이 나타낸다.

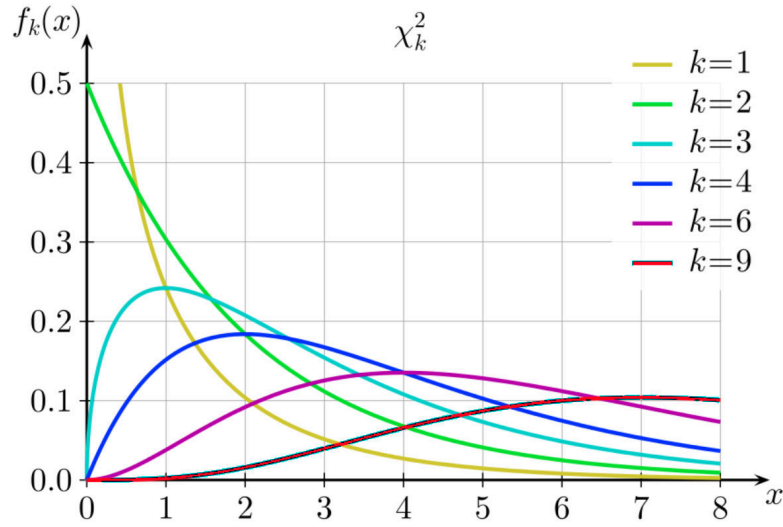
$$Z_1^2 + \dots + Z_k^2 = V \sim \chi^2(k)$$

그리고 확률밀도함수는 다음과 같다.<sup>10</sup>

$$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad (x > 0)$$

<sup>9</sup>당연히,  $n$ 이 클수록 근사는 정확해진다.

<sup>10</sup> $\Gamma$  는 감마함수(Gamma function)을 나타내는 기호이다.



$\chi^2$ -분포의 형태 ( $k$ : 자유도)

**정의 9.14.**  $V \sim \chi^2(k)$  일 때,  $P(V > v) = \alpha$  인  $v$ 의 값을  $\chi_\alpha^2(k)$  로 정의한다.

**정리 9.15.** (카이제곱분포의 가법성)  $V_1, V_2$ 가 서로 독립이면 다음이 성립한다.

- (1)  $V_1 \sim \chi^2(k_1), V_2 \sim \chi^2(k_2)$  이면  $V_1 + V_2 \sim \chi^2(k_1 + k_2)$
- (2)  $V_1 \sim \chi^2(k_1), V_1 + V_2 \sim \chi^2(k_1 + k_2)$  이면  $V_2 \sim \chi^2(k_2)$

**정리 9.16.**  $X_1, \dots, X_n$ 이  $\mathcal{N}(\mu, \sigma^2)$ 의 랜덤포본일 때, 표본분산  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 에 대하여 다음이 성립한다.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

**증명.** 모든  $i$ 에 대해  $\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$  이고 서로 독립이므로 카이제곱분포의 정의로부터

$$\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

이 성립한다. 그런데, 표본분산  $S^2$ 의 정의로부터

$$\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

이고  $\left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi^2(1)$  이므로 가법성에 의해  $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ . □

**정리 9.17.** 분산이 동일한 두 정규모집단  $\mathcal{N}(\mu_1, \sigma^2)$ ,  $\mathcal{N}(\mu_2, \sigma^2)$  에서 각각 뽑은 랜덤표본  $X_1, \dots, X_{n_1}$  과  $Y_1, \dots, Y_{n_2}$  이 서로 독립이라고 하자.  $S_1^2, S_2^2$ 를 각각  $X_i, Y_i$  의 표본분산이라 할 때, **합동표본분산**(pooled sample variance)

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

에 대하여 다음이 성립한다.

