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. $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2$ $\big($분산은 [변량^{2}의 평균] - [평균의 제곱]$\big)$

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): $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$
- **표본분산**(sample variance): $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$
- **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**: $\widehat{MAD} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left| x_i - \widehat{Q_2} \right|$
- **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) $(1+x)^n = \displaystyle\sum_{r=0}^{n} \binom{n}{r} x^r = \binom{n}{0} + \binom{n}{1} x + \cdots + \binom{n}{n} x^n \quad$ (for all $x \in \mathbb{C}$)

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

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

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

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

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

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) $A_i \cap A_j = \varnothing \ \big(\text{for all } 1 \leq i \neq j \leq n\big)$ (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 $\mathrm{P}(B \mid A)$ 와 같이 나타낸다. 이는 다음과 같이 계산한다.

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

연습문제 5.2. 주사위 한 개를 던지는 시행에서 소수의 눈이 나오는 사건을 A, 홀수의 눈이 나오는 사건을 B라 할 때, 다음을 구하여라.

(1) $\mathrm{P}(A \cap B)$

(2) $\mathrm{P}(B \mid A)$

(3) $\mathrm{P}\big(A^C \mid B\big)$

정리 5.3. (확률의 곱셈정리) 공사건이 아닌 두 사건 A, B에 대하여 다음이 성립한다.

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

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, B에 대하여 사건 A가 일어났을 때의 사건 B의 조건부확률이 사건 B가 일어날 확률과 같을 때, 즉

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

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.

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, B에 대하여 다음 조건은 서로 동치이다.

(1) A, B가 서로 독립이다.

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

정리 5.8. 공사건이 아닌 두 사건 A, B에 대하여 다음이 성립한다.

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

연습문제 5.9. 다음을 증명하여라.

(1) 정리 5.8.

(2) 공사건이 아닌 두 사건 A, B가 서로 배반사건이면, 사건 A, B는 종속이다.

연습문제 5.10. 두 사건 A, B가 서로 독립이고 $\mathrm{P}(A) = 0.25, \mathrm{P}(B) = 0.4$ 일 때, 다음을 구하여라.

(1) $\mathrm{P}(A \cap B)$

(2) $\mathrm{P}(A^C \cap B)$

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

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

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

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

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

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

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

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

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

연습문제 5.14. 사건 A, B, C에 대하여 $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$ 이지만 A, B, C 가 상호 독립은 아닌 A, B, C의 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 A가 일어날 확률을 p 라 할 때, 이 시행을 n회 반복하는 독립시행에서 사건 A가 r번 일어날 확률은 다음과 같다.

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

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

(1) $\mathrm{P}(A)$를 구하여라.

(2) 주사위를 20번 던지는 시행에서 사건 A가 12번 일어날 확률을 구하여라.

(3) 주사위를 100번 던지는 시행에서 사건 A가 r번 일어날 확률을 구하여라.

정리 5.18. (**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. $\mathrm{P}(B) = \mathrm{P}\left(\bigcup_{i=1}^{n} (B \cap A_i)\right) = \sum_{i=1}^{n} \mathrm{P}(B \cap A_i) = \sum_{i=1}^{n} \mathrm{P}(B \mid A_i)\, \mathrm{P}(A_i)$ □

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 $A_1, \ldots, A_n$이 표본공간 S의 분할이고 $\mathrm{P}(B) > 0$ 이면 다음이 성립한다.

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

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 X의 값에 따라 확률이 어떻게 흩어져 있는지를 합이 1인 양수로써 나타낸 것을 X의 **확률분포**(probability distribution)라고 한다.

정의 6.2. 확률변수 X가 어떤 값 x를 취할 확률은 기호로 $\mathrm{P}(X = x)$, a 이상 b 이하의 값을 취할 확률은 기호로 $\mathrm{P}(a \leq X \leq b)$ 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 X라 할 때, 확률변수 X의 확률분포를 표로 나타내어라.

X		합계
$\mathrm{P}(X = x)$		1

연습문제 6.5. 검은 공 2개와 흰 공 4개가 들어 있는 주머니에서 동시에 3개의 공을 꺼낼 때 나오는 검은 공의 개수를 X라 하자. 확률변수 X의 확률분포를 표로 나타내어라.

정의 6.6. 이산확률변수 X가 취할 수 있는 값이 $x_1, \ldots, x_n$ 일 때, X의 각 값에 [X가 이 값을 취할 확률 $p_1, \ldots, p_n$] 을 대응시키는 함수

$$\mathrm{P}(X = x_i) = p_i \qquad (i = 1, \ldots, n)$$

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

(1) $0 \leq \mathrm{P}(X = x_i) \leq 1$

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

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

연습문제 6.7. 5개의 제비 중에 3개의 당첨 제비가 있다. 임의로 뽑은 2개의 제비 중에 있는 당첨 제비의 개수를 X라고 할 때, 확률변수 X의 확률질량함수를 구하여라.

연습문제 6.8. 주어진 이산확률변수 X에 대해 다음 값을 구하여라.

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

(1) $a + b$

(2) $\mathrm{P}(X = 1 \cup X = 3)$

(3) $\mathrm{P}(0 \le X \le 2)$

정의 6.9. 이산확률변수 X가 취하는 값이 $x_1, \ldots, x_n$일 때,

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

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

정의 6.11. 이산확률변수 X와 함수 $g(x)$에 대하여, 다음이 성립한다.

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

연습문제 6.12. 확률변수 X에 대하여 다음을 보여라. (단, a, b는 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 X에 대하여 $\mathbf{E}(5X+4), \mathbf{V}(5X-1), \sigma(5X+3)$ 을 각각 구하여라.

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

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

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

를 이산확률변수 X, Y의 **결합확률밀도함수**라 하며, 이 함수는 다음 조건을 만족해야 한다.

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

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

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

정의 6.16. 이산확률변수 X, Y의 결합확률밀도함수 $\mathrm{P}(X = x_i, Y = y_j)$에서 확률변수 X와 Y의 확률분포를 얻을 수 있다.

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

이를 각각 X, Y의 **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×100)원

이 때, A의 수입을 X, B의 수입을 Y라 하자. X, Y의 결합확률분포, 주변분포를 구하고, 확률변수 $Z = X + Y$로 정의할 때, Z의 확률분포와 $\mathbf{E}(Z)$를 구하여라.

$x \setminus y$				행의 합
열의 합				1

z						합계
$\mathrm{P}(Z = z)$						1

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

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

연습문제 6.19. 확률변수 X, Y에 대하여 $\mathbf{E}(aX+bY) = a\mathbf{E}(X)+b\mathbf{E}(Y)$ 임을 보여라. (단, a, b는 상수)

정의 6.20. 이산확률변수 X, Y가 주어져 있다. 임의의 $x_i \in X, y_j \in Y$에 대해 다음이 성립하면 확률변수 X, Y는 서로 **독립**이라 한다.

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

연습문제 6.21. 확률변수 X, Y가 서로 독립일 때, 다음이 성립함을 보여라.

(1) $\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$

(2) $\mathbf{V}(X \pm Y) = \mathbf{V}(X) + \mathbf{V}(Y)$

연습문제 6.22. 위 연습문제의 역은 성립하지 않음을 보여라.[1]

정리 6.23. (큰 수의 법칙) 어떤 시행에서 사건 A가 일어날 확률이 p일 때, n번의 독립시행에서 사건 A가 일어나는 횟수를 X라 하면, 임의의 양수 h에 대하여 다음이 성립한다.

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

[1] Hint: $-1, 1$을 각각 $1/2$의 확률로 취하는 확률변수 X에 대해 X와 X^2을 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, X가 성공 확률이 p인 베르누이 분포를 따를 때, $X \sim \mathrm{Berr}(p)$ 와 같이 나타낸다.[2]

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

정의 7.4. 한 번의 시행에서 사건 A가 일어날 확률이 p로 일정할 때, n번의 독립시행에서 사건 A가 일어나는 횟수를 X라고 하면 확률변수 X의 확률분포를 **이항분포**(binomial distribution)라 하고 기호로 $\mathrm{B}(n, p)$ 와 같이 나타낸다.

