

목차

1

자료의 요약

2

확률변수와 확률분포

3

다양한 이산확률분포

4

연속확률변수

5

다양한 연속확률분포

6

모집단과 표본

7

표본과 구간추정

8

가설검정

1 자료의 정리와 요약

1.1 모집단과 표본

우리는 일상생활에서 여론조사 결과, 출산율, 실업률과 같은 다양한 지표들을 마주하게 된다. 또한 여러분이 의료계나 이공계에 진학하게 되더라도, 환자들의 데이터를 분석하고, 실험 결과를 정리할 필요성이 있다. 수많은 데이터로부터 어떻게 의미있으면서도 간결한 정보를 얻을 수 있을까? 이에 대한 해답을 제공하는 것이 바로 **통계학**(Statistics)이다.

모집단 : 연구의 대상이 되는 모든 가능한 대상의 집합

표본 : 모집단에서 자료 조사를 위하여 임의로 선택된 일부 집합

문제 1. 1. 서울과학고등학교 학생 중 아이돌을 좋아하는 학생이 300명 있다고 하자. 그 중 40명을 뽑아 무슨 아이돌을 좋아하는지 확인하였다. 모집단의 크기와 표본의 크기는 각각 얼마인가?

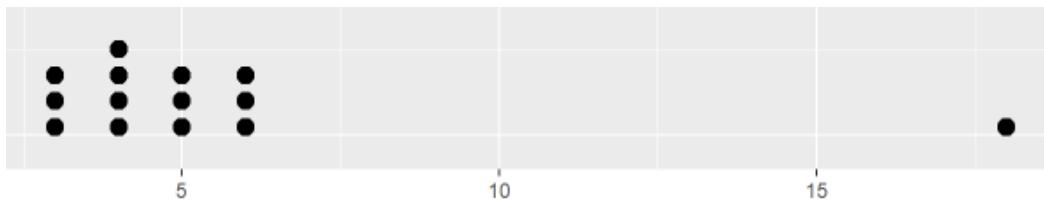
몇몇 경우에는 모집단 전체에 대하여 조사하기도 하며, 이를 전수조사라고 한다. 예를 들면, **문제 1. 1.**에서 300명 전체에 대해 좋아하는 아이돌을 물어본 것이다. 그러나 이렇게 모두 물어보기에는 굉장히 오랜 시간이 걸리기 때문에, 표본을 고르는 것이 거의 필수적이다. 일반적으로, 표본은 무작위로 골라진다.

기술통계학 : 자료를 어떻게 기술할지를 다루는 통계학의 한 분야로, 자료를 정리하여 그래프나 표를 만들거나, 자료의 분포 형태, 대푯값, 변동의 크기 등을 요약하는 방식에 대한 것이다.

추측통계학 : 표본으로부터 모집단의 성질을 어떻게 추측할지를 다루는 통계학의 한 분야로, 통계적 모형을 설정한 후 그 합리성을 평가하고, 이를 바탕으로 관측되지 않은 것에 대한 잠정적 결론을 내리는 것을 목표로 한다.

문제 1. 2. 중학교 때 배운 '도수분포표'는 기술통계학과 추측통계학 중 어느 분야에 속하겠는가?

1.2 점도표



위 그림과 같이 수직선 위에 각 원소에 대응되는 점을 찍어 나타낸 그림을 **점도표**라 한다. 점도표에서 가장 오른쪽에 다른 자료들과는 확실히 구분되는 점이 있는데, 이와 같이 자료의 전체적인 경향성으로부터 벗어난 점을 **이상점**이라 부른다.

문제 1. 3. 위의 점도표는 서울과학고등학교 학생들 중 몇 명을 뽑아 그들의 하루 핸드폰 사용 시간을 기록한 것이다. 표본의 크기는?

1.3 줄기와 잎 그림

만약 자료가 수치형이면서, 자릿수가 2개 이상이라면 줄기와 잎 그림을 이용하여 이를 쉽게 표현할 수 있다. 자료를 두 부분으로 나누어 앞 부분을 줄기라 하고, 뒷 부분을 잎으로 두자. 그렇다면 같은 줄기에 달린 자료들은 어느 정도의 유사성을 가지면서, 뒷 부분에 의해 서로 구분될 수 있다.

1		34
2		399
3		27
4		12889
5		1
6		
7		
8		
9		3

만약 한 줄기에 잎이 너무 많다면, 줄기 내부에서도 이를 세분화하는 방식으로 정리할 수도 있다.

문제 1. 4. 위 줄기와 잎 그림은 걸그룹 우주소녀 팬클럽 우정에게 물어 조사한 우주소녀 앨범 구매 개수이다. 이 자료에서 이상점은?

1.4 도수분포표

도수는 자료에서 원하는 값, 혹은 원하는 범위에 속하는 원소의 개수를 의미한다. **도수분포표**는 자료의 값이 가지는 범위를 적당한 간격으로 나누어 몇 개의 계급을 만들고, 각각의 도수를 구함으로써 만든 표를 말한다. 일반적으로, 계급의 왼쪽을 포함하고 오른쪽은 포함하지 않는다.

계급	도수
0곡~10곡	2
10곡~20곡	2
20곡~30곡	16
30곡~40곡	10
40곡~50곡	6
계	36

도수분포표는 아래의 순서대로 만든다.

- 1) 자료의 최댓값과 최솟값을 찾는다.
- 2) 적절한 계급의 개수를 정한다. 계급의 개수가 너무 적으면 분포상태를 알기 어렵고, 너무 많으면 자료의 전반적인 형태를 알기 쉽지 않다. 일반적으로, $\sqrt{\text{자료의 수}} \pm 3$ 내부에서 정한다.
- 3) 중복되지 않고 동일한 간격을 갖도록 계급 구간을 정한다.
- 4) 각 계급에 속하는 관측값의 개수를 세어 도수를 구한다.

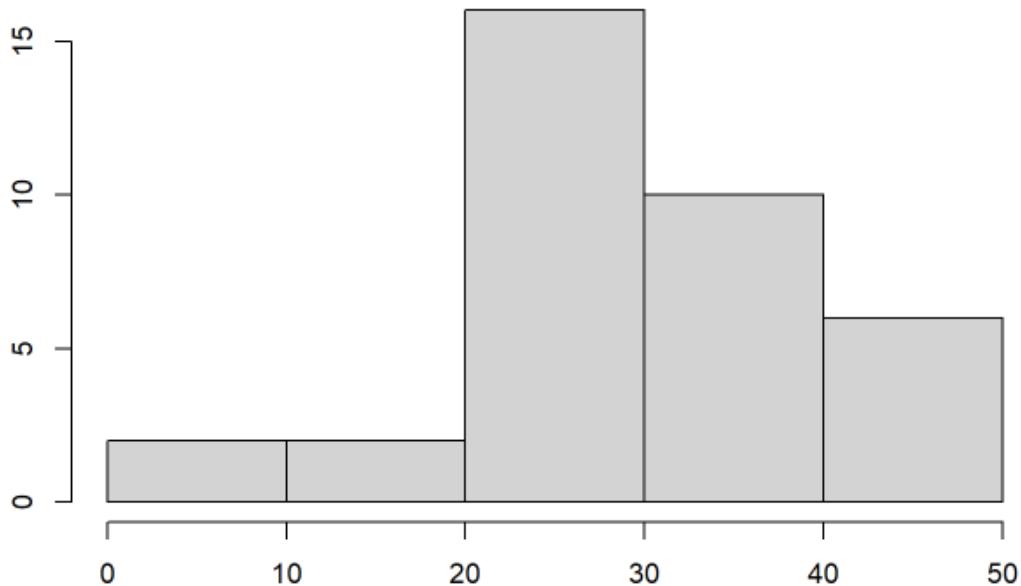
이때, 계급구간의 길이를 계급의 **나비**라고 부른다. 예를 들어, 위의 도수분포표에서는 나비가 10이다.

도수분포표에서 계를 함께 적어주기는 하지만, 비율이 어느 정도인지 아는 것이 분포의 전반적인 형태를 아는 것에 더욱 도움이 될 수도 있다. **상대도수**는 전체 자료의 수로 도수를 나누어 그 계급에 속하는 자료들의 비율을 나타내며, **상대누적도수**는 해당 계급 이하의 모든 상대도수를 합쳐 구한 값으로, 그것의 이하에 얼마나 많은 자료들이 속하는지를 표시한다.

문제 1. 5. 위의 도수분포표는 서울과학고등학교 학생들을 대상으로 90년대 노래를 몇 곡 아는지를 조사하여 나타낸 도수분포표이다. 90년대 노래를 40곡 이상 50곡 미만 아는 학생들의 상대도수는 얼마인가?

1.5 히스토그램

히스토그램은 도수분포표를 시각화하여 자료의 분포를 한눈에 알아볼 수 있도록 만든 그림이다. 가로축은 계급구간을, 세로축은 도수를 의미한다.



히스토그램은 일반적으로 균등한 계급구간의 길이를 전제로 하나, 계급의 크기가 다를 경우에는

$$\text{사각형의 높이} = \frac{\text{계급의 상대도수}}{\text{계급간격}}$$

로 한다. 즉, 면적이 상대도수에 비례하게 한다. **상대도수밀도**는 한 계급의 상대도수를 그 계급의 폭으로 나눈 것을 말한다.

문제 1. 6. 히스토그램에서 그려진 막대들의 넓이 합은 (계급의 나비) \times (도수의 총합)임을 보여라. 단, 계급의 나비는 모든 계급에 대해 같다고 하자.