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

**증명.** 다음을 이용하면 가법정리에 의해 자명하다.

$$S_p^2 = \frac{(n_1 - 1)}{(n_1 - 1) + (n_2 - 1)} S_1^2 + \frac{(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} S_2^2$$

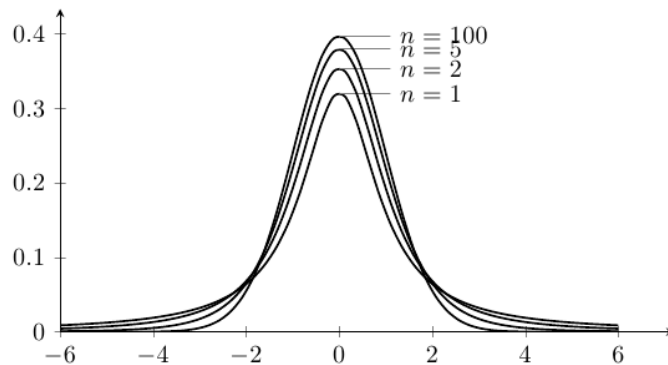
□

**정의 9.18.**  $Z \sim \mathcal{N}(0, 1)$  과 이와 독립인 확률변수  $V$ 가 자유도가  $k$ 인 카이제곱분포를 따를 때,  $T = \frac{Z}{\sqrt{V/k}}$  의 분포를 자유도가  $k$ 인  **$t$ -분포**( $t$ -distribution)라 한다.<sup>11</sup> 기호로는 다음과 같이 나타낸다.

$$\frac{Z}{\sqrt{V/k}} = T \sim t(k)$$

그리고 확률밀도함수는 다음과 같다.

$$\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi} \Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} \quad (-\infty < x < \infty)$$



$t$ -분포의 형태 ( $n$ : 자유도)

<sup>11</sup>스튜던트의(Student's)  $t$ -분포 라고 불리기도 한다.

**정리 9.19.**  $t$ -분포 곡선은 0을 중심으로 좌우 대칭의 밀도곡선을 가지며, 자유도  $k$ 가 커지면  $\mathcal{N}(0, 1)$ 과 비슷하며, 일반적으로 표준정규분포보다 더 두꺼운 꼬리를 갖고 있다.

**정의 9.20.**  $T \sim t(k)$  일 때,  $P(T > t) = \alpha$  인  $t$ 의 값을  $t_\alpha(k)$  로 정의한다.

**정리 9.21.**  $X_1, \dots, X_n$  이  $\mathcal{N}(\mu, \sigma^2)$ 의 랜덤포본일 때, 다음이 성립한다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

**증명.** 다음 변형을 이용한다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu) / \left(\frac{\sigma}{\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2} / (n-1)}}$$

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$  이고,  $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$  이며,  $\bar{X}$  와  $S^2$ 은 서로 독립임이 알려져 있다.  $t$ -분포의 정의에 의해 성립한다.  $\square$

**정리 9.22.** 분산이 동일한 두 정규모집단  $\mathcal{N}(\mu_1, \sigma^2)$ ,  $\mathcal{N}(\mu_2, \sigma^2)$  에서 각각 뽑은 랜덤포본  $X_1, \dots, X_{n_1}$  과  $Y_1, \dots, Y_{n_2}$  이 서로 독립이라고 하자.  $X_i, Y_i$  의 표본분산  $S_1^2, S_2^2$ , 합동표본 분산  $S_p^2$ 에 대하여 다음이 성립한다.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

**증명.** 다음 변형을 이용한다.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{[(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)] / \left(\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}{\sqrt{\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} / (n_1 + n_2 - 2)}}$$

분모는  $\mathcal{N}(0, 1)$  을 따르고, 정리 9.17에 의해  $t$ -분포의 정의를 만족한다.  $\square$

**정의 9.23.**  $V_1 \sim \chi^2(k_1)$ ,  $V_2 \sim \chi^2(k_2)$  이고  $V_1, V_2$ 가 서로 독립일 때,

$$F = \frac{V_1/k_1}{V_2/k_2}$$

의 분포를 자유도  $(k_1, k_2)$  인  $F$ -분포라고 한다. 기호로는 다음과 같이 나타낸다.

$$\frac{V_1/k_1}{V_2/k_2} = F \sim F(k_1, k_2)$$

그리고 확률밀도함수는 다음과 같다.<sup>12</sup>

$$\frac{\sqrt{\frac{(k_1 x)^{k_1} k_2^{k_2}}{(k_1 x + k_2)^{k_1 + k_2}}}}{x B\left(\frac{k_1}{2}, \frac{k_2}{2}\right)} \quad (x > 0)$$

