Introductory Statistics

Sungchan Yi

January 2019

Contents

1 자료의 생성

통계학(statistics)이란, 주어진 문제에 대하여 합리적인 답을 줄 수 있도록 숫자로 표시되는 정보를 **수집**하고 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 $\sim$ 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. N개의 추출단위로 구성된 유한모집단에서 n개의 추출단위를 비복원추출할 때, ${}_N\mathrm{C}_n$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 $x_1, \cdots, x_n$에 대하여

정의 2.1. (산술)**평균**(mean μ)은 변량의 총합을 변량의 개수로 나눈 값이다.

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

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

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

정리 2.3. $\sigma^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 - \mu^2$ (분산은 [변량^{2}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_i가 **최댓값**(maximum)이면 모든 j에 대해 $x_i \geq x_j$ 이고, x_i가 **최솟값**(minimum)이면 모든 j에 대해 $x_i \leq x_j$ 이다. 최댓값에서 최솟값을 뺀 값을 **범위**(range R)라 한다.

정의 2.10. Q_3에서 Q_1을 뺀 값을 **사분위수범위**(interquartile range IQR)로 정의한다.

$$IQR = Q_3 - Q_1$$

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

정의 2.11.

- **표본평균**(sample mean): $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **표본분산**(sample variance): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- **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.

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

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**: $\widehat{MAD} = \dfrac{1}{n}\sum_{i=1}^{n}\left|x_i - \widehat{Q_2}\right|$
- **표본의 범위**: $\widehat{R} = x_{(n)} - x_{(1)}$
- **표본의 사분위수범위**: $\widehat{IQR} = \widehat{Q_3} - \widehat{Q_1}$

3 순열과 조합

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

정의 3.2. 서로 다른 n개의 원소에서 서로 다른 r개를 택하여 일렬로 배열하는 것을 n개에서 r개를 택하는 **순열**(permutation)이라 하고, 기호로 ${}_n\mathrm{P}_r$ 와 같이 나타낸다.

정리 3.3. ${}_n\mathrm{P}_r = n(n-1)\cdots(n-r+1) = \dfrac{n!}{(n-r)!} \quad (단,\ 0 \leq r \leq n)$

정의 3.4. 서로 다른 n개의 원소에서 순서를 생각하지 않고 r개를 택하는 것을 n개에서 r개를 택하는 **조합**(combination)이라 하고, 기호로 ${}_n\mathrm{C}_r$ 또는 $\dbinom{n}{r}$ 과 같이 나타낸다.

정리 3.5. $\dbinom{n}{r} = \dfrac{{}_n\mathrm{P}_r}{r!} = \dfrac{n!}{r!(n-r)!} \quad (단,\ 0 \leq r \leq n)$

정리 3.6. (조합의 성질)

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

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

정의 3.7. 서로 다른 n개의 원소에서 중복을 허락하여 r개를 택하는 순열을 n개에서 r개를 택하는 **중복순열**이라 하고, 기호로 ${}_n\Pi_r$ 과 같이 나타낸다.

정의 3.8. 서로 다른 n개의 원소에서 중복을 허락하여 r개를 택하는 조합을 n개에서 r개를 택하는 **중복조합**이라 하고, 기호로 ${}_n\mathrm{H}_r$ 과 같이 나타낸다.

정리 3.9. ${}_n\Pi_r = n^r, \quad {}_n\mathrm{H}_r = \binom{n+r-1}{r}.$

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

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

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

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

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

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

정리 3.12. $n \in \mathbb{N}$ 에 대하여,

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

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

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

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

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

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

단, $g = \gcd(p_1, \ldots, p_k)$, $d > 0$ 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**($\varnothing$)이라 한다.

정의 4.3. 두 사건 A, B에 대하여, A 또는 B가 일어나는 사건을 A와 B의 **합사건**이라 하고, $A \cup B$ 로 나타낸다. 그리고 A와 B가 동시에 일어나는 사건을 A와 B의 **곱사건**이라 하고, $A \cap B$ 로 나타낸다.

정의 4.4. 표본공간 S의 부분집합인 두 사건 A, B에 대하여 $A \cap B = \varnothing$ 이면 A와 B는 서로 **배반사건**(disjoint)이라 한다. 또, 사건 A가 일어나지 않는 사건을 사건 A의 **여사건**이라 하고, A^C로 나타낸다.