2 대푯값

주어진 자료가 어떤 값을 중심으로 분포되어 있는지 나타내는 척도인 **대푯값**에는 평균, 중앙값, 최빈수, 백분위수와 같이 다양한 것들이 있다. 앞으로 우리는 n 개의 관측값으로 이루어진 데이터를 x_1, x_2, \dots, x_n 으로 표시할 것이며, x_i 는 i 번째 관측값을 의미하는 것이다.

2.1 평균

자료 x_1, x_2, \dots, x_n 이 주어졌을 때, 자료의 평균은 \bar{x} 혹은 m 으로 표시하며,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

으로 계산한다.

만약 자료가 도수분포표 형태로 주어졌을 때는, 다른 방식을 사용할 수 있다. 변량별로 도수를 주욱 늘어놓게 된다면, 이들은 도수의 총합 N 이 n 의 자리를 대신하며 뒷여서 표시된 자료들이 각각의 x_i 로 표현된다. 따라서, 변량 x_1, x_2, \dots, x_n 에 대하여 각각의 도수가 f_1, f_2, \dots, f_n 이라 한다면

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_n f_n}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i}$$

인 것이다. 만약 계급이 값이 아니라 구간으로 주어졌을 때는, 계급의 중앙값인 **계급값**을 각 계급의 대푯값으로서 사용해 변량 자리에 각 구간의 계급값을 넣어 계산을 수행한다.

도수분포표에서 여러 개의 숫자를 직접 계산하는 데에는 계산량이 많다는 단점이 있다. 따라서 **가평균**을 이용하여 계산량을 조금이나마 줄일 수 있다. 도수분포표의 경우, 아래 식을 통해 계산할 수 있다고 교과서에 언급되어 있다. 아래에서, 계급의 나비가 c 이며 각 변량은 x_i , 각각의 도수는 f_i 로 표시되었다. 가평균은 x_0 로 선택되었다.

$$\begin{aligned} u_i &= \frac{x_i - x_0}{c} \quad (i = 1, 2, 3, \dots, n) \\ \rightarrow x_i &= x_0 + cu_i \end{aligned}$$

로 정한다는 언급이 나와 있는데, 이로부터 우리는 u_i 가 의미하는 바가 속해 있는 계급이 가평균으로부터 얼마나 떨어져 있는지를 의미하는 것임을 알 수 있다. 그렇다면,

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i} \\ &= \frac{\sum_{i=1}^n (x_0 + cu_i) f_i}{\sum_{i=1}^n f_i} \\ &= \frac{x_0 \sum_{i=1}^n f_i}{\sum_{i=1}^n f_i} + \frac{c \sum_{i=1}^n u_i f_i}{\sum_{i=1}^n f_i} \\ &= x_0 + c\bar{u} \end{aligned}$$

라는 것이다. 예를 들면, 자료 3, 4, 5, 6, 7의 평균을 구하기 위해서, 3+4+5+6+7를 하는 대신, 가평균으로 5를 잡은 이후 각각을 -2, -1, 0, 1, 2로 두어 그 합이 0인 것을 이용하겠다는 것이다. 이로써 평균이 5임을 쉽게 구할 수 있게 된다.

문제 1. 7. $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$ 임을 보여라.

2.2 중앙값

자료의 **중앙값**은 \hat{x} 로 표시하며, 이는 자료를 크기에 따라 늘어놓을 때 정중앙에 속하는 값을 말한다. 자료의 개수가 짝수면 정중앙에 속하는 수가 없으므로, 정중앙의 두 개를 골라 그 평균을 구함으로써 정의한다. 만약 이상점이 있는 데이터라면 평균이 왜곡될 가능성이 크기에, 분포가 극단적이라면 중앙값을 주로 선호하는 경향이 있다.

2.3 사분위수와 백분위수

중앙값이 중앙에 속하는, 즉 상위 50퍼센트의 값을 이야기한 것이라면 이를 확대할 수도 있다. 제 p 백분위수는 자료를 작은 값부터 늘어놓아 오름차순일 때, 적어도 $p\%$ 의 관측값들이 그 값보다 작거나 같게 되고, $(100 - p)\%$ 의 관측값들이 크거나 같게 되도록 정해진 것이다. 또한, 사분위수는 자료를 4등분하여 그 분기가 어디인지를 표시한 것이다. 즉 제 25백분위수는 제 1사분위수이며, Q_1 이라 표시한다. 같은 이유로 50백분위수는 중앙값이자 제 2사분위수로, Q_2 이고 75백분위수는 제3사분위수이며 Q_3 로 표시한다.

문제 1. 8. 3쪽의 히스토그램에서 제1사분위수와 제3사분위수는 각각 얼마인가?

2.4 그 외의 척도들

최빈값 : 도수가 가장 많은 관측값

범위의 중앙값 : 최댓값과 최솟값의 산술평균

절사평균 : 작은 관측값과 큰 관측값 일부를 버린 후 나머지 관측값에 대한 산술평균

문제 1. 9. 함수 $f : \mathbb{R} \rightarrow \mathbb{R}$ 이 있다. 주어진 자료 x_1, x_2, \dots, x_n 에 대하여, 처리한 자료 $f(x_1), f(x_2), \dots, f(x_n)$ 을 생각하자. $f(x) := ax + b$ 일 때의 $\bar{f}(x)$ 를 a, b, \bar{x} 를 이용하여 표시하라.

문제 1. 10. 위의 문제에서 f 가 단조함수라고 생각하자. 흔수 개의 자료 $f(x_1), f(x_2), \dots, f(x_n)$ 의 중앙값을 함수 f 와 \hat{x} 를 이용하여 표시하라.

문제 1. 11. 위의 문제에서 $f(x_1), f(x_2), \dots, f(x_n)$ 의 최빈값을 함수 f 와 원래 자료의 유일한 최빈값 x_0 를 이용하여 표시하라.

3 산포도

3.1 분산과 표준편차

산포도는 자료가 얼마나 퍼져 있는지를 나타낸 척도이다. 평균이 같은 자료라도, 이 값에 따라서 분포 양상이 완전히 달라질 수 있다. n 개의 자료 x_1, x_2, \dots, x_n 이 주어져 있을 때, 평균으로부터 떨어진 정도인 $x_i - \bar{x}$ 를 i 번째 편차라 부른다. 편차가 음수이면 평균보다 작은 변량, 편차가 양수이면 평균보다 큰 변량이라고 생각할 수 있다. 아마도, 편차의 절댓값이 클수록 해당 자료가 평균으로부터 멀리 떨어져 있을 수 있음을 알 수 있다.

분산 : 편차의 제곱의 평균이다. σ^2 으로 표시한다.

$$\sigma^2 = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_n - m)^2}{n}$$

표준편차 : 분산의 양의 제곱근으로, σ 로 표시한다.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_n - m)^2}{n}}$$

문제 1. 12. 함수 $f(k) = \frac{(x_1 - k)^2 + (x_2 - k)^2 + \cdots + (x_n - k)^2}{n}$ 의 최소가 되는 k 의 값을 구하여라.

편차의 제곱을 쓰는 이유는 바로 위의 식에서 나온다. 분산은 평균을 중심으로 계산할 수 있는 최적의 방식이며, 제곱이라는 다향식을 기반으로 하기에 미적분을 통해 다루기에도 쉽다. **평균편차**는 제곱이 아닌 절댓값들의 평균, 즉 $\frac{1}{n}(|x_1 - m| + |x_2 - m| + \cdots + |x_n - m|)$ 로 정의한다.

문제 1. 13. 함수 $g(k) = \frac{|x_1 - k| + |x_2 - k| + \cdots + |x_n - k|}{n}$ 의 최소가 되는 k 의 값을 구하여라.

도수분포표에서는 평균을 도수를 하중으로 하여 가중평균을 구함으로써 얻어냈다. 분산을 구하는 과정도 이와 유사하다. 즉, 변량 x_1, x_2, \dots, x_n 에 대하여 각각의 도수가 f_1, f_2, \dots, f_n 이라 한다면

$$\sigma^2 = \frac{(x_1 - m)^2 f_1 + (x_2 - m)^2 f_2 + (x_3 - m)^2 f_3 + \cdots + (x_n - m)^2 f_n}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n (x_i - m)^2 f_i}{\sum_{i=1}^n f_i}$$

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

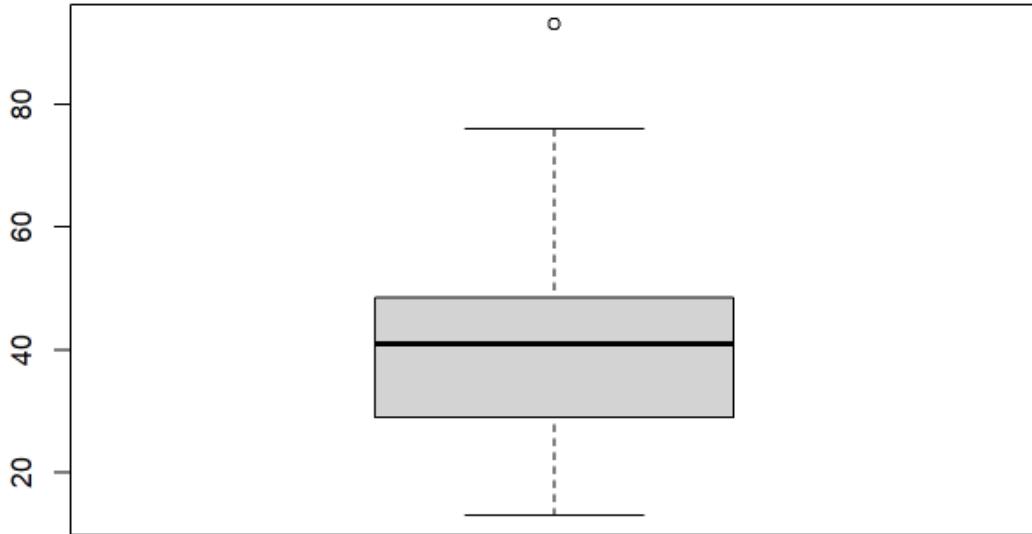
이게 될 것이다. 앞으로도 도수분포표에 대해서는, 각각의 도수만큼의 하중이 실린다고 보면 된다. 따라서, 앞으로는 이것을 연습문제로 남긴다.