정의 7.5. 성공 확률이 p인 베르누이 시행을 n번 독립적으로 반복 시행할 때, 성공 횟수의 분포를 **이항분포**라 한다. 즉, $i = 1, \ldots, n$ 에 대하여 $X_i \sim_{i.i.d} \mathrm{Berr}(p)$일 때,[3] 이항분포는 n 개의 베르누이 확률변수의 합으로 정의된다.[4]

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

---

[2] X가 분포 $\mathcal{A}$ 를 따를 때, $X \sim \mathcal{A}$ 와 같이 표기한다.
[3] $i.i.d$: independent and identically distributed.
[4] 따라서 $\mathrm{Berr}(p) = \mathrm{B}(1, p)$ 이다.

정의 7.6. $X \sim \mathrm{B}(n,p)$일 때, X의 확률질량함수는 다음과 같다.

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

연습문제 7.7. 다음 확률변수 X가 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 X라 하자. $\mathrm{P}(X \le 9)$ 의 값을 구하여라.

연습문제 7.9. $X \sim \mathrm{B}(n,p)$ 이면, $\mathbf{E}(X) = np$, $\mathbf{V}(X) = np(1-p)$ 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 Y라 하자. $\mathbf{E}(X)$, $\mathbf{E}(Y)$ 를 구하여라.

연습문제 7.11. $X \sim \mathrm{B}(100,p)$ 라 하자. X의 분산이 최대일 때, $\mathbf{E}(X)$의 값을 구하여라.

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

$$\mathrm{P}(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad \left(\text{단, } \max\{0, n-N+D\} \le x \le \min\{n, D\}\right)$$

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

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

정리 7.14. $X \sim \mathrm{H}(N, D, n)$ 일 때, $N \gg n$ 이면 X는 근사적으로 $\mathrm{B}(n, D/N)$ 를 따른다.

정의 7.15. 성공 확률이 p인 베르누이 시행을 반복하여 최초로 성공할 때 까지의 시행 횟수를 X라 하자. 이 때, 확률변수 X는 **기하분포**(geometric distribution)을 따른다. 기하분포의 확률질량함수는 다음과 같이 주어진다.

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

연습문제 7.16. 성공 확률이 p인 기하분포를 따르는 확률변수 X에 대하여 다음이 성립함을 보여라.

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

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

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

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

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

정리 7.19. $X \sim \mathrm{B}(n, p)$ 일 때, n이 충분히 크면 $X \sim \mathrm{Poi}(np)$ 이다.[5]

연습문제 7.20. $X \sim \mathrm{Poi}(\lambda)$ 일 때, $\mathbf{E}(X) = \lambda$, $\mathbf{V}(X) = \lambda$ 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.

[5] 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 X가 구간 $[\alpha, \beta]$에 속하는 모든 실수 값을 취하고,

$$\mathrm{P}(a \le X \le b) = \int_a^b f(x)dx \quad (\alpha \le a \le x \le b \le \beta)$$

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

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

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

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

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

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

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

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

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

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

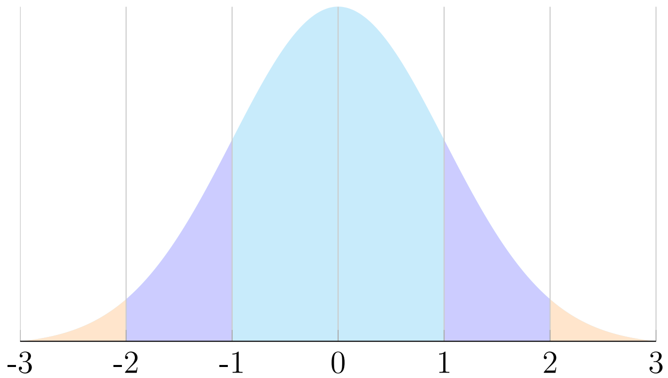
X의 확률분포를 **정규분포**(normal distribution)라 하고, 평균이 μ, 분산이 σ^2인 정규분포를 기호로 $\mathcal{N}(\mu, \sigma^2)$와 같이 나타낸다.

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.

표준정규분포 $\mathcal{N}(0, 1)$



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

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

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.

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

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

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

이로부터 다음이 성립함을 알 수 있다. 임의의 실수 a, b 에 대하여 $(a \le b)$

(1) $\mathrm{P}(Z \ge a) = \mathrm{P}(Z \le -a)$.

(2) $\mathrm{P}(Z \ge a) = 1 - \mathrm{P}(Z < a)$

(3) $\mathrm{P}(a \le Z \le b) = \mathrm{P}(Z \le b) - \mathrm{P}(Z < a)$

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

z	$\mathrm{P}(0 \le Z \le z)$
0.5	0.1915
1.0	0.3413
1.5	0.4332
2.0	0.4772

(1) $\mathrm{P}(X \le 21)$

(2) $\mathrm{P}(29 \le X \le 35)$

(3) $\mathrm{P}(X \ge 25)$

정의 8.10. $\mathcal{N}(0,1)$ 의 $100(1-\alpha)$ 백분위수를 z_α 로 나타낸다. 즉 $\mathrm{P}(Z \ge z_\alpha) = \alpha$ 이다.[6]

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

[6] **상방백분위수**라고도 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. $\mathrm{P}\left(|X - \mu| < \sigma\right) = 0.6826$.
- 95.44% 의 관측값들이 $[\mu - 2\sigma, \mu + 2\sigma]$ 에 있다. $\mathrm{P}\left(|X - \mu| < 2\sigma\right) = 0.9544$.
- 99.74% 의 관측값들이 $[\mu - 3\sigma, \mu + 3\sigma]$ 에 있다. $\mathrm{P}\left(|X - \mu| < 3\sigma\right) = 0.9974$.

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	$\mathrm{P}(0 \le Z \le z)$
0.25	0.1
0.52	0.2
1.28	0.4

연습문제 8.15. 다음은 9등급제의 산출 방식 중 일부이다.

등급	z	$\mathrm{P}(Z \ge z)$
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**) $X \sim \mathrm{B}(n,p)$ 일 때, n이 충분히 크면[7] X는 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. $X \sim \mathrm{B}(n,p)$ 일 때,

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

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

[7] 일반적으로, $np \ge 5, n(1-p) \ge 5$ 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 n인 표본을 $X_1, \ldots, X_n$ 이라 할 때,

- **표본평균**(sample mean): $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **표본분산**: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

표본을 뽑을 때마다 표본평균, 표본분산은 달라질 수 있으므로 $\overline{X}, S^2$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 $\overline{X}$의 확률분포가 다음과 같다.

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

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

정의 9.4. X_i 를 i-번째로 뽑힌 추출단위의 특성값을 나타내는 확률변수라 하자. 다음 조건을 만족하는 $X_1, \ldots, X_n$ 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 μ이고 모표준편차가 σ인 모집단에서[8] 복원추출하여 뽑은 크기가 n인 표본의 표본평균 $\overline{X}$에 대하여 다음이 성립한다.

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

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

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

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

□

참고: 모평균이 μ, 모분산이 σ^2, 크기가 N인 유한모집단에서 크기 n인 표본을 비복원추출하는 경우에는 다음이 성립한다.

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

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

연습문제 9.6. 연습문제 9.3 에서 구한 $\overline{X}$ 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 $\overline{X}$는 $\mathcal{N}(\mu, \sigma^2/n)$ 을 따른다.

[8]무한모집단의 경우 복원추출이나 비복원추출이나 큰 차이가 없다.