**정의 9.24.**  $F \sim F(k_1, k_2)$  일 때,  $P(F > f) = \alpha$  인  $f$ 의 값을  $F_\alpha(k_1, k_2)$  로 정의한다.

**정리 9.25.**

- (1)  $F \sim F(k_1, k_2)$  이면,  $1/F \sim F(k_2, k_1)$  이다.
- (2)  $F_{1-\alpha}(k_2, k_1) = 1/F_\alpha(k_1, k_2)$
- (3)  $T \sim t(k) \iff T^2 \sim F(1, k)$

**정리 9.26.** 두 정규모집단  $\mathcal{N}(\mu_1, \sigma_1^2)$ ,  $\mathcal{N}(\mu_2, \sigma_2^2)$  에서 각각 뽑은 랜덤표본  $X_1, \dots, X_{n_1}$  과  $Y_1, \dots, Y_{n_2}$  이 서로 독립이라고 하자.  $X_i, Y_i$  의 표본분산  $S_1^2, S_2^2$ 에 대하여 다음이 성립한다.

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

**증명.** 다음과 같이 변형한다.

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\frac{(n_1 - 1)S_1^2}{\sigma_1^2}/(n_1 - 1)}{\frac{(n_2 - 2)S_2^2}{\sigma_2^2}/(n_2 - 1)}$$

$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$ ,  $\frac{(n_2 - 2)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$  이므로  $F$ -분포의 정의를 만족한다.  $\square$

---

<sup>12</sup>B 는 베타함수(Beta function)을 나타낸다.

## 10 통계적 추론

**정의 10.1.** 표본으로부터의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어내는 과정을 **통계적 추론**(statistical inference)이라 한다. 모집단의 특성치(모수)에 대한 추측값을 제공하고 그 오차의 한계를 제시하는 과정을 **추정**(estimation)이라 하고, 다음 두 가지 종류가 있다.

- **점추정**(point estimation): 모수의 참값이라고 추측되는 하나의 추정값을 제공
- **구간추정**(interval estimation): 모수의 참값이 속할 것으로 기대되는 범위를 추측

### 정의 10.2. 정의 및 표기법

- **모수**(population parameter)  $\theta$ : 모집단의 특성을 나타내는 수치적 측도
- 랜덤포본은  $X_1, \dots, X_n$ , 랜덤포본의 관측값은  $x_1, \dots, x_n$  으로 표기한다.
- **추정량**(estimator): 미지의 모수  $\theta$ 의 추정에 사용되는 통계량으로,  $\hat{\theta}(X_1, \dots, X_n)$  혹은  $\hat{\theta}$ 으로 표기한다.
- **추정값**(estimate): 추정량  $\hat{\theta}(X_1, \dots, X_n)$ 의 관측값으로,  $\hat{\theta}(x_1, \dots, x_n)$ 로 표기한다.

추정량과 추정값의 관계는 확률변수와 그 관측값의 관계이다.

모평균  $\mu$ 를 추정하기 위해 추정량으로는 표본평균  $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  을 사용하고, 그 추정값으로는 관측한 결과인  $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  를 사용한다.

**정의 10.3.**  $E(\hat{\theta}) = \theta$  를 만족하는 추정량  $\hat{\theta}$ 를 **불편추정량**(unbiased estimator)이라 한다. 추정량  $\hat{\mu} = \bar{X}$  의 경우  $E(\bar{X}) = \mu$  이므로 불편추정량이다.<sup>13</sup>

**정의 10.4.**  $\theta$ 의 추정량  $\hat{\theta}$ 의 표준편차를 **표준오차**(standard error)라 한다. 즉,  $SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}$  이다. 표준오차는 추정량  $\hat{\theta}$ 의 흩어짐의 정도를 나타낸다.

---

<sup>13</sup>표본분산  $\widehat{\sigma^2} = S^2$  도 불편추정량이다. 사실 불편추정량이 되도록  $n - 1$  로 나눈 것이다...