아마 우리가 중학교에서 분산을 배웠다면, 제평평제라는 단어를 들어보았을 것이다. 분산이 제곱의 평균에서 평균의 제곱을 뺀 것과 같다는 이야기인데, 우리는 문제 1.7.에서 이미 $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$ 임을 보였다. 분산이 왼쪽의 값을 n 으로 나눈 것임을 고려하면, 분산이 제평평제라는 것을 쉽게 보일 수 있을 것이다. 편차를 하나하나 구하는 것이 매우 귀찮은 과정이기에, 제곱을 해서 계산량을 줄이는 것은 꽤 효과적인 아이디어라고 할 수 있다.

문제 1. 14. 문제 1.9.와 문제 1.7.을 이용하여 가평균 x_0 이 주어졌을 때 분산을 구해 보아라.

4 데이터 시각화

데이터 시각화는 숫자나 문자 형태로 주어진 데이터를 가공해 시각적 자료 형태로 만들므로써 독자가 쉽게 이해할 수 있는 방법이다. 앞서 다룬 점도표, 히스토그램 등은 모두 데이터 시각화의 예이다. 대푯값과 산포도를 모두 다룰 수 있는 데이터 시각화의 한 예시는 **상자 그림**이다.



상자 그림에서 중심에 있는 상자는 제 1사분위수로부터 제 3사분위수 사이의 영역을 의미한다. 중간에 있는 선은 제 2사분위수가 된다. 상자의 크기, 즉 제 1 사분위수와 제 3사분위수 사이의 거리를 **사분위 범위 (IQR)**라 부르며, 이는 산포도의 한 예시이다. 또한 상자의 양 끝에서 점선으로 뻗어나오는 부분이 있는데, 그것은 각각 **LIF**와 **UIF**이며, 각 사분위수로부터 1.5IQR 만큼 벌어진 것이다. 그림에 표시되지 않았지만 **LOF**와 **UOF**도 있는데, 이들은 3IQR 만큼 벌어진 것이다. 추측할 수 있듯이, L은 lower, U는 upper, I는 interior, O는 outer, F는 fence이다. 그림에서 나오지는 않지만 **범위**는 최댓값과 최솟값 사이의 점으로 정의한다.

문제 1. 15. 위의 상자 그림에서 **UIF**를 넘어서는 점은 몇 개인가?

5 연습문제

문제 1. 16. 자료 x_1, x_2, \dots, x_n 의 평균이 m , 표준편차가 σ 이다. 자료 $3x_1^2 - 2, 3x_2^2 - 2, \dots, 3x_n^2 - 2$ 의 평균을 구하여라.

문제 1. 17. 평균과 표준편차를 계산할 때, 계산량을 줄이기 위하여 점화식을 이용하는 경우가 있다. 아래 식들이 성립함을 보여라.

$$\bar{x}_{j+1} = \bar{x}_j + \frac{x_{j+1} - \bar{x}_j}{j+1}$$

$$\sigma_{j+1}^2 = \frac{j}{j+1} \sigma_j^2 + j(\bar{x}_{j+1} - \bar{x}_j)^2$$

1 확률변수와 확률분포

1.1 확률변수와 확률분포

표본공간 : S 로 주로 표기하며, 어떠한 시행으로부터 얻어질 수 있는 모든 결과값을 모은 집합
확률변수 : 표본공간을 정의역으로 하고 실수를 치역으로 하는 함수의 일종으로, 특정 확률로 표본 공간 내의 한 원소가 나올 때 그것을 어떠한 실수값으로 대응시켜주는 역할을 한다.
확률분포, 혹은 분포 : 확률변수의 대응관계

문제 2. 1. 주사위를 던져 나오는 수를 확률변수 X 라고 하자. X 의 확률분포를 구하여라.

위와 같이, 확률분포는 $P(X = x)$ 의 형태나, $P(X \in A)$ 등의 형태로 나타낼 수 있다. 물론, 여기서 A 는 S 의 부분집합이다. 중간고사 범위에서 다룬 확률의 공리와 같이, 아래 성질들이 만족한다.

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

$$\sum_{x \in S} P(X = x) = 1$$

$$\sum_{x \in A} P(X = x) = P(X \in A)$$

$$P(X \in A) + P(X \in (S - A)) = 1$$

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

문제 2. 2. X 를 주사위를 던져 나오는 수로 정의하자. X 가 짝수일 확률을 구하여라.

1.2 확률변수의 종류

만약 S 가 셀 수 있는 집합이라면, X 를 **이산확률변수**라 부른다. (셀 수 있는 집합 : 자연수 집합으로부터 이 집합으로 가는 전단사함수가 존재할 때, 이를 셀 수 있는 집합이라 부른다. 그렇지 못하면, 셀 수 없는 집합이라 부른다. 즉 S 를 주욱 늘어놓을 수 있을 때에만, X 가 이산확률변수가 된다.)

확률질량함수 : X 가 취할 수 있는 값인 $x_1, x_2, \dots, x_n, \dots$ 에 대하여 확률 $P(X = x_1), P(X = x_2), \dots, P(X = x_n), \dots$ 을 대응시키는 함수를 말한다.

이산확률변수에 대해선, 그 확률을 $P(X = x) = f(x)$ ($f(x)$ 는 확률질량함수)로 나타내거나 확률분포표를 통해 아래와 같이 나타낼 수 있다.

X	x_1	x_2	x_3	\dots	x_n	\dots	합계
$P(X = x_1)$	p_1	p_2	p_3	\dots	p_n	\dots	1

만약 위와 같이 확률분포표로 나타내지 못하게 S 가 셀 수 없는 집합일 경우에는, **연속확률변수**를 통해 현상을 분석하며, 이를 다음 시간에 다룰 것이다.

문제 2. 3. 단어 DOGS에서 임의로 한 알파벳을 뽑아 그 알파벳의 알파벳상 순서를 확률변수 X 라 정의하자. 예를 들어, D 를 뽑았다면 $X = 4$ 이다. X 의 확률분포를 구하라.

도수분포표에서 무한하게 도수의 합을 늘렸다면, 아마 그 상대도수는 그 값이 나올 확률에 점점 더 가까워질 것이다. 그렇다면, 도수분포표에서 누적상대도수를 구하였듯이 확률분포에서도 누적분포를 생각할 수 있다. **누적분포함수**를 $F(x)$, 혹은 $P(X \leq x)$ 로 표시하며, $X \leq x$ 를 x 보다 작은 수들의 집합 $t : t \leq x$ 에 대응시킨다는 생각을 해보면 이것이 $\sum_{t \leq x} P(X = t)$ 와 같다고도 말할 수 있다. 아래 성질 역시 성립한다. 사실 이는 위에서 다룬 확률의 공리를 잘 비틀면 알아낼 수 있다.

$$0 \leq F(x) \leq 1$$

$F(x)$ 는 단조증가함수이다.

$$\lim_{t \rightarrow x+} F(t) = F(x)$$

하나 주목할 점은, 마지막 성질은 우극한일 땐 항상 성립하지만 좌극한일 때는 성립하지 않는다는 것이다. 아이러니하게도, 이를 이용하면 $F(x)$ 로부터 $f(x)$ 를 구할 수 있다.

$$\begin{aligned} f(x) &= P(X = x) = \lim_{h \rightarrow 0} P(x - h < X \leq x) \\ &= \lim_{h \rightarrow 0} (P(X \leq x) - P(X \leq x - h)) \\ &= \lim_{h \rightarrow 0} (F(x) - F(x - h)) \\ &= F(x) - \lim_{h \rightarrow 0} F(x - h) \end{aligned}$$

이다. 그렇다면, 연속확률변수일 때는 미분을 이용해서 구할 수 있다는 것 역시 추측할 수 있다.

문제 2. 4. **문제 2.3.**에서 다룬 확률변수 X 에 대하여, 누적분포함수의 그래프를 그려라.

2 이산확률변수의 기댓값과 분산

2.1 기댓값

기댓값은 확률변수 위에서 정의된 평균이나 마찬가지다. 즉, 이산확률변수 X 의 확률분포가 주어졌을 때, 기댓값 $E(X)$ 는 변량과 확률의 곱들의 합인 $\sum_{k=1}^n x_i p_i$ 로 주어진다.

문제 2. 5. 두 개의 주사위를 던졌을 때 두 주사위 눈의 합을 확률변수 Y 라고 하자. $E[Y]$ 의 값을?

2.2 분산과 표준편차

이산확률변수 X 가 주어져 있고, 그 기댓값이 m 이라 하자. 그렇다면 이산확률변수의 분산 $V(X)$ 는 편차의 제곱의 기댓값과 같다. 즉

$$V(X) = E[(X - m)^2] = \sum_{k=1}^n (x_i - m)^2 p_i$$

또한 분산의 양의 제곱근을 표준편차라 부르며, $\sigma(X)$ 로 표시한다. 즉

$$\sigma(X) = \sqrt{V(X)} = \sqrt{\sum_{k=1}^n (x_i - m)^2 p_i}$$

또한, 우리는 n 개의 자료에 대하여 분산이 제평평제임을 알고 있었다. 이산확률변수에 대해서도, 동일하게 이것이 성립한다. 연습문제에서 이것을 보여 보자.

문제 2. 6. 이산확률변수의 분산도 제평평제를 만족함을 보여라. 이로부터, $E[X^2]$ 를 $V(X)$ 와 $E(X)$ 를 이용해 표시해 보아라.