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

연습문제 9.10. 어느 전기 회사에서 생산하는 전구의 수명을 나타내는 확률변수 X에 대하여, $X \sim \mathcal{N}(2000, 200^2)$ 이라고 한다. 이 회사가 생산한 전구 중 임의추출한 n개 전구의 평균 수명을 $\overline{X}$라 할 때, $\mathrm{P}(1900 \le \overline{X} \le 2100) \ge 0.9$ 가 성립하기 위한 n의 최솟값을 구하여라. 단, $z_{0.05} = 1.65$ 이다.

정리 9.11. (**중심극한정리** - Central Limit Theorem) 평균이 μ이고 분산이 σ^2인 임의의 무한모집단에서 표본의 크기 n이 충분히 크면, 랜덤표본의 표본평균 $\overline{X}$는 근사적으로 정규분포 $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ 을 따른다.[9]

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 $Z_1, \ldots, Z_k$ 이 $\mathcal{N}(0,1)$의 랜덤표본일 때, $V = Z_1^2 + \cdots + Z_k^2$ 의 분포를 **자유도**(degrees of freedom) k인 **카이제곱분포**(χ^2-distribution)라고 한다. 기호로는 다음과 같이 나타낸다.

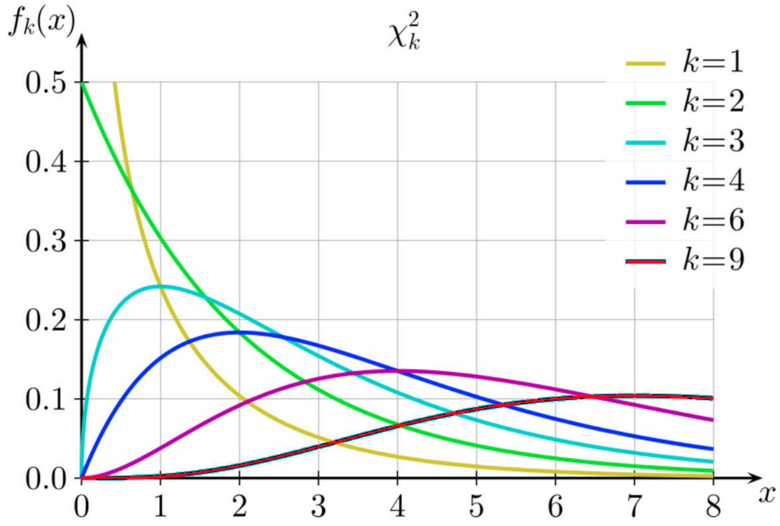
$$Z_1^2 + \cdots + Z_k^2 = V \sim \chi^2(k)$$

그리고 확률밀도함수는 다음과 같다.[10]

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

[9]당연히, n이 클수록 근사는 정확해진다.

[10]Γ 는 감마함수(Gamma function)을 나타내는 기호이다.



χ^2-분포의 형태 (k: 자유도)

정의 9.14. $V \sim \chi^2(k)$ 일 때, $\mathrm{P}(V > v) = \alpha$ 인 v의 값을 $\chi^2_\alpha(k)$ 로 정의한다.

정리 9.15. (카이제곱분포의 가법성) V_1, V_2가 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. $X_1, \ldots, X_n$이 $\mathcal{N}(\mu, \sigma^2)$의 랜덤표본일 때, 표본분산 $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$ 에 대하여 다음이 성립한다.

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

증명. 모든 i에 대해 $\dfrac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$ 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^2의 정의로부터

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

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

정리 9.17. 분산이 동일한 두 정규모집단 $\mathcal{N}(\mu_1, \sigma^2)$, $\mathcal{N}(\mu_2, \sigma^2)$ 에서 각각 뽑은 랜덤표본 $X_1, \ldots, X_{n_1}$ 과 $Y_1, \ldots, Y_{n_2}$ 이 서로 독립이라고 하자. S_1^2, S_2^2를 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.

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

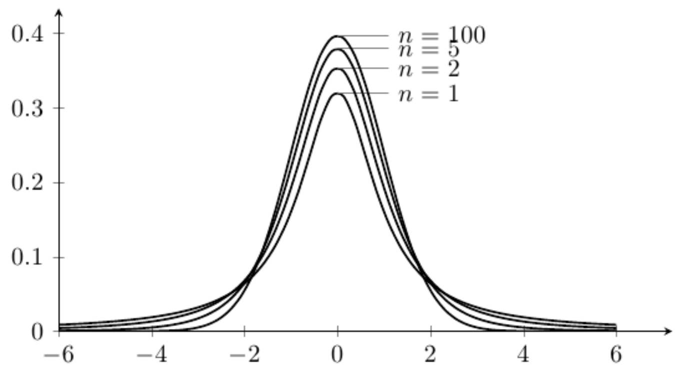
□

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

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

그리고 확률밀도함수는 다음과 같다.

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



t-분포의 형태 (n: 자유도)

[11] 스튜던트의(Student's) t-분포 라고 불리기도 한다.

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

정의 9.20. $T \sim t(k)$ 일 때, $\mathrm{P}(T > t) = \alpha$ 인 t의 값을 $t_\alpha(k)$ 로 정의한다.

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

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

증명. 다음 변형을 이용한다.

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

$\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ 이고, $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 이며, $\overline{X}$ 와 S^2은 서로 독립임이 알려져 있다.[12] t-분포의 정의에 의해 성립한다. □

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

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

증명. 다음 변형을 이용한다.

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

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

[12] 일반적으로는 독립이 아니다. 이 정리의 가정 하에서만 독립이다. See Basu's Theorem.

정의 9.23. $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$ 이고 V_1, V_2가 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.[13]

$$\frac{\sqrt{\dfrac{(k_1x)^{k_1}k_2^{k_2}}{(k_1x+k_2)^{k_1+k_2}}}}{x\mathrm{B}\left(\dfrac{k_1}{2}, \dfrac{k_2}{2}\right)} \qquad (x > 0)$$

정의 9.24. $F \sim F(k_1, k_2)$ 일 때, $\mathrm{P}(F > f) = \alpha$ 인 f의 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 $X_1, \ldots, X_{n_1}$ 과 $Y_1, \ldots, Y_{n_2}$ 이 서로 독립이라고 하자. X_i, Y_i 의 표본분산 S_1^2, S_2^2에 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.

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

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

[13] 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) θ: 모집단의 특성을 나타내는 수치적 측도
- 랜덤표본은 $X_1, \ldots, X_n$, 랜덤표본의 관측값은 $x_1, \ldots, x_n$ 으로 표기한다.
- **추정량**(estimator): 미지의 모수 θ의 추정에 사용되는 통계량으로, $\hat{\theta}(X_1, \ldots, X_n)$ 혹은 $\hat{\theta}$으로 표기한다.
- **추정값**(estimate): 추정량 $\hat{\theta}(X_1, \ldots, X_n)$의 관측값으로, $\hat{\theta}(x_1, \ldots, x_n)$로 표기한다.

추정량과 추정값의 관계는 확률변수와 그 관측값의 관계이다.
모평균 μ를 추정하기 위해 추정량으로는 표본평균 $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 을 사용하고, 그 추정값으로는 관측한 결과인 $\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 를 사용한다.

정의 10.3. $\mathbf{E}(\hat{\theta}) = \theta$ 를 만족하는 추정량 $\hat{\theta}$를 **불편추정량**(unbiased estimator)이라 한다.
추정량 $\hat{\mu} = \overline{X}$ 의 경우 $\mathbf{E}(\overline{X}) = \mu$ 이므로 불편추정량이다.[14]

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

[14]표본분산 $\widehat{\sigma^2} = S^2$ 도 불편추정량이다. 사실 불편추정량이 되도록 $n-1$ 로 나눈 것이다...

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

$$\mathrm{P}(L(X_1, \ldots, X_n) < \theta < U(X_1, \ldots, X_n)) = 1 - \alpha$$

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

$$\mathrm{P}(\theta \le L(X_1, \ldots, X_n)) = \mathrm{P}(\theta \ge U(X_1, \ldots, X_n)) = \alpha/2$$

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

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

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

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

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

이므로

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

□

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

예제. 정규모집단에서의 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간은

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

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 μ에 대한 95% 신뢰구간을 구하여라.