정의 4.5. 표본공간 S의 공사건이 아닌 사건 $A_1, \ldots, A_n$이 다음 조건을 만족하면,

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

(2) $A_i \cap A_j = \varnothing$ $\big(\text{for all } 1 \le i \ne j \le n\big)$ (pairwise disjoint)

사건 $A_1, \ldots, A_n$을 S의 **분할**(partition)이라 한다.

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

정의 4.7. (수학적 확률) 어떤 시행의 표본공간 S가 m개의 근원사건으로 이루어져 있고, **각 근원사건이 일어날 가능성이 모두 같은 정도로 기대될 때**, 사건 A가 r개의 근원사건으로 이루어져 있으면 사건 A가 일어날 확률은 다음과 같다.

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

정의 4.8. (통계적 확률) 같은 시행을 n번 반복하여 사건 A가 일어난 횟수를 r_n이라고 하자. 이 때, 시행 횟수 n이 한없이 커짐에 따라 그 상대도수 r_n/n 은 $\mathrm{P}(A)$에 가까워진다.

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

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

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

정의 4.10. (확률의 공리 - Axioms of Probability) 표본공간 S와 사건 A에 대하여,

(1) $0 \le \mathrm{P}(A) \le 1$

(2) $\mathrm{P}(S) = 1$

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

정리 4.11. (확률의 기본 성질) 사건 A, B에 대하여 다음이 성립한다.

(1) $\mathrm{P}(\varnothing) = 0$

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

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

정리 4.12. 사건 A, B, C에 대하여 다음이 성립한다.

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

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

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

5 조건부확률

정의 5.1. 확률이 0이 아닌 두 사건 A, B에 대하여 사건 A가 일어났을 때, 사건 B가 일어날 확률을 사건 A가 일어났을 때의 사건 B의 **조건부확률**(conditional probability)이라 하고, 기호로 $\mathrm{P}(B \,|\, A)$ 와 같이 나타낸다. 이는 다음과 같이 계산한다.

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

정리 5.2. (확률의 곱셈정리) 공사건이 아닌 두 사건 A, B에 대하여 다음이 성립한다.

$$\mathrm{P}(A \cap B) = \mathrm{P}(B \,|\, A)\mathrm{P}(A) = \mathrm{P}(A \,|\, B)\mathrm{P}(B)$$

정의 5.3. 두 사건 A, B에 대하여 사건 A가 일어났을 때의 사건 B의 조건부확률이 사건 B가 일어날 확률과 같을 때, 즉

$$\mathrm{P}(B \,|\, A) = \mathrm{P}(B \,|\, A^C) = \mathrm{P}(B)$$

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

정리 5.4. 공사건이 아닌 두 사건 A, B에 대하여 다음 조건은 서로 동치이다.

(1) A, B가 서로 독립이다.
(2) $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$

정리 5.5. 공사건이 아닌 두 사건 A, B에 대하여 다음이 성립한다.

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

정의 5.6. 공사건이 아닌 사건 $A_1, \ldots, A_n$에 대하여,

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

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

정의 5.7. 공사건이 아닌 사건 $A_1, \ldots, A_n$에 대하여,

$$\mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathrm{P}(A_i)$$

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

정의 5.8. 동일한 시행을 반복할 때, 각 시행에서 일어나는 사건이 서로 독립이면 이러한 시행을 **독립시행**이라고 한다.

정리 5.9. (독립시행의 확률) 1회의 시행에서 사건 A가 일어날 확률을 p 라 할 때, 이 시행을 n회 반복하는 독립시행에서 사건 A가 r번 일어날 확률은 다음과 같다.

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

정리 5.10. (**전확률공식** - Law of Total Probability) 표본공간 S의 분할인 사건 $A_1, \ldots, A_n$에 대하여 다음이 성립한다.

$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(B \mid A_i)\, \mathrm{P}(A_i)$$

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

정리 5.11. (**베이즈 정리** - Bayes' Theorem) 사건 $A_1, \ldots, A_n$이 표본공간 S의 분할이고 $\mathrm{P}(B) > 0$ 이면 다음이 성립한다.

$$\mathrm{P}(A_k \mid B) = \frac{\mathrm{P}(B \mid A_k)\, \mathrm{P}(A_k)}{\displaystyle\sum_{i=1}^{n} \mathrm{P}(B \mid A_i)\, \mathrm{P}(A_i)}$$

증명. 조건부확률의 정의와 전확률공식으로부터 자명. □

6 확률변수와 확률분포