문제 2. 7. 주사위를 던져 나오는 눈을 X 라는 확률변수라 하자. X 의 분산을 구하여라.

2.3 평균, 분산, 표준편차의 성질

- 1) $E[aX + b] = aE[X] + b$ (기댓값의 선형성)
- 2) $V(aX + b) = a^2V(X)$
- 3) $\sigma(aX) = |a|\sigma(X)$

증명은 독자에게 남긴다.

문제 2. 8. 위 박스의 증명을 하여라.

2.4 마코프의 부등식, 체비셰프 부등식

사실 이 부분은 서울과학고 기초통계학을 기준으로 이항분포를 다루며 나온다. 하지만 이는 지금도 증명이 가능하고, 이항분포만이 아니라 전체 이산확률변수에 대해서도 가능하기에 지금 여기서 증명하기로 한다.

마코프의 부등식

X 가 음이 아닌 값만을 가지는 이산확률변수라면, 임의의 양수 $a > 0$ 에 대하여

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

이 성립한다.

pf) 우리는 이제 실수의 부분집합 A 와 B 를 아래와 같이 정의할 것이다.

$$A = \{i : x_i < a\} \quad B = \{i : x_i \geq a\}$$

그렇다면

$$\begin{aligned} E[X] &= \sum_{k=1}^n x_i p_i \\ &= \sum_{i \in A} x_i p_i + \sum_{i \in B} x_i p_i \\ &\geq \sum_{i \in B} x_i p_i \\ &\geq \sum_{i \in B} a p_i \\ &= a P\{x \geq a\} \end{aligned}$$

이므로 증명이 완료된다.

체비셰프 부등식

X 가 기댓값 μ 와 분산 σ^2 을 가지는 이산확률변수라면, 임의의 양수 $k > 0$ 에 대하여

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

이 성립한다.

pf) 확률변수 $(X - \mu)^2$ 을 생각하자. 이는 당연하게도 음이 아닌 값을 가지는 새로운 확률변수이기에, 양수 k^2 에 대해 마코프의 부등식을 적용할 수 있다. 그러면

$$P\{(X - \mu)^2 \geq k^2\} \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

이다. 따라서 이를 다시 정리해주면

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

임을 쉽게 알 수 있다. 사실상 마코프 부등식의 따름정리나 마찬가지다.

사실 이들은 앞으로 큰 수의 법칙을 배우게 될 때 매우 중요하게 사용될 것이다.

3 결합확률분포

이산확률변수가 유한 개의 값 x_1, x_2, \dots, x_m 을 취하고 또 다른 이산확률변수 Y 가 유한 개의 값 y_1, y_2, \dots, y_n 을 가질 때, (X, Y) 를 이변량 이산확률변수라고 한다. 즉 두 개의 확률변수로 이루어진 새로운 체계를 만들어주는 것이나 마찬가지다. 이들의 확률분포는 결합확률질량함수라 불린다.

$1 \leq i \leq m$ 과 $1 \leq j \leq n$ 에 대하여, $P(X = x_i, Y = y_j)$ 를 X 와 Y 의 결합확률질량함수라 부른다.

이들은 아래의 성질을 가진다.

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

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

이를 통해 가지는 대응관계를 결합확률분포라 하며, 이를 표로 표현한 것을 결합확률분포표라 부른다.

주변확률질량함수는 주어진 결합확률질량함수에 대해 한 확률변수에 초점을 맞추고 그에 대한 확률변수를 구하고자 할 때 사용하는 함수이다. 즉 X 의 주변확률질량함수는

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

로 주어지며, Y 의 주변확률질량함수는

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

이다.

문제 2. 9. 90년대 CD 5장과 최신가요 CD 4장이 들어있는 음반에서 임의로 3장의 CD를 뽑고자 한다. 뽑힌 90년대 CD의 수를 확률변수 X , 최신가요 CD의 수를 확률변수 Y 라고 할 때, X 와 Y 의 결합확률질량함수를 구하여라.

3.1 두 확률변수의 독립

사실상 기초통계학을 배움에 있어, 독립을 이해하는 것은 가장 중요하다고 봐도 과언이 아니다. 확률변수의 독립성은 거의 모든 검정에 가정으로서 포함되며, 확률변수를 이용한 문제에서도 주요하게 이용된다. 과연 독립이란 무엇일까?

다음 식을 만족하면 두 확률변수 X 와 Y 가 **독립**이라고 한다.

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

즉 두 확률변수의 주변확률질량함수를 곱하면 그 두 값에 대응되는 결합확률질량함수가 나온다는 것이다. 우리가 알고 있던 사건의 독립의 정의, 즉 X 가 결정되는 사건이 Y 의 결정에 영향을 미치지 않는다는 사실에 부응하는 것처럼 보인다. 혹은 조건부확률로도 이를 해석할 수 있으나, 기초통계학에서는 그 단계까지 뺀어나가지는 않는다.

문제 2. 10. n 번째 주사위에서 나오는 수를 확률변수 X_n 이라 두자. $X_1 + X_2$ 와 X_1 은 독립인가?

3.2 결합확률변수의 기댓값

확률변수 X 와 Y 의 결합확률질량함수가 $P(X = x_i, Y = y_i)$ 일 때 다음이 성립한다.

- 1) $E[g(X, Y)] = \sum_{i=1}^m \sum_{j=1}^n g(x_i, y_i) P(X = x_i, Y = y_i)$
- 2) X 와 Y 가 독립이면 $E[XY] = E[X]E[Y]$
- 3) $E[aX + bY] = aE[X] + bE[Y]$ (기댓값의 선형성)

두번째 식은 아래와 같이 보일 수 있다. 이를 따르면 세번째 식 역시 증명해볼 수 있다. 그것은 연습문제로 남긴다.

$$\begin{aligned} E[XY] &= \sum_{i=1}^m \sum_{j=1}^n x_i y_i P(X = x_i, Y = y_i) \\ &= \sum_{i=1}^m \sum_{j=1}^n x_i y_i P(X = x_i) P(Y = y_i) \\ &= \sum_{i=1}^m x_i P(X = x_i) \sum_{j=1}^n y_i P(Y = y_i) \\ &= E[X] \sum_{i=1}^m x_i P(X = x_i) \\ &= E[X]E[Y] \end{aligned}$$

문제 2. 11. 세 번째 식, 즉 $E[aX + bY] = aE[X] + bE[Y]$ 을 보여라.

위의 결과를 일반화한다면, n 개의 확률변수에 대해서

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n]$$

이 성립함을 알 수 있으며, 그것들이 모두 독립이라면

$$E[X_1X_2 \cdots X_n] = E[X_1]E[X_2] \cdots E[X_n]$$

임을 알 수 있다. 이때 모두 독립이라는 것은 pairwise하게 독립이라는 것이다.

3.3 공분산

한 개의 확률변수에 대해서는 분산을 구했었다. 두 개의 확률변수의 대해서는 공분산이라는 개념을 적용하여 두 확률변수가 어떻게 관계를 맺고 있는지에 대해 알 수 있다.

공분산 : 두 확률변수 X 와 Y 의 상호 변동을 나타내는 척도

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

이를 다르게 계산하면,

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

그리고 $\text{Cov}(X, X) = V(X)$ 임도 알고 있다.

특히나, 만약 X 와 Y 가 독립이라면 $E[XY] = E[X]E[Y]$ 이므로 $\text{Cov}(X, Y) = 0$ 임을 알 수 있다. 그러나 그 역은 성립하지 않는다. 가장 간단한 예시로, 주사위에서 1,2가 나오면 $X = 1, Y = 0$, 3,4가 나오면 $X = 0, Y = 1, 5,6$ 이 나오면 $X = -1, Y = 0$ 이라 하자. 그러면 $E[XY] = 0$ 이고, $E[X]E[Y] = 0$ 이다. 그러나 X 와 Y 는 독립이 아니다. 그 외에도 다양한 예시가 있을 수 있다. 중요한 것은 문제가 '두 확률변수 X 와 Y 가 독립임을 보여라'라면, 우리가 선택 가능한 방법은 독립의 정의를 이용하는 것만이 유일하다는 것이다.

다음으로 확률변수 $X + Y$ 의 분산을 구하여 보자. 정의를 이용하여 구하는 것이 가장 효율적일 것이다.

$$\begin{aligned}V(X + Y) &= E[(X + Y - E[X + Y])^2] \\ &= E[(X - E[X])^2 + 2(X - E[X])(Y - E[Y]) + (Y - E[Y])^2] \\ &= E[(X - E[X])^2] + 2E[(X - E[X])(Y - E[Y])] + E[(Y - E[Y])^2] \\ &= V(X) + V(Y) + 2\text{Cov}(X, Y)\end{aligned}$$

만약 확률변수 X 와 Y 가 독립이라면 $\text{Cov}(X, Y) = 0$ 이다. 따라서 $V(X + Y) = V(X) + V(Y)$ 이다.

문제 2. 12. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$ 임을 보여라.

3.4 상관계수

상관계수 : 두 변수의 상관관계를 나타내는 척도

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

이때, 상관계수의 절댓값이 1에 가까울수록 선형관계가 두드러지며, 0에 가까울수록 둘 사이의 선형관계가 명확하지 않다고 풀이된다. 또한, 그것이 양이면 X 가 증가하면 Y 가 증가하는 양의 상관관계, 음이면 음의 상관관계가 나타난다. 단, 주의할 것은 상관계수는 둘 사이의 선형관계만을 알려줄 뿐 관계가 없다는 것을 의미하지는 않는다. 예를 들어, 둘 사이의 관계가 이차함수 개형이라면 선형적인 관계성은 적게 나타날 것이다. 이 값은 항상 -1과 1 사이에 있다.