정리 10.9. 모평균 μ에 대한 추정 <u>(σ를 알 때)</u>

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

모집단이 정규분포가 아닌 경우에는 n이 충분히 클 때 근사적으로 성립한다.

연습문제 10.10. 표준편차가 5인 모집단의 평균을 신뢰도 99%로 추정할 때, 모평균 μ와 표본평균 $\overline{X}$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 $n \geq \left(\dfrac{z_{\alpha/2} \cdot \sigma}{d}\right)^2$ 인 최소의 정수이다.
증명. 연습문제로 남긴다. □

이제 모표준편차 σ를 모를 때 모평균 μ를 추정하는 방법을 살펴보자. 이 경우에는 σ를 그 추정량인 표본표준편차 S로 대신해야 한다. 이렇게 표본표준편차를 이용해 표준화하는 과정을 **스튜던트화**(Studentize)라고 한다. 9장에서 t-분포를 배우면서 표본평균을 표본표준편차를 이용해 표준화하면 자유도가 $n-1$인 t-분포를 따름을 배웠다.

정리 10.12. 모평균 μ에 대한 추정 (σ를 모를 때), 표본표준편차 S

- 가정: $X_1, \ldots, X_n \sim_{i.i.d} \mathcal{N}(\mu, \sigma^2)$, 모표준편차 σ는 미지의 값
- 추정량과 추정값: $\hat{\mu}(X_1, \ldots, X_n) = \overline{X}$, $\hat{\mu}(x_1, \ldots, x_n) = \overline{x}$
- $100(1-\alpha)\%$ 오차한계: $t_{\alpha/2}(n-1) \cdot \dfrac{S}{\sqrt{n}}$
- $100(1-\alpha)\%$ 신뢰구간: $\left(\overline{X} - t_{\alpha/2}(n-1) \cdot \dfrac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2}(n-1) \cdot \dfrac{S}{\sqrt{n}}\right)$
- $100(1-\alpha)\%$ 신뢰구간의 길이: $2 \cdot t_{\alpha/2}(n-1) \cdot \dfrac{S}{\sqrt{n}}$

모집단이 정규분포가 아닌 경우에는 n이 충분히 클 때 근사적으로 성립한다.

증명. 신뢰구간만 증명한다.
$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ 이므로 $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$ 이고,

$$\mathrm{P}(T > t_{\alpha/2}(n-1)) = \mathrm{P}(T < -t_{\alpha/2}(n-1)) = \alpha/2$$

이므로

$$\begin{aligned} 1 - \alpha &= \mathrm{P}\left(\left|\frac{\overline{X} - \mu}{S/\sqrt{n}}\right| \le t_{\alpha/2}(n-1)\right) \\ &= \mathrm{P}\left(\overline{X} - t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}\right) \end{aligned}$$

□

주의. 위 방법들은 모집단이 정규분포를 따를 때 사용할 수 있다. 따라서 표본으로 자료가 주어진 경우에는 **정규분포 분위수 대조도**(normal distribution quantile-quantile plot)을 통해 자료분포의 분위수와 정규분포의 분위수를 비교하여 모집단의 분포가 정규분포라는 가정을 검토해야 한다.

참고. t-분포표에서 자유도가 큰 경우에는 모든 자유도에 대해 확률 값이 계산되어 있지 않다. 이 경우에는 표의 자유도 중 실제 자유도보다 더 작으면서 가장 가까운 값을 사용한다.

연습문제 10.13. 정규분포를 따르는 모집단으로부터 64개의 자료를 관측한 결과 표본평균이 27, 표본표준편차가 5 였다. 모평균 μ에 대한 99% 신뢰구간을 구하여라.

통계적 추론을 할 때는 **통계적 가설**(statistical hypothesis)을 세워 모수에 대해 예상하거나 추측을 하고, 이를 **검정**한다.

정의 10.14. 자료로부터의 강력한 증거에 의해 입증하고자 하는 가설을 **대립가설**(alternative hypothesis)이라 하고 H_1 으로 나타낸다. 그리고 반대 증거를 찾기 위해 상정된 가설을 **귀무가설**(null hypothesis)이라 하고, H_0 으로 나타내며, 대립가설에 반대되는 가설이다. 두 가설 중 어느 가설을 택할지 통계적으로 결정하는 과정을 **가설 검정**(hypothesis test)이라 한다.

예제. 어느 전구 공장의 기존 공정에서 전구의 수명이 평균 1200분, 표준편차 100분인 것으로 알려져 있다. 새로운 공정에 대하여 25개의 표본을 조사한 결과 표본평균이 1240분이었다. 새로운 공정으로 생산한 전구의 평균 수명 μ에 대해 조사하려 한다.

(1) 새로운 공법과 기존의 공법의 평균 수명이 다르다고 할 수 있는가?
조사의 목적이 평균 수명이 달라졌다는 증거가 있는지 알아보기 위함이므로,

$$H_0\colon \mu = 1200 \text{ (평균 수명이 같다)} \qquad H_1\colon \mu \neq 1200 \text{ (평균 수명이 다르다)}$$

(2) 새로운 공법이 전구의 평균수명을 증가시켰다고 할 수 있는가?

$$H_0\colon \mu = 1200 \text{ (평균 수명이 같다)} \qquad H_1\colon \mu > 1200 \text{ (평균 수명이 증가했다)}$$

정의 10.15. 비교하는 값의 양쪽을 뜻하는 가설을 **양측가설**(two-sided hypothesis)이라 하고, 한쪽을 뜻하는 가설을 **단측가설**(one-sided hypothesis)이라 한다.

정의 10.16. 귀무가설에 대한 반증의 강도를 제공하는 과정을 **유의성검증**(test of significance)이라 한다. 이 과정에서 사용되는 통계량을 **검정통계량**(test statistic)이라 한다.

귀무가설과 대립가설은 모수에 관한 가설로 주어지므로, 유의성검증에서 찾고자 하는 증거는 그 모수의 추정량을 이용하여 찾게 된다. 따라서, 귀무가설에 반대되며 대립가설을 지지하는 증거는 모수의 추정값이 귀무가설 H_0에서 주어지는 모수의 값으로부터 대립가설 H_1의 방향으로 멀리 떨어질수록 강해진다.

위 예제의 (2)에서 귀무가설과 대립가설이 각각 모평균에 관한 가설이므로, 표본평균 $\overline{X}$를 이용하여 유의성검증에서의 증거를 찾게 된다. 그렇다면 가설에 대한 반증의 강도는 어떻게 측정할지 의문이 들기 마련이다. 다음과 고려해 보자. **귀무가설 H_0가 사실일 때, 실제 관측값보다 더욱 H_0에 반대되며, H_1을 지지하는 조사 결과를 얻을 확률은 얼마인가?**

위 예제에서 실제 관측값보다 더욱 H_0에 반대되며 H_1을 지지하는 조사의 결과는 $\overline{X} \geq 1240$이다. 이에 대한 확률을 계산해 보면

$$\begin{aligned} \mathrm{P}\left(\overline{X} \geq 1240 \,|\, H_0 \text{ 가 참}\right) &= \mathrm{P}\left(\frac{\overline{X}-1200}{100/\sqrt{25}} \geq \frac{1240-1200}{100/\sqrt{25}} \,\middle|\, \mu = 1200\right) \\ &= \mathrm{P}(Z \geq 2) = 0.0228 \end{aligned}$$

이다. 따라서, 무한히 같은 조사를 해도, 실제 조사 결과 $\overline{x} = 1240$ 보다 더욱 강한 H_0의 반증을 얻을 기회는 2.28% 이므로, 위 확률은 이 관측값이 얼마나 일어나기 어려운 것인지 나타내 준다. 즉, 귀무가설 H_0가 사실이 아님을 강하게 시사한다고 할 수 있다.