**정의 10.5.** 랜덤표본  $X_1, \dots, X_n$  으로부터 얻어진 두 추정량  $L(X_1, \dots, X_n), U(X_1, \dots, X_n)$  에 대하여

$$P(L(X_1, \dots, X_n) < \theta < U(X_1, \dots, X_n)) = 1 - \alpha$$

가 성립할 때, 구간  $(L(X_1, \dots, X_n), U(X_1, \dots, X_n))$ 를  $\theta$ 에 대한  $100(1 - \alpha)\%$  **구간추정량**(interval estimator) 또는 **신뢰구간**(confidence interval)이라 한다. 주로

$$P(\theta \leq L(X_1, \dots, X_n)) = P(\theta \geq U(X_1, \dots, X_n)) = \alpha/2$$

를 만족하는  $L(X_1, \dots, X_n), U(X_1, \dots, X_n)$  를 사용한다.

**정리 10.6.**  $X_1, \dots, X_n \sim_{i.i.d} \mathcal{N}(\mu, \sigma^2)$  일 때, 모평균  $\mu$  에 대한  $100(1 - \alpha)\%$  신뢰구간은

$$\left( \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

**증명.**  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$  이므로  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$  이고,

$$P(Z > z_{\alpha/2}) = P(Z < -z_{\alpha/2}) = \alpha/2$$

이므로

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right) = P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

□

**정의 10.7.** 구간추정량의 관측값  $L(x_1, \dots, x_n), U(x_1, \dots, x_n)$  을 **구간추정값**(interval estimate) 또는 **신뢰구간**이라 한다.

**예제.** 정규모집단에서의 모평균  $\mu$ 에 대한  $100(1 - \alpha)\%$  신뢰구간은

$$\left( \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

**참고.**  $z_{0.05} = 1.645, z_{0.025} = 1.96, z_{0.005} = 2.756$  임을 알아두면 좋다.

**연습문제 10.8.** 미국임상영양학회지에 실린 한 기록에 의하면, 중앙아메리카의 원주민을 대상으로 49명의 표본조사를 한 결과 혈청 내의 콜레스테롤 양이 157mg/L 였다고 한다. 이들 원주민의 혈청 내 콜레스테롤 양이 표준편차가 30인 정규분포를 따를 때, 모평균  $\mu$ 에 대한 95% 신뢰구간을 구하여라.

**정리 10.9. 모평균  $\mu$ 에 대한 추정 ( $\sigma$ 를 알 때)**

- 가정:  $X_1, \dots, X_n \sim_{i.i.d} \mathcal{N}(\mu, \sigma^2)$ , 모표준편차  $\sigma$ 는 알려진 값
- 추정량과 추정값:  $\hat{\mu}(X_1, \dots, X_n) = \bar{X}$ ,  $\hat{\mu}(x_1, \dots, x_n) = \bar{x}$
- 표준오차:  $SE(\hat{\mu}) = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}}$
- $100(1 - \alpha)\%$  오차한계:  $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
- $100(1 - \alpha)\%$  신뢰구간:  $\left( \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$
- $100(1 - \alpha)\%$  신뢰구간의 길이:  $2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

모집단이 정규분포가 아닌 경우에는  $n$ 이 충분히 클 때 근사적으로 성립한다.

**연습문제 10.10.** 표준편차가 5인 모집단의 평균을 신뢰도 99%로 추정할 때, 모평균  $\mu$ 와 표본평균  $\bar{X}$ 의 차이가 0.5 이하가 되도록 하려면 적어도 몇 개의 표본을 조사해야 하는가?

**정리 10.11.**  $100(1 - \alpha)\%$  오차한계를  $d$  이하로 또는  $100(1 - \alpha)\%$  신뢰구간의 길이를  $2d$  이하로 하기 위한 최소 표본의 크기는  $n \geq \left( \frac{z_{\alpha/2} \cdot \sigma}{d} \right)^2$  인 최소의 정수이다.

**증명.** 연습문제로 남긴다. □

위