4 연습문제

문제 2. 13. $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$ 임을 보여라.

문제 2. 14. 아래 식을 보여라.

$$\text{Cov} \left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j \right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$$

문제 2. 15. 아래 식을 보여라.

$$V \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n V(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n Cov(X_i, X_j)$$

문제 2. 16. 다섯 개의 트랜지스터가 테스트를 앞두고 있다. 무작위로 그들을 배열하고, 그들이 불량품인지 확인하려 한다. 다섯 개 중 3개의 트랜지스터가 불량이라고 할 때, N_1 을 첫 불량품이 밝혀질 때까지 수행되는 테스트의 횟수, N_2 를 그 이후 두번째 불량품이 밝혀질 때까지 수행되는 테스트의 횟수라고 하자. N_1 과 N_2 의 결합확률질량함수를 구하여라.

문제 2. 17. 매일 밤 기상학자들은 다음 날 비가 올 확률을 제시한다. 그들의 예측을 평가하기 위하여, 아래와 같이 점수를 매기기로 했다. 만약 기상학자가 p 의 확률로 비가 내린다고 했다면, 비가 내일 경우 $1 - (1 - p)^2$ 점을, 비가 내리지 않을 경우 $1 - p^2$ 점을 받는다. 이를 수행하여 가장 높은 점수를 받은 사람을 최고의 기상학자로 생각하기로 했다. 그들이 그들의 점수 기댓값을 최대한으로 만들기 위해서 어떤 전략을 취해야 할까? 단, 그들이 실제로 생각하는 비가 올 확률은 p^* 이라고 두자.

문제 2. 18. 보험회사는 교통사고가 일어났을 때 대물보험 가입자에게 A 원을 보상한다. 만약 일년 동안 가입자가 교통사고를 일으킬 확률을 p 라 예상하고 있었다면, 보험사의 기대수익이 A 의 10퍼센트가 되려면 연 보험료가 얼마인가?

문제 2. 19. 명반을 찾는 TV쇼에서 148명을 데려다 놓고 락밴드 A, B, C, D 에 대한 선호도 조사를 진행했다. 각각 40, 33, 25, 50명이 A, B, C, D 를 선호한다고 밝혔다. 148명 중 한 명의 방청객이 임의로 선택되었을 때, 확률변수 X 를 해당 사람이 좋아하는 밴드의 선호인 수라고 정의하자. 반대로, 밴드 중 하나를 임의로 선택한 다음 그 밴드의 선호인 수를 확률변수 Y 라고 두자. $E[X]$ 와 $E[Y]$ 를 비교하여라.

문제 2. 20. $p_i = P\{X = i\}$ 이며 $p_1 + p_2 + p_3 = 1$ 이다. 만약 $E[X] = 2$ 라면, $V(x)$ 를 최소화/최대화시키는 p_1, p_2, p_3 의 값을 각각 무엇인가?

문제 2. 21. X, Y 의 분산을 각각 σ_X^2, σ_Y^2 이라 둔 상태에서, $V\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \geq 0$ 임을 이용하여 $Corr(X, Y) \geq -1$ 임을 보여라. 동일한 방식으로, $Corr(X, Y) \leq 1$ 임도 보여라.

문제 2. 22. n 번의 독립시행에 대하여, 각 결과는 1, 2, 3중에 하나로 p_1, p_2, p_3 의 확률로 등장한다. 만약 N_i 를 i 라는 결과가 나온 시행 횟수라고 둔다면, $Cov(N_1, N_2) = -np_1p_2$ 임을 보여라. 또한, $Corr(N_1, N_2)$ 의 부호를 결정하고 이를 직관적으로 설명하여라.

문제 2. 23. 만약 확률변수 X_1, X_2 가 동일한 확률질량분포함수를 가지고 있다면, $Cov(X_1 - X_2, X_1 + X_2) = 0$ 임을 보여라.

문제 2. 24. 합수 $\phi(t)$ 를

$$\phi(t) = E[e^{tX}] = \sum_x e^t x P(X = x)$$

라고 정의하자.

$$\phi'(0) = E[X]$$

$$\phi''(0) = E[X^2]$$

임을 보여라. 수학적 귀납법을 이용하여, $n \leq 1$ 인 자연수에 대해

$$\phi^n(0) = E[X^n]$$

임을 보여라.

문제 2. 25. 위에서 정의한 함수 $\phi(t)$ 를 적률생성함수라 한다. 두 확률 변수 X 와 Y 의 적률생성함수를 $\phi_X(t)$, $\phi_Y(t)$ 라고 하자. 만약 X 와 Y 가 독립이라면, $X + Y$ 의 적률생성함수 $\phi_{X+Y}(t)$ 는

$$\phi_{X+Y}(t) = \phi_X(t) + \phi_Y(t)$$

임을 보여라.

문제 2. 26. 확률변수 X 와 Y 가 독립이라면, 확률변수 $h(X)$ 와 $g(Y)$ 도 독립임을 보여라.

문제 2. 27. 모두 같은 확률분포를 가지는 X_i 의 대푯값이, $Cov\left(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n X_i}{n} - X_1\right)$ 의 값을 구하여라.
단, $n \geq 2$ 라고 한다.

문제 2. 28. 최근 EBS에서 문해력에 관한 다큐를 촬영하기 위해, 25명의 초등학생을 모아두고 수업 중 모르는 단어를 모두 보고하라고 이야기했다. 선행연구 결과, 학생들은 평균 40개의 단어를 몰랐으며, 모르는 단어의 표준편차는 20개였다.

25명의 모르는 단어 수를 모두 합친 것이 1100을 넘을 확률이 10/11 이하임을 증명하여라.

문제 2. 29. 서울과학고의 기초통계학 학점 부여 기준을 보면, 30퍼센트의 학생은 A, 30퍼센트의 학생은 B, 20퍼센트의 학생은 C, 그리고 20퍼센트의 학생은 D나 F를 받게 되어 있다. 수학 선생님께서는 과목별로 독립적으로 학점을 부여하신다. 진섭이는 이번 학기에 수학 과목을 세 개 듣는다. 확률변수 X 를 진섭이가 받은 A의 수라고 하자.

- 1) X 의 확률질량함수를 구하여라.
- 2) X 의 누적확률질량함수를 구하여라.

문제 2. 30. 확률변수 X 가 확률질량함수를 $P(X = n) = \frac{1}{2^n}$ 으로 가진다. 단, $n \geq 1$ 이다. $E[X]$ 의 값은?

문제 2. 31. 확률변수 Y 가 확률질량함수를 $P(Y = n) = P(Y = \frac{1}{n}) = \frac{1}{2^{n+1}}$ 로 가진다. 단 $n \geq 2$ 일 때며, $n = 1$ 이라면 확률은 0.5이다. $E[Y]$ 의 값을 구하여라.

문제 2. 32. i) 산화률변수 X 에 대하여, $E[X^2] < \infty$ 라면 $E[X] < \infty$ 임을 보여라.

문제 2. 33. 유리수의 집합 \mathbb{Q} 는 셀 수 있는 집합이라고 알려져 있다. 또한, 두 실수 사이에는 항상 유리수가 존재한다. 이제 구간 $(0, 1)$ 에 있는 모든 유리수를 r_1, r_2, \dots , 와 같은 식으로 나열한 이후에 확률변수 X 를 확률질량함수가 $P(X = r_n) = \frac{1}{2^n}$ 이라 정의하자. 그렇다면, $(0, 1)$ 의 어떠한 열린 부분구간에서도 X 의 누적확률밀도함수가 상수함수가 아님을 보여라.

문제 2. 34. 음이 아닌 정수를 값으로 갖는 이산확률변수 X 를 생각하자. X 의 기댓값은 유한하다고 가정하자.

- 1) $E[X] = \sum_{k=0}^{\infty} (1 - F_X(k))$ 임을 보여라. 단, $F_X(k)$ 는 X 의 누적확률질량함수이다.
- 2) $E[X^2] - E[X] = 2 \sum_{k=1}^{\infty} kP(X > k)$ 임을 보여라.

1 이항분포

1.1 베르누이 분포

베르누이 시행 : 시행의 결과가 두 가지로 구분되는 시행

베르누이 분포 : 성공확률이 p 인 베르누이 시행에서 성공은 1, 실패는 0으로 대응되는 확률변수 X 가 따르는 확률분포

확률질량함수는 $P(X = x) = p^x(1 - p)^{1-x}$, $x = 0$ or 1로 정의된다.

베르누이 분포의 평균과 분산을 구해보자.

$$E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

$$V(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$$

서로 독립인 베르누이분포가 여러 개 누적되면 어떻게 될까? 예를 들어, n 번의 베르누이 시행을 시행하여 성공 횟수를 X 라는 확률변수에 대응하는 것이다. 그렇다면 서로 독립인 n 개의 베르누이 분포를 더한 것이다. 그것이 다음으로 다룰 이항분포다.

1.2 이항분포

이항분포 : n 회의 독립시행 중 일어날 확률이 p 인 사건이 일어나는 횟수가 가지는 확률분포

: 이때, 확률변수 X 가 이항분포를 따른다면 $X \sim B(n, p)$ 로 표시한다.

: 이로부터 이항분포의 변수는 n 과 p 두 개가 유일함을 알 수 있다.

$$: P(X = r) =_n C_r p^r (1 - p)^{n-r}, r = 0, 1, 2, \dots, n$$

문제 3. 1. 이항정리를 이용하여 이항분포에서 모든 가능한 값이 나올 확률의 합이 1임을 보여라.

Note : 주어진 확률질량함수에 대해서, 그들의 합이 1임은 항상 당연하다.