정의 10.17. 검정통계량이 실제 관측된 값보다 대립가설을 지지하는 방향으로 더욱 치우칠 확률로서 귀무가설 H_0하에서 계산된 값을 **유의확률**(significance probability) 또는 P-값(p-value)라 한다. 유의확률이 작을수록 H_0에 대한 반증이 강한 것을 뜻하며, 유의확률을 계산할 때는 H_0 하에서 검정통계량의 분포를 이용한다.

반증이 "강하다"는 상대적인 표현이므로, 가설검정을 할 때, 미리 기준값을 정해두고 유의확률을 그 기준값과 비교한다.

정의 10.18. 귀무가설 H_0에 대한 반증의 강도에 대하여 미리 정해둔 기준값을 **유의수준**(significance level)이라 부르고, 흔히 α로 나타낸다.

유의수준으로는 주로 $\alpha = 0.1,\ 0.05,\ 0.01$ 을 사용하며, 유의수준이 α인 것은 [귀무가설에 대한 반증이 조사 결과보다 강하게 나타날 확률]이 α 이하일 것을 요구하는 것이다. 유의확률이 지정된 유의수준 α 이하로 나타나면, **유의수준 α 에서 유의하다**(statistically significant)고 하며, 이는 귀무가설에 대한 반증의 강도가 지정된 수준보다 강함을 의미한다. 따라서 귀무가설을 **기각**(reject)하고 대립가설을 채택한다.

정의 10.19. 가설검정에서 오류의 종류

검정결과\실제상황	H_0가 참	H_1이 참
H_0 채택	옳은 결정	**제 2종의 오류**(Type II Error β)
H_0 기각	**제 1종의 오류**(Type I Error α)	옳은 결정

유의수준을 α로 지정한다는 것은 제 1종의 오류를 범할 확률의 허용한계를 α로 미리 정해 주는 것이다.

정의 10.20. 정해진 유의수준에 따라, [귀무가설을 기각하게 되는 검정통계량의 관측값]의 범위를 **기각역**(rejection region)이라 한다.[15]

가설검정의 절차

- 귀무가설과 대립가설을 설정한다. 유의성검증은 **귀무가설에 대한 반증의 강도**를 알아보기 위함임을 고려한다.
- 유의수준 α를 지정한다. 귀무가설에 대한 반증의 강도가 어느 정도이어야 하는지 고려한다.
- 자료로부터 검정통계량의 관측된 값을 계산한다. 검정통계량은 가설에 관계되는 모수의 추정량을 이용한다.
- (**유의확률** 사용) 검정통계량의 관측값으로부터 유의확률을 계산하여 유의수준보다 작으면 H_0를 기각한다. 그렇지 않으면 H_0를 채택한다.
- (**기각역** 사용) 유의수준에 대한 기각역을 찾아 검정통계량의 관측값이 기각역에 속하면 H_0를 기각한다. 그렇지 않으면 H_0를 채택한다.

정리 10.21. 모평균 μ에 대한 유의성검증 (σ를 알 때) - Z 검정

- 가정: $X_1, \ldots, X_n \sim_{i.i.d} \mathcal{N}(\mu, \sigma^2)$, 모표준편차 σ는 알려진 값. 혹은 모집단이 정규분포는 아니지만 n이 충분히 큰 경우 (중심극한정리)
- 귀무가설 H_0: $\mu = \mu_0$

[15]기각역과 기각역이 아닌 영역을 나누는 기준이 되는 경계값을 critical value 라고도 한다.

- **검정통계량**: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim_{H_0} \mathcal{N}(0,1)$, **관측값**: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$
- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu > \mu_0$	$z \geq z_\alpha$	$\mathrm{P}(Z \geq z)$
H_1: $\mu < \mu_0$	$z \leq -z_\alpha$	$\mathrm{P}(Z \leq z)$
H_1: $\mu \neq \mu_0$	$\lvert z\rvert \geq z_{\alpha/2}$	$\mathrm{P}(\lvert Z\rvert \geq \lvert z\rvert)$

연습문제 10.22. 한 제약 회사의 연구개발부에서 특정 약품의 주성분의 농도가 평균이 μ 이고 표준편차가 0.0123 인 정규분포를 따른다고 한다. 이 연구개발부에서는 이론적으로 주성분의 농도가 0.85 일 것이라고 예상하는 제조법을 창안하였다. 제조와 측정의 기술적, 비용적 측면을 고려하여 3번 반복 측정한 결과 주성분 농도의 평균이 $\overline{x} = 0.8404$ 이었다. 이 제조법에 의한 주성분의 실제 농도 μ가 0.85 가 아니라고 의심할 만한 증거가 있는가? 유의수준 0.05에서 검정하여라.

정리 10.23. 모평균 μ에 대한 유의성검증 (<u>σ를 모를 때</u>) - t 검정

- 가정: $X_1, \ldots, X_n \sim_{i.i.d} \mathcal{N}(\mu, \sigma^2)$, 모표준편차 σ는 미지의 값. 혹은 모집단이 정규분포는 아니지만 n이 충분히 큰 경우 (중심극한정리)
- 귀무가설 H_0: $\mu = \mu_0$
- **검정통계량**: $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim_{H_0} t(n-1)$, **관측값**: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ (s: 표본표준편차)
- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu > \mu_0$	$t \geq t_\alpha(n-1)$	$\mathrm{P}(T \geq t)$
H_1: $\mu < \mu_0$	$t \leq -t_\alpha(n-1)$	$\mathrm{P}(T \leq t)$
H_1: $\mu \neq \mu_0$	$\lvert t\rvert \geq t_{\alpha/2}(n-1)$	$\mathrm{P}(\lvert T\rvert \geq \lvert t\rvert)$

연습문제 10.24. 전구를 생산하는 회사에서 현재 생산하는 전구의 수명은 평균이 1950 시간인 정규분포를 따른다고 알려져 있다. 새롭게 개발중인 전구의 평균수명 μ가 기존의 전구보다 수명이 더 길다고 할 수 있는지 확인하기 위해 9개의 시제품을 생산하여 그 수명 시간을 조사한 결과가 다음과 같다.

$$2000,\ 1975,\ 1900,\ 2000,\ 1950,\ 1850,\ 1950,\ 2100,\ 1975$$

적절한 가설을 세우고, 유의수준 5%에서 검정하여라.

정의 10.25. 제 2종의 오류를 범하지 않을 확률을 **검정력**(power)이라 한다. 제 2종의 오류를 범할 확률을 주로 β로 두며, 이때 검정력은 $1-\beta$가 된다.

연습문제 10.26. 연습문제 10.22 에서 실제로 $\mu = 0.84$ 일 때, 검정력을 구하여라.

정리 10.27. 표준편차를 알고있는 정규모집단 $\mathcal{N}(\mu, \sigma)$ 에서의 표본에 대해, 유의수준 α인 H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$ 의 검정에서 실제로 $\mu = \mu_1(> \mu_0)$ 일 때, 제 2종의 오류를 범할 확률이 β 이하가 되기 위한 표본의 크기 n은 다음을 만족하는 최소의 정수이다.

$$n \geq \left(\frac{z_\alpha + z_\beta}{(\mu_1 - \mu_0)/\sigma} \right)^2$$

증명. 제 2종의 오류를 범할 확률이 β 이하가 되어야 하므로,

$$\begin{aligned} \mathrm{P}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \,\middle|\, \mu = \mu_1 \right) &= \mathrm{P}\left(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \,\middle|\, \mu = \mu_1 \right) \\ &= \mathrm{P}\left(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} < z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, \mu = \mu_1 \right) < \beta \end{aligned}$$

가 성립해야 한다. $\dfrac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ 이므로,

$$-z_\beta \geq z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$$

이 성립하고, 이를 n에 대해 정리하면 원하는 결론을 얻는다. $\square$

연습문제 10.28. 한 제약회사에서 생산하고 있는 기존의 진통제는 진통 효과가 나타나는 시간이 평균 30분, 표준편차 5분인 것으로 알려져 있다. 회사의 연구소에서 진통 효과가 더 빨리 나타날 것으로 기대되는 새로운 진통제를 개발하였다. 새로운 진통제의 진통효과가 더 빠른가를 확인하기 위하여, 50명의 환자를 랜덤추출하여 새로운 진통제에 의해 그 효과가 나타나는 시간을 측정한 결과 평균이 28.5분이었다. 새로운 진통제에 의한 진통 효과가 나타나는 시간이 표준편차 5분인 정규분포를 따른다고 하자. 다음 물음에 답하여라.

(1) 적절한 가설을 유의수준 5%에서 검정하여라.

(2) 유의수준 5%에서 검정할 때, 제 1종의 오류를 범할 확률을 구하여라.

(3) 실제로 $\mu = 28$ 일 때, 제 2종의 오류를 범할 확률이 $\beta = 0.10$ 이하가 되도록 하며, 유의수준 5%인 검정법을 사용하려고 한다. 이 때, 요구되는 표본의 최소 크기를 구하여라.

11 분포에 대한 추론

통계적 추론을 필요로 하는 많은 문제들은 하나의 모집단에 관한 것이라기 보다는 여러 모집단을 비교하기 위한 경우가 더 많다.

예제. 두 종류의 진통제에 대한 상대적 효과의 척도로서, 복용 후 숙면할 수 있는 정도를 비교하려고 한다. 비교 실험에 참여하는 환자 6명을 랜덤추출 했으나 각 환자들의 건강상태에는 상당한 차이가 있으므로, 각 환자에게 두 종류의 진통제를 각각 1회씩 복용하게 하여 숙면시간의 차이를 이용하여 두 진통제의 효과를 비교하기로 하였다.

환자	1	2	3	4	5	6
진통제 A	4.8	4.0	5.8	4.9	5.3	7.4
진통제 B	4.0	4.2	5.2	4.9	5.6	7.1

두 진통제의 효과에 차이가 있는가?

정의 11.1.

- **실험 단위**(experimental unit): 비교의 목적을 위해 그 매개체로 사용되는 대상
- **처리**(treatment): 실험 단위에 적용되어 특성치를 결정지어주는 것
- **처리 효과**(treatment effect): 비교 대상인 특성치
- **대응비교** 또는 **쌍체비교**(paired comparison): 두 모집단의 평균을 비교할 때 실험 단위를 동질적인 쌍으로 묶은 다음, 각 쌍에 두 처리를 임의로 적용하고, 각 쌍에서 모은 관측값의 차로 처리효과의 차에 관한 추론을 하는 방법

연습문제 11.2. 위 **예제**의 상황에 사용되는 방법이 대응비교이다. 예제에서 실험단위, 처리, 처리효과를 찾아보아라.

정리 11.3. 대응비교의 자료구조 및 모형

- 자료구조: 랜덤표본 $(X_1, Y_1), \ldots, (X_n, Y_n)$

쌍	처리 1	처리 2	처리효과의 차
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n	$D_n = X_n - Y_n$

- 처리효과: $\mathbf{E}(X_i) = \mu_1$, $\mathbf{E}(Y_i) = \mu_2$
- 처리효과의 차: $\delta = \mu_1 - \mu_2$
- 차에 대한 가정: $D_i = X_i - Y_i \sim_{i.i.d} \mathcal{N}(\delta, \sigma_D^2)$, $i = 1, \ldots, n$. (σ_D 는 미지의 값)
- $\mu_1 - \mu_2$ 에 대한 추정량: $\widehat{\mu_1 - \mu_2} = \hat{\delta} = \overline{D} = \dfrac{1}{n}\sum_{i=1}^{n} D_i$
- $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\left(\overline{d} - t_{\alpha/2}(n-1) \cdot \frac{s_D}{\sqrt{n}},\ \overline{d} + t_{\alpha/2}(n-1) \cdot \frac{s_D}{\sqrt{n}}\right)$$

단, $s_D^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$.

참고. $D_i \sim_{i.i.d} \mathcal{N}(\delta, \sigma_D^2)$ 이므로 $\overline{D} \sim \mathcal{N}(\delta, \sigma_D^2/n)$ 이다. 모표준편차를 알지 못하므로, 추정을 위해 표본표준편차를 사용하며, 분포는 t-분포를 사용한다. 표준화하면,

$$T = \frac{\overline{D} - \delta}{S_D/\sqrt{n}} \sim t(n-1)$$

가 됨을 알 수 있다. 이로부터 신뢰구간을 구하면 된다.

연습문제 11.4. 위 **예제**에서 주어진 자료를 이용하여 두 진통제 A, B에 의한 평균 숙면시간의 차이에 대한 95% 신뢰구간을 구하여라.

정리 11.5. 대응비교에 의한 두 모평균의 비교 - t 검정

- 가정: 대응비교의 자료구조에서의 가정과 동일
- 귀무가설 H_0: $\mu_1 - \mu_2 = \delta_0$
- **검정통계량**: $T = \dfrac{\overline{D} - \delta_0}{S_D/\sqrt{n}} \sim_{H_0} t(n-1)$, **관측값**: $t = \dfrac{\overline{d} - \delta_0}{s_D/\sqrt{n}}$ (s_D: 표본표준편차)
- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu_1 - \mu_2 > \delta_0$	$t \geq t_\alpha(n-1)$	$\mathrm{P}(T \geq t)$
H_1: $\mu_1 - \mu_2 < \delta_0$	$t \leq -t_\alpha(n-1)$	$\mathrm{P}(T \leq t)$
H_1: $\mu_1 - \mu_2 \neq \delta_0$	$\lvert t\rvert \geq t_{\alpha/2}(n-1)$	$\mathrm{P}(\lvert T\rvert \geq \lvert t\rvert)$

연습문제 11.6. 예제에서 주어진 자료를 활용하여 두 진통제의 효과에 차이가 있는지 유의수준 1%에서 검정하여라. 그리고 유의확률도 구하여라.

두 가지 처리를 비교하는 경우에는 대응비교와는 달리 실험 단위를 두 그룹으로 나누어 그룹 별로 서로 다른 처리를 적용하여 그 결과를 이용하여 두 처리를 비교할 수도 있다. 한 실험 단위에 두 처리를 모두 적용하기 어려운 경우에는 대응비교를 사용할 수 없다.

정리 11.7. 이표본에 의한 모평균의 비교 - 가정

- 모형 가정: 각 그룹에서의 관측값은 각 모집단에서의 랜덤 표본이고, 서로 다른 그룹에서의 관측값들은 독립적으로 관측된 것이다.
- 분포 가정: $X_1, \ldots, X_{n_1}$ 와 $Y_1, \ldots, Y_{n_2}$ 는 각각 $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$ 의 랜덤표본

정리 11.8. $\mu_1 - \mu_2$ 에 대한 추정

- 추정량: $\widehat{\mu_1 - \mu_2} = \widehat{\mu}_1 - \widehat{\mu}_2 = \overline{X} - \overline{Y}$
- 추정값: $\widehat{\mu_1 - \mu_2} = \overline{x} - \overline{y}$
- 표본분포: $\overline{X} \sim \mathcal{N}\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right), \overline{Y} \sim \mathcal{N}\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right), \overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$

참고. 모형 가정에 의해 표본이 서로 독립이므로,

$$\mathbf{V}(\overline{X}-\overline{Y})=\mathbf{V}(\overline{X})+\mathbf{V}(\overline{Y})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

따라서 표준화하면,

$$\frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim\mathcal{N}(0,1)$$

정리 11.9. 이표본에 의한 모평균의 비교 (σ_1, σ_2 **를 아는 경우**)