또한, 이항분포가 베르누이 분포 n 개를 합친 것으로부터 다음을 알 수 있다.

확률변수 X 가 $B(n, p)$ 를 따르는 이항분포라고 하자. 그렇다면 i 번째 시행이 성공하면 $X_i = 1$, 실패하면 $X_i = 0$ 이라 할 때 $X = X_1 + X_2 + \dots + X_n$ 이라 표시할 수 있다. 따라서

$$E[X] = E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n] = np$$

$$V(X) = V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n) = np(1 - p)$$

문제 3. 2. 이항분포의 평균과 분산은 책에 나온 것처럼 콤비네이션을 잘 조작하여 증명해낼 수도 있다. 혹은, 2장 연습문제에서 다른 적률생성함수를 구함으로써 알아낼 수도 있다. 이 연습문제에서는 콤비네이션을 조작하여 기댓값과 분산이 각각 np 임과 $np(1 - p)$ 임을 보여라.

문제 3. 3. 책에서는 미분법을 이용한 증명도 소개하고 있다. 미분법을 이용하여 증명하여라.

위의 문제들에서 다른 방법은 계산이 꽤 복잡한 반면, 베르누이 분포를 이용하면 비교적 쉽게 증명해낼 수 있기에 이를 추천한다. 이때, 박스에서는 간소화하였지만 X_i 와 X_j 가 독립이라는 조건을 언급해주는 것이 매우매우 중요하다. 그것이 없다면 공분산을 무시해줄 수 없기 때문이다.

문제 3. 4. 이항분포 $B(n, p)$ 의 표준편차 σ 가 $\sqrt{np(1 - p)}$ 임을 보여라. 이항분포에서 n 만 알고 p 를 모를 때, 최대의 표준편차는 얼마만큼인가?

1.3 큰 수의 법칙

만약 야구나 축구, 배구나 농구와 같은 구기스포츠를 즐긴다면 캐스터가 '아 000선수 시즌 타율이 0.333 인데요... 3타석에 한 번은 안타를 친다 이건데요. 오늘 2타수 무안타입니다. 하나 칠때가 되었네요?'라고 말하는 걸 들어봄직하다. 사실 독립시행임을 고려하면 이는 도박사의 오류에 지나지 않지만, 어찌하였든 시행이 반복될수록 확률변수 X 의 값을 n 으로 나눈 값은 p 에 가까워질 것이라 추측할 수 있다. 즉, 이를 식으로 표시하면 충분히 작은 양수 h 에 대하여

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

이라는 것이다. 사실은, 이것이 이항분포만이 아니라 다른 확률분포에도 적용될 수 있다.

X_1, X_2, \dots, X_n 을 서로 독립이며, 동일한 확률분포를 가지는 확률변수라 하자. 또한, $E[X_i] = \mu$ 이며 X_i 의 분산은 유한하다. 그렇다면, 임의의 양수 $\varepsilon > 0$ 에 대하여,

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| > \varepsilon\right) = 0$$

이다. 우리가 본 것은 X_i 가 확률 μ 인 베르누이 분포였을 때이다.

pf) 2장에서 배운 것과 같이, 확률변수 $\frac{X_1 + X_2 + \dots + X_n}{n}$ 의 기댓값과 분산을 구해보자.

$$E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{n} = \mu$$

$$V\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{V(X_1) + V(X_2) + \dots + V(X_n)}{n^2} = \frac{\sigma^2}{n}$$

그러면, 여기서 채비셰프 부등식을 사용하면

$$P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

이므로 n 이 무한으로 감에 따라 해당하는 값은 항상 0으로 간다. 증명이 따라서 완료된다.

문제 3. 5. 채비셰프 부등식을 이용한 큰 수의 법칙의 증명에서, $\varepsilon = 1$ 이고 $\sigma = 1$ 일 때 확률

$$P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| > \varepsilon\right)$$

이 0.01보다 작음이 보장되려면 n 은 무엇보다 커야 하는가?

2 다른 이산확률분포들

여기서는 기하분포, 초기하분포, 푸아송분포에 대해 다룰 것이다. 하나 유념할 것은, 적률생성함수는 이들 모두에 사용 가능하지만 (심지어는 푸아송분포에 대해서는 적률생성함수가 훨씬 편리함에도!) 사용하지 않을 것이다. 범위에 들어가지 않으니 최대한 사용하는 것을 지양하고, 정말 궁지에 몰렸을 때만 몰래 사용하도록 하자.

2.1 기하분포

확률변수 X 의 확률질량함수 $P(X = n) = (1 - p)^{n-1}p$ ($n = 1, 2, 3, \dots$) 일 때, X 는 **기하분포**를 따른다고 하고, 이것을

$$X \sim GEO(p)$$

라고 나타낸다. X 는 일어날 사건이 p 인 사건이 처음으로 일어나기 위해 필요한 시행의 횟수를 의미한다고 해석할 수도 있다.

문제 3. 6. p 의 값에 관계없이, 기하분포의 확률질량함수는 확률질량함수의 조건을 만족시킴을 보여라.

우리가 보통 등비수열이라고 부른 수열을 영어로는 'geometric series'라고 부르는데, 이는 도형에서 닮음인 도형들을 계속 만들다 보면 닮음비가 공비인 등비수열이 나타나는 것으로부터 추측이 가능하다. 그렇기에 이처럼 확률질량함수가 등비수열인 이러한 분포도 기하분포라고 불리는 것이다. 우리는 등비수열에 대해서는 계산을 하는 데 큰 어려움을 겪지 않는다. 기댓값과 분산을 구하여 보자.

$$\begin{aligned} E[X] &= \sum_{n=1}^{\infty} np(1-p)^{n-1} \\ &= p \sum_{n=1}^{\infty} \frac{n(1-p)^{n-1} - n(1-p)^n}{p} \\ &= \sum_{n=1}^{\infty} (1-p)^{n-1} + ((n-1)(1-p)^{n-1} - n(1-p)^n) \\ &= \sum_{n=1}^{\infty} (1-p)^{n-1} + \sum_{n=1}^{\infty} ((n-1)(1-p)^{n-1} - n(1-p)^n) \\ &= \frac{1}{1 - (1-p)} = \frac{1}{p} \end{aligned}$$

사실 성공 확률이 p 인 것을 계속 시행하는 것이기에, 성공하기 위해 필요한 시행 수는 1을 그것으로 나눈 $1/p$ 임이 직관적이긴 하다. 증명을 물을 일은 거의 없을 것 같으니, 그렇게만 외워 두어도 된다. 사실 거의 모든 부분에 있어 증명보다는 평균/분산이 어떻게 나오는지를 기억해 두는 것이 중요하다. 적어도 통계학 시험을 보는 입장에서는.

$$\begin{aligned}
V(X) &= E[X^2] - (E[X])^2 \\
&= \sum_{n=1}^{\infty} n^2 p(1-p)^{n-1} - \frac{1}{p^2} \\
&= p \sum_{n=1}^{\infty} n^2 \frac{(1-p)^{n-1} - (1-p)^n}{p} - \frac{1}{p^2} \\
&= \sum_{n=1}^{\infty} (2n-1)(1-p)^{n-1} + \sum_{n=1}^{\infty} ((n-1)^2(1-p)^{n-1} - n^2(1-p)^n) - \frac{1}{p^2} \\
&= \frac{1}{p} \sum_{n=1}^{\infty} (2n-1)p(1-p)^{n-1} - \frac{1}{p^2} \\
&= \frac{1}{p} \left(\frac{2}{p} - 1 \right) - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}$$

분산도 위와 같이 구해질 수 있다. 미분법을 사용하는 것이 가장 편리하긴 하나, 거듭제곱급수의 미분을 염밀하게 다루려면 꽤 많은 지식이 필요하기에 여기에서는 지양하려 한다. 다만, 무한에 대한 어느 정도의 넘겨짚음은 용인하자.

가장 특이한 것은 기하분포의 **무기역성**이라는 것이다. 앞서 도박사의 저주에 대해 언급하였었는데, 타자가 안타를 치는 첫 타석이 기하분포로서 나타내진다는 점을 고려하면 독립시행은 꽤 크게 다가온다. 무기역성은 타자는 이전 타석에 대해 기억하지 않고, 매 타석을 첫 타석인 것처럼 다룬다는 것이다. 사실상 어떠한 선행조건이 없어도 항상 동일한 분포를 보인다는 점에서, 마치 확률변수의 독립을 떠올리게 한다. 독립이면 주변확률밀도함수로 결합확률밀도함수를 표현할 수 있음을 배웠는데, 이를 가지고 여러 가지 조작을 하는 데 매우 간편한 성질로서 작용한다.

책에서, $P(X \in A | X \in B)$ 와 같은 표기를 사용하고 있는데, 통계학에서 $|$ 는 조건부확률을 의미한다고 봐도 무리가 없다. 즉, 저것의 함의는 X 가 B 안에 있는 값으로 나왔을 때, X 가 A 안에 포함되어 있을 확률이며, 기준에 하던 계산과 동일하게 할 수 있다. 사실은 결합확률분포를 다룰 때 특정 Y 값에 대하여 X 값의 분포가 가지는 확률질량함수를 $P_{X|Y}(X = x)$ 와 같은 형태로 쓰기도 하는데, 이 책에서는 굳이 다루지 않겠다.

어찌하였든, 아래 식을 보면

$$\begin{aligned}
P(X > i+j | X > j) &= \frac{P(X > i+j)}{P(X > j)} \\
&= \frac{p(1-p)^{i+j}}{\frac{p}{p(1-p)^j}} \\
&= (1-p)^i = \frac{pq^i}{p} \\
&= P(X > i)
\end{aligned}$$

이다. 즉, $X \sim GEO(p)$ 일 때 무기역성이 성립한다.