표본의 크기 n_1, n_2 가 충분히 큰 경우, 정규 모집단 가정이 생략 가능 (중심극한정리)

- $\mu_1-\mu_2$ 에 대한 $100(1-\alpha)\%$ 신뢰구간: $(\overline{x}-\overline{y})\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
- H_0: $\mu_1-\mu_2=\delta_0$ 에 대한 검정
  - 검정통계량: $Z=\dfrac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim_{H_0}\mathcal{N}(0,1)$
  - 관측값: $z=\dfrac{\overline{x}-\overline{y}-\delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$
- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu_1-\mu_2>\delta_0$	$z\geq z_\alpha$	$\mathrm{P}\left(Z\geq z\right)$
H_1: $\mu_1-\mu_2<\delta_0$	$z\leq -z_\alpha$	$\mathrm{P}\left(Z\leq z\right)$
H_1: $\mu_1-\mu_2\neq\delta_0$	$\lvert z\rvert\geq z_{\alpha/2}$	$\mathrm{P}\left(\lvert Z\rvert\geq\lvert z\rvert\right)$

그러나 실제 문제에서는 모표준편차 σ_1, σ_2 를 모르는 경우가 많다. 따라서 모표준편차 대신 그 추정량인 표본표준편차를 사용하여 스튜던트화된 표본평균의 차인

$$T=\frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$$

를 사용하여 추론을 하게 된다.

정리 11.10. 이표본에 의한 모평균의 비교 (σ_1, σ_2 **를 모르고,** $\sigma_1 \neq \sigma_2$, $n_1, n_2 \gg 1$)
표본의 크기가 충분히 크면 T의 분포가 표준정규분포에 가까워진다.

- $\mu_1 - \mu_2$ 에 대한 $100(1-\alpha)\%$ 신뢰구간: $(\overline{x} - \overline{y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- H_0: $\mu_1 - \mu_2 = \delta_0$ 에 대한 검정
  - 검정통계량: $T = \dfrac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim_{H_0} \mathcal{N}(0,1)$
  - 관측값: $t = \dfrac{\overline{x} - \overline{y} - \delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
- 대립가설의 형태에 따른 기각역과 유의확률[16]

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu_1 - \mu_2 > \delta_0$	$t \geq z_\alpha$	$\mathrm{P}(T \geq t)$
H_1: $\mu_1 - \mu_2 < \delta_0$	$t \leq -z_\alpha$	$\mathrm{P}(T \leq t)$
H_1: $\mu_1 - \mu_2 \neq \delta_0$	$\lvert t\rvert \geq z_{\alpha/2}$	$\mathrm{P}(\lvert T\rvert \geq \lvert t\rvert)$

연습문제 11.11. 지역 환경에 따라 학력에 차이가 있는가를 알아보기 위하여, 두 도시의 중학교 1학년 학생 중에서 각각 90명과 100명을 랜덤추출하여 동일한 시험을 시행한 결과가 다음과 같았다. 두 도시의 중학교 1학년 학생 전체의 평균 성적에 차이가 있는지 유의수준 1%에서 검정하고, 유의확률을 구하여라. 그리고 평균 성적의 차이에 대한 신뢰수준 99%의 신뢰구간도 구하여라.

	도시 1	도시 2
표본 크기	90	100
평균	76.4	81.2
표준편차	8.2	7.6

[16] 표본의 크기가 크므로, 근사적으로 $T \sim \mathcal{N}(0,1)$ 임에 주의한다.

정리 11.12. 이표본에 의한 모평균의 비교 (σ_1, σ_2 **를 모르고,** $\sigma_1 \neq \sigma_2$)

n_1, n_2 가 충분히 크지 않고(5 이상), 정규모집단인 경우이다.

- $df = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$ (Welch–Satterthwaite Equation)

- $\mu_1 - \mu_2$ 에 대한 $100(1-\alpha)\%$ 신뢰구간: $(\overline{x} - \overline{y}) \pm t_{\alpha/2}(df)\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

- H_0: $\mu_1 - \mu_2 = \delta_0$ 에 대한 검정
  - 검정통계량: $T = \dfrac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim_{H_0} t(df)$
  - 관측값: $t = \dfrac{\overline{x} - \overline{y} - \delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu_1 - \mu_2 > \delta_0$	$t \geq t_\alpha(df)$	$\mathrm{P}(T \geq t)$
H_1: $\mu_1 - \mu_2 < \delta_0$	$t \leq -t_\alpha(df)$	$\mathrm{P}(T \leq t)$
H_1: $\mu_1 - \mu_2 \neq \delta_0$	$\lvert t\rvert \geq t_{\alpha/2}(df)$	$\mathrm{P}(\lvert T\rvert \geq \lvert t\rvert)$

연습문제 11.13. 16 마리의 쥐를 대상으로 진행한 다음 실험에 대하여 질산칼륨의 과다 섭취가 성장을 저해하는지 유의수준 5% 에서 검정하여라.

	질산칼륨 섭취	규정식 섭취
표본 크기	9	7
평균	15.07	19.27
표준편차	3.56	8.05

위에서 소개한 방법들은 전부 근사적인 방법이다. 또한, 두 모집단의 분산이 다르다는 가정 하에서 사용할 수 있는 방법들이었다. 이와 같이 **모집단의 분산이 다를 때** 사용하는 t 검정을 **비합동 이표본 t 검정**(nonpooled two sample t-test)이라 한다. 두 **모집단의 분산이 같은 경우** 사용하는 더욱 효율적인 추론 방법인 **합동 이표본 t-검정**(pooled two sample t-test)에 대해 알아보자.

정리 11.14. 공통분산의 추정 ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)
두 표본분산 S_1^2, S_2^2 의 자유도인 $n_1 - 1$, $n_2 - 1$ 을 가중치로 하여 이들의 가중치 평균인

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

를 공통분산 σ^2 의 추정량으로 생각할 수 있다. 이를 **합동 표본 분산**(pooled sample variance)라고 하고 S_p^2 으로 나타내며, 자유도는 $n_1 + n_2 - 2$ 이다.

	모집단 1	모집단 2	합동표본분산
표본분산	S_1^2	S_2^2	$S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$
자유도	$n_1 - 1$	$n_2 - 1$	$n_1 + n_2 - 2$

두 모집단의 분산이 같을 때, $\mathbf{V}(\overline{X} - \overline{Y}) = \dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}$ 이다. 표본평균의 차를 표준화하면

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

를 얻고, 모분산 σ^2 대신 그 추정량인 S_p^2 를 이용하여 스튜던트화 하면

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

를 얻는다. 이를 이용하여 모평균을 비교할 수 있다.

정리 11.15. 이표본에 의한 모평균의 비교 (σ_1, σ_2 **를 모르고,** $\sigma_1 = \sigma_2$)
공통분산의 정규모집단인 경우이다.

- $\mu_1 - \mu_2$ 의 $100(1-\alpha)\%$ 신뢰구간: $(\overline{x} - \overline{y}) \pm t_{\alpha/2}(n_1 + n_2 - 2)s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$

- H_0: $\mu_1 - \mu_2 = \delta_0$ 에 대한 검정
  - 검정통계량: $T = \dfrac{\overline{X} - \overline{Y} - \delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim_{H_0} t(n_1 + n_2 - 2)$
  - 관측값: $t = \dfrac{\overline{x} - \overline{y} - \delta_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$
- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\mu_1 - \mu_2 > \delta_0$	$t \geq t_\alpha(n_1 + n_2 - 2)$	$\mathrm{P}(T \geq t)$
H_1: $\mu_1 - \mu_2 < \delta_0$	$t \leq -t_\alpha(n_1 + n_2 - 2)$	$\mathrm{P}(T \leq t)$
H_1: $\mu_1 - \mu_2 \neq \delta_0$	$\lvert t\rvert \geq t_{\alpha/2}(n_1 + n_2 - 2)$	$\mathrm{P}(\lvert T\rvert \geq \lvert t\rvert)$

연습문제 11.16. 두 약물 A, B에 대하여 효과가 나타나기까지 걸리는 시간을 조사하기 위해 건강 상태가 비슷한 지원자 12명 중에서 6명을 랜덤추출하여 약물 A를, 나머지 6명에게는 약물 B를 주사하여 측정한 결과가 다음과 같았다. 두 약물의 효과가 나타나기까지 걸리는 시간은 분산이 동일한 정규분포를 따른다고 할 때, 약물 효과가 나타나기까지 걸리는 평균 시간에 차이가 있는지 유의수준 5%에서 검정하고, 유의확률을 구하여라. 또한, 평균 시간의 차이에 대한 95% 신뢰구간도 구하여라.