문제 3. 7. 이대호 선수가 안타를 치기 위해 필요한 타석의 수 X 는 $GEO(0.3)$ 을 따른다고 한다. 오늘 이대호 선수가 앞선 두 타석에서 안타를 치지 못했고, 세번째 타석에 들어섰다. 이대호 선수가 네 번째 타석까지 소화한다고 할 때, 오늘 안에 안타를 칠 확률은?

2.2 초기하분포

확률변수 X 의 확률질량함수가

$$P(X = k) = \frac{{}^M C_k {}^{N-M} C_{n-k}}{{}^N C_n} \quad (k = 0, 1, \dots, n)$$

일 때, X 는 **초기하분포**를 따른다고 하고,

$$X \sim HYP(N, M, n)$$

과 같이 나타낸다. 이는 N 개의 공이 든 박스에서 원하는 공이 M 개 들어 있을 때, n 개의 공을 뽑았을 때 원하는 공이 포함된 개수라고 해석할 수 있다. 따라서 $N \geq M$ 임이 항상 가정되며, $n \leq M$ 임도 가정해야 한다. 그렇지 않으면 분포의 모양이 조금 달라질 수 있게 된다.

기하분포에서 '초'를 뜻하는 'hyper'가 붙게 된 분포이다. 기하분포에서는 복원추출, 즉 시행을 거듭해도 원하는 것을 얻을 확률이 p 로 항상 일정했지만, 초기하분포에서는 비복원추출을 시행하기에 시행을 거듭하면 할수록 그 결과에 따라 원하는 것을 얻을 확률 역시 달라진다.

문제 3. 8. 초기하분포의 확률질량함수가 확률질량함수의 조건을 만족시킴을 보여라.

이제 초기하분포의 평균과 분산을 구해보자.... 그러나 이는 계산이 너무 복잡하여, 증명하기에 매우 귀찮은 면이 있다. 외우고 다니는 것을 적극 추천한다.

$$\begin{aligned}
 E[X] &= \sum_{k=0}^n k \times \frac{M C_{kN-M} C_{n-k}}{N C_n} \\
 &= \sum_{k=0}^n k M C_k \frac{N-M C_{n-k}}{N C_n} \\
 &= \sum_{k=1}^n M_{M-1} C_{k-1} \frac{N-M C_{n-k}}{N C_n} \\
 &= M \sum_{k=1}^n \frac{M-1 C_{k-1 N-M} C_{n-k}}{N C_n} \\
 &= \frac{nM}{N} \sum_{k=1}^n \frac{M-1 C_{k-1 N-M} C_{n-k}}{N-1 C_{n-1}} \\
 &= n \frac{M}{N}
 \end{aligned}$$

여기서도 결국 원하는 공의 비율인 M/N 에 시행의 수 n 을 곱한 것이 기댓값으로 나온다. 따라서, 기댓값을 구하는 것은 생각보다 나쁘지 않다. 이제 분산을 구하여 보자.

$$\begin{aligned}
 V(X) &= E[X^2] - (E[X])^2 \\
 &= E[X(X-1)] + E[X] - (E[X])^2 \\
 &= \sum_{k=0}^n k(k-1) M C_k \frac{N-M C_{n-k}}{N C_n} + E[X] - (E[X])^2 \\
 &= M(M-1) \sum_{k=2}^n M-2 C_{k-2} \frac{N-M C_{n-k}}{N C_n} + E[X] - (E[X])^2 \\
 &= \frac{M(M-1)n(n-1)}{N(N-1)} \sum_{k=2}^n \frac{M-2 C_{k-2 N-M} C_{n-k}}{N-2 C_{n-2}} + E[X] - (E[X])^2 \\
 &= \frac{M(M-1)n(n-1)}{N(N-1)} + n \frac{M}{N} - n^2 \frac{M^2}{N^2} \\
 &= \frac{N-n}{N-1} \times n \left(\frac{M}{N} \right) \left(1 - \frac{M}{N} \right)
 \end{aligned}$$

즉 이항분포의 분산에, 계수 $\frac{N-n}{N-1}$ 을 곱한 것이다. n 은 항상 1 이상이기에, 분산은 이항분포보다는 작게 나옴을 알 수 있다. 이는 비복원추출의 특성에 의한 것으로, 이 계수를 유한수정계수라 부르기로 한다. 만약 N 이 충분히 커진다면 그것이 1로 수렴하기에, 공이 무한히 많아 뽑을 때마다 딱히 확률이 변하지 않는다면 이것이 이항분포로 근사될 수 있음을 의미한다.

문제 3. 9. X 의 분포가 $HYP(N, M, n)$ 일 때, i 번째 시행에서 원하는 공이 나오면 1, 아니면 0을 가지는 확률변수 X_i 를 생각하자. 그러면 $X_1 + X_2 + \dots + X_n = X$ 라고 생각할 수도 있을 것이다. 모든 i 에 대하여, $P(X_i = 1) = \frac{M}{N}$ 임을 보여 보아라.

2.3 푸아송 분포

가장 중요한 푸아송 분포의 차례이다. 여러분이 아마 공대에 진학한다면, 특히 컴퓨터이나 전기전자와 같은 과에 진학한다면 확률과정에 대해 이해하는 것은 매우 중요하다. 푸아송 분포는 푸아송 과정으로부터 얻어지는 분포 정도로만 지금은 알고 있자. 이산확률분포에서는 이항분포와 더불어 가장 많이 사용되는 분포 중 하나다. 음이항분포와 같은 다른 분포들은 연습문제에서 다루기로 한다. 푸아송 분포만 다루고 정규 내용은 마치자.

푸아송 과정은 아래의 조건들을 만족하는 것을 말한다.

- 1) 어떤 구간에서 발생하는 사건의 수는, 그것과 서로소인 다른 구간에서 발생하는 사건의 수와 독립이다.
- 2) 어떤 구간에서 사건의 평균 발생 횟수는 구간의 길이가 같으면 시작 시점에 관계없이 일정하고, 구간의 길이가 커지면 비례하여 커진다. 즉, 사건의 발생 횟수는 오직 구간의 길이에만 비례한다.
- 3) 어느 한 구간에서 두 개 혹은 그 이상의 사건이 발생할 확률은 구간의 길이가 충분히 짧아지면 0으로 수렴한다. 즉, 길이가 h 인 구간에서 발생하는 사건의 수를 $N(h)$ 라고 한다면,

$$\lim_{h \rightarrow 0} \frac{P(N(h) \geq 2)}{h} = 0$$

이다.

라고 언급을 하고 있으나, 사실은 중요한 두 가정을 조금 빼놓은 감이 있다. 우리는 $N(0) = 0$ 임과 미소한 시간 동안 사건이 발생하는 속도가 λ 라는 가정을 놓고 갈 것이다. 즉,

$$\lim_{h \rightarrow 0} \frac{P(N(h) = 1)}{h} = \lambda$$

까지 생각하자. 그렇다면, 우리가 해당하는 시간 t 를 n 개의 구간으로 나눈다면, 해당 구간에서 사건이 일어날 확률은 약 $\lambda t/n$ 일 것이며, 일어나지 않을 확률은 $1 - \lambda t/n$ 이다. 2회 이상 일어나는 것은 0으로 간다고 가정한 상태이다. 따라서, 총 일어나는 사건의 수가 k 개일 확률은

$${}_nC_k \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k}$$

일 것이다. 즉, 이항분포와 대략적으로 비슷한 분포를 띠고 있다는 것이다. 증명은 않겠으나, 만약 n 이 무한으로 가고, λt 를 시간 t 동안 평균 λ 회의 사건이 일어난다고 해석하면

확률변수 X 가

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

이라는 확률질량함수를 가진다면 이를 **푸아송분포**를 따른다 하며, $X \sim POI(\lambda)$ 라고 표시한다.

일반적으로, 일어날 확률은 적고 관찰 시간이 길은 경우에 푸아송분포가 매우 잘 근사한다. 이는 푸아송 과정의 가정들을 잘 생각해보면 참 잘 부합한다.

문제 3. 10. 푸아송분포의 확률질량함수가 그 정의에 잘 맞음을 보여라.

또한, 위에서 푸아송 분포가 이항분포로 근사될 가능성이 있음을 보였다. 반대로, 이항분포도 푸아송 분포로 근사가 가능하다. 콤비네이션을 계산하는 것보다 지수함수 값을 계산하는 편이 훨씬 낫기에, 이러한 근사는 꽤 효과적으로 이용된다.

확률변수 X 가 이항분포 $B(n, p)$ 를 따르고, n 이 충분히 크며 p 가 아주 작을 때, $B(n, p)$ 는 $POI(np)$ 로 근사될 수 있다. 일반적으로, $np < 5$ 를 기준으로 삼는다.

정확한 근사를 여기서 보이지는 않겠으나, 위에서 $\lambda/n = p$ 라고 둔 것이나 마찬가지였으므로 $np = \lambda$ 로 두는 것은 매우 타당해 보인다!

문제 3. 11. 책에 나온 대로 이항분포를 푸아송 근사 해보아라. 즉,

$$P(X = x) =_n C_x p^x (1 - p)^{n-x}$$

가

$$P(X = x) = \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

로 변형된 이후, n 이 무한으로 감에 따라

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

가 됨을 이용해 이것의 $\frac{e^{-\lambda} \lambda^x}{x!}$ 로 근사됨을 보이라는 것이다.

마지막으로 푸아송 분포의 평균과 분산을 구해 보자. 이때는 지수함수의 정의를 적극적으로 이용하면 좋다. 쉬우니, 증명은 연습문제로 남긴다.

$$X \sim POI(\lambda) \text{일 때},$$

$$E[X] = \lambda, V(X) = \lambda$$

특이한 점은, 평균과 분산이 같다는 것이다.