약물 A	19.54	14.47	16.00	24.83	26.39	11.49
약물 B	15.95	25.89	20.53	15.52	14.18	16.00

	약물 A	약물 B
표본 크기		
평균		
표준편차		

합동 이표본 t-검정을 사용할 때에는 **두 모집단의 분산이 같아야 한다**는 전제 조건에 유의해야 한다. 특히 표본의 크기 n_1, n_2가 매우 다를 때에는 전제 조건이 충족되지 않으면 검정의 유효성이 쉽게 상실된다는 것이 알려져 있다.

현실에서는 분산이 지나치게 큰 상황도 문제가 될 수 있다. 이제 모평균 μ 와 모표준편차 σ 가 모두 미지인 **정규모집단**에서[17] 모분산에 대한 유의성검증을 하는 방법을 알아보자.
모분산 σ^2 의 추정량인 표본분산 S^2 을 이용하여 추정을 하고, 표본분산의 표본분포는

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

임을 이용하여 모분산에 관한 추정이나 유의성검증을 할 수 있다.

정리 11.17. 모분산의 구간추정 (정규모집단의 경우)
모분산 σ^2 에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

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

증명. 연습문제로 남긴다.

정리 11.18. 모분산 σ^2 에 관한 추론 (정규모집단의 경우) - χ^2 검정

- H_0: $\sigma^2 = \sigma_0^2$ 에 대한 검정
  - 검정통계량: $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim_{H_0} \chi^2(n-1)$
  - 관측값: $\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$
- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\sigma^2 > \sigma_0^2$	$\chi_0^2 \geq \chi_\alpha^2(n-1)$	$\mathrm{P}(\chi^2 \geq \chi_0^2)$
H_1: $\sigma^2 < \sigma_0^2$	$\chi_0^2 \leq \chi_{1-\alpha}^2(n-1)$	$\mathrm{P}(\chi^2 \leq \chi_0^2)$
H_1: $\sigma^2 \neq \sigma_0^2$	$\chi_0^2 \geq \chi_{\alpha/2}^2(n-1)$ 또는 $\chi_0^2 \leq \chi_{1-\alpha/2}^2(n-1)$	$2\mathrm{P}(\chi^2 \geq \chi_0^2)$ 또는 $2\mathrm{P}(\chi^2 \leq \chi_0^2)$

양측 검정의 경우 유의확률을 계산하고 2가지 중 1보다 작은 값을 사용한다.

[17]정규모집단이라는 가정에 추론의 타당성이 깊게 의존하고 있다.

연습문제 11.19. 플라스틱 판을 생산하는 한 공장에서 생산되는 판 두께의 표준편차가 1.5mm 를 상회하면 공정에 이상이 있는 것으로 간주한다. 점검을 위해 10개의 판을 랜덤 추출하여 두께를 측정한 결과가 mm 단위로 다음과 같이 주어졌다. 과거의 기록에 의하면 판 두께의 분포가 정규분포를 따른다고 할 때, 공정에 이상이 있는지 유의수준 5%에서 검정하고 유의확률을 구하여라. 또한, 판 두께의 표준편차에 대한 95% 신뢰구간도 구하여라.

$$226, 228, 226, 225, 232, 228, 227, 229, 225, 230$$

다음으로는 두 모집단의 표준편차를 비교하는 방법에 대하여 알아보자.

정리 11.20. 모분산의 비 σ_1^2/σ_2^2 에 관한 추론 (정규모집단의 경우)

- 자료구조: $X_1, \ldots, X_{n_1}$ 는 $\mathcal{N}(\mu_1, \sigma_1^2)$ 의 랜덤표본, $Y_1, \ldots, Y_{n_2}$ 는 $\mathcal{N}(\mu_2, \sigma_2^2)$ 의 랜덤표본으로 서로 독립이다.
- 추정량: $\widehat{\sigma_1^2/\sigma_2^2} = S_1^2/S_2^2$ (S_1^2 은 X_i 의 표본분산, S_2^2 은 Y_i 의 표본분산)
- 표본분포: $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$
- σ_1^2/σ_2^2 의 $100(1-\alpha)\%$ 신뢰구간:

$$\left(\frac{S_1^2/S_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{S_1^2}{S_2^2} \cdot F_{\alpha/2}(n_2 - 1, n_1 - 1) \right)$$

증명. 신뢰구간의 증명은 연습문제로 남긴다.

정리 11.21. 모분산의 비 σ_1^2/σ_2^2 에 관한 추론 <u>(정규모집단의 경우)</u> - F 검정

- H_0: $\sigma_1^2 = \sigma_2^2$ 에 대한 검정
  - 검정통계량: $F = \dfrac{S_1^2}{S_2^2} \sim_{H_0} F(n_1 - 1, n_2 - 1)$
  - 관측값: $f = \dfrac{s_1^2}{s_2^2}$

- 대립가설의 형태에 따른 기각역과 유의확률

대립가설	유의수준 α에서의 기각역	유의확률
H_1: $\sigma_1^2 > \sigma_2^2$	$f \geq F_\alpha(n_1-1, n_2-1)$	$\mathrm{P}(F \geq f)$
H_1: $\sigma_1^2 < \sigma_2^2$	$f \leq 1/F_\alpha(n_2-1, n_1-1)$	$\mathrm{P}(F \leq f)$
H_1: $\sigma_1^2 \neq \sigma_2^2$	$f \geq F_{\alpha/2}(n_1-1, n_2-1)$ 또는 $f \leq 1/F_{\alpha/2}(n_2-1, n_1-1)$	$2\mathrm{P}(F \geq f)$ 또는 $2\mathrm{P}(F \leq f)$

양측 검정의 경우 유의확률을 계산하고 2가지 중 1보다 작은 값을 사용한다.

연습문제 11.22. 두 폴리머를 사용할 때의 주입 압축률에 대한 자료이다. 두 폴리머의 압축률의 산포가 다르다는 증거가 있는가를 유의수준 5%에서 검정하고, 이들의 표준편차의 비에 대한 95% 신뢰구간을 구하여라.

Epoxy	1.75	2.12	2.05	1.97
MMA Prepolymer	1.77	1.59	1.70	1.69

연습문제 11.23. 정신분열증환자 9명과 비슷한 연령의 정상인 9명의 해부로부터 뇌세포의 견본도에 대해 특정한 실험을 행하였다. 이 실험에서 각 세포조직에 대하여 1시간 동안에 특정한 효소활동에 의해 형성되는 물질의 양을 측정한 결과가 다음과 같았다. 물음에 답하여라.

	평균	표준편차
정상인	39.8	8.16
환자	35.5	6.93

(1) 정상인보다 환자의 평균 뇌세포 활동이 저조한지 유의수준 5%에서 검정하여라.

(2) 두 그룹의 사람들의 평균 뇌세포 활동의 차에 대한 95% 신뢰구간을 구하여라.

(3) 정상인의 세포활동의 표준편차가 환자보다 크다고 할 수 있는지 유의수준 5%에서 검정하여라.

(4) 두 모표준편차의 비에 대한 90% 신뢰구간을 구하여라.