문제 3. 12. 위 박스를 증명하라.

3 연습문제

문제 3. 13. 인기 예능 신서유기에서는 n 명의 멤버를 데리고 90년대 노래 퀴즈를 진행한다. 각각이 정답을 맞출 확률은 p 이며, 각자의 결과에 독립적이다. 나PD는 최소 반 이상이 정답을 맞췄을 때 식사를 제공하기로 했다.

- 1) p 가 어느 값 이상이어야 5명의 멤버가 도전했을 때 3명의 멤버가 도전했을 때보다 더 성공 확률이 높을 것인가?
- 2) p 가 어느 값 이상이어야 $2k+1$ 명의 멤버가 도전했을 때 $2k-1$ 명의 멤버가 도전했을 때보다 효과적일 것인가?

문제 3. 14. 나PD는 다음 게임으로 n 명의 멤버들을 방에 불러 두고 그들의 모자를 모두 모아 섞었다. 그 다음, 연장자 순으로 무작위의 모자를 고르고 가져갔다. X 를 자신의 모자를 되찾은 멤버의 수라고 하자. $E[X] = 1$ 임을 보여라.

문제 3. 15. $X \sim POI(\lambda_1), Y \sim POI(\lambda_2)$ 이다. 그리고 이 두 확률변수는 독립이다. $X + Y$ 가 따르는 분포를 구하여라. 단, convolution은 가정하고 풀어도 좋다.

문제 3. 16. 가수 신대철은 자신의 기타 연주회에 N 명의 대중을 초청하는데, N 은 공교롭게도 평균이 λ 인 푸아송분포를 따른다. 그런데, 초청된 관객은 p 의 확률로 신대철보다는 그의 아버지인 신중현을 더 좋아하고, $1 - p$ 의 확률로 신대철을 더욱 좋아한다. 물론, 각 관객의 선호는 독립적이다. 각각의 수를 N_1, N_2 라고 하자. 즉, $N_1 + N_2 = N$ 이라고 하자. 그렇다면, N_1 과 N_2 역시도 푸아송 분포를 따름을 보이고, 그 기댓값을 구하여라.

문제 3.17. 문제 3.9.에서 우리는 초기하 분포를 n 개의 작은 확률변수로 나누어서 분석할 수 있음을 배웠다.
그를 가정하고, $i \neq j$ 일 때

$$E[X_i X_j] = \frac{(M)(M-1)}{(N)(N-1)}$$

임을 보여라. 이를 통해

$$Cov(X_i, X_j) = \frac{-M(N-M)}{N^2(N-1)}$$

임을 보여라.

문제 3. 18. 이어서, $V(X)$ 가 우리가 아는 것처럼 $\frac{N-n}{N-1} \times n\left(\frac{M}{N}\right)\left(1-\frac{M}{N}\right)$ 의 됨을 $E[X_i], V(X_i), Cov(X_i, X_j)$ 를 이용하여 보여 보아라.

문제 3. 19. X 가 이항분포 $B(n, p)$ 를 따른다고 하자. 아래를 보여라.

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

2) k 가 0에서 n 으로 감에 따라, $P(X = k)$ 의 값은 증가하다가 감소한다. 이때, 이 값이 최대가 되는, 즉 증가세에서 감소세로 돌아서는 k 의 값은 $(n+1)p$ 이하의 자연수 중 최대의 자연수이다.

문제 3. 20. 솔리데어 게임은 52개의 카드로 이루어진 텍을 잘 섞어 만든 카드 더미를 통해 이루어진다. 첫 카드를 뒤집기 전, 에이스라고 말한다. 두번째 카드를 뒤집기 전, 2라고 말한다. 같은 방법으로 계속 하되, 킹까지 모두 말한 이후, 즉 14번째 카드를 뒤집을 때는 다시 에이스부터 시작하여 반복하는 식으로 진행한다. 당신은 당신이 말한 카드와 실제 카드가 일치할 때 패배한다. 푸아송분포를 이용해 52개 카드를 모두 말할 동안 패배하지 않을 확률을 근사하라.

문제 3. 21. X 는 평균 λ 의 푸아송분포를 따른다. $P(X = i)$ 는 i 가 증가함에 따라 처음엔 증가했다가 그 이후 감소함을 보여라. 또한, 최댓값은 i 가 λ 이하의 자연수 중 가장 큰 것일 때 도달함을 보여라.

문제 3. 22. 배구선수 김연경의 리시브효율은 p 이며, 연달아서 서브를 계속 받고 있다. 또한, 각 서브리 시브의 성공확률인 p 는 이전/이후 서브 시도의 영향을 받지 않는다. 확률변수 X 를 성공까지 필요한 시도 횟수라고 하자. 즉, X 가 k 라는 말은 $k - 1$ 번의 리시브 실패 후 k 번째 서브를 잘 받아냈다는 것을 의미한다.

- 1) X 는 어떤 확률분포를 따르는가? 2) X 의 평균과 분산은 각각 얼마인가?

문제 3. 23. 평소 연습량이 많은 김연경 선수는 r 번의 성공을 얻기 전까지 리시브 훈련을 계속한다. 이처럼, 특정 성공 횟수를 얻기 위한 시행의 수를 Y 라고 할 때, Y 는 음이항분포를 따른다고 한다.

- 1) $k = r, r+1, \dots$ 에 대하여 $P(Y = k)$ 을 구하여라.

(Hint : 기하분포의 무기억성을 잘 기억해 보아라. $r-1$ 번째 사건 발생 이후 r 번째 사건이 발생하기 위해 필요한 시간과, 1번째 사건이 발생하기 위해 필요한 시간은 같은가? 다른가?)

- 2) $E[Y] = r/p$ 임을 보여라.

문제 3. 24. $X \not\sim B(n, p)$ 를 따른다고 하자.

- 1) Y 를 $B(n - 1, p)$ 를 따르는 이항확률변수라고 하자. $E[X^k] = npE[(Y + 1)^{k-1}]$ 임을 보여라.
- 2) X 의 기댓값과 분산을 구하되, 꼭 1)을 이용하여라.

문제 3. 25. X 가 $HYP(G + B, G, n)$ 을 따른다고 하자. 그러면 $G + B$ 가 충분히 커질 때, $B(n, \frac{G}{G + B})$ 로 근사 가능함을 보여라.

문제 3. 26. 두 확률변수 X, Y 에 대하여 항상 $Y > X$ 이고, 결합확률질량함수가

$$P(X = i, Y = j) =_j C_i e^{-2\lambda} \frac{\lambda^j}{j!}$$

이라고 정의하자.

- 1) Y 의 주변확률질량함수를 구하여라.
- 2) $E[X]$ 의 값을 구하여라.
- 3) $Z = Y - X$ 의 확률질량함수를 구하여라.

문제 3. 27. 호수에 총 b 마리의 물고기가 있다. 처음에 그물로 a 마리를 잡은 후에, 표식을 단 후 다시 호수에 방생하였다. 그 다음에 충분한 시간이 지나고 표식을 단 물고기를 m 마리 잡기 위해 다시 잡아야 할 물고기의 총 개체수를 X 라 하자. (단, $a << m \leq X << b$)

1) X 의 확률질량함수를 구하여라.

2) $E[X]$ 가 $m \frac{b+1}{a+1}$ 임을 보여라.

3) $V(X)$ 을 구하여라.

문제 3. 28. 멘델이 유전법칙을 발견하기 위해 완두콩을 이용하였는데, 완두콩에는 열성 형질인 노란색과 우성 형질인 초록색이 있다. 멘델은 자신의 법칙을 증명하기 위해 우성 중 순종의 비율이 $1/3$ 이 됨을 보이고자 했다. 멘델은 우성 표현형을 가진 2세대 완두콩 600개를 임의로 뽑은 후 자가수분을 통해 2세대 완두콩 각각에게서 10개의 3세대 완두콩을 만들고 10개 모두 우성인 경우 2세대 부모 완두콩이 순종이라 간주하였다. 실제로 600개 중 200개가 순종이었다고 하자.

- 1) X 를 우성으로 판단되는 2세대 부모 완두콩의 수라고 하자. X 의 확률질량함수를 구하여라.
- 2) $P(X \leq 201)$ 을 구하여라.

문제 3. 29. 동전을 던져 앞면이면 H , 뒷면이면 T 라고 표시해 n 번의 시행을 통해 문자열 w 를 얻자. 예를 들어, $w = HHTTT$ 이다. 그리고, X_n 을 해당하는 문자열에서 나온 앞면의 개수라고 하자.

- 1) X_5 의 확률질량함수를 구하여라.
- 2) X_n 의 확률질량함수를 구하여라. $P(X_n = k) = p_n(k)$ 라고 정의할 때, $p_n(k)$ 의 재귀식을 구하여라.

문제 3. 30. X 와 Y 가 독립이며, 동일한 성공 확률 p 를 가진 기하분포를 따른다. $P(X = k|X + Y = n)$ 은 무엇인가?

문제 3. 31. 확률변수 X, Y 가 유한한 기댓값을 가지고 있다.

$$E[\max(X, Y)] = E[X] + E[Y] - E[\min(X, Y)]$$

임을 보여라.

문제 3. 32. 다음을 보여라. 1)

$$\sum_{j=0}^k {}_{a+k-j-1}C_{k-j} \cdot {}_{b+j-1}C_j = {}_{a+b+k-1}C_k$$

2)

$$\sum_{r=0}^{m-1} {}_{n+m-1}C_{n+r} p^{n+r} (1-p)^{m-1-r} = \sum_{k=n}^{n+m-1} {}_{k-1}C_{n-1} p^n (1-p)^{k-n} = \sum_{j=0}^{m-1} {}_{n-1+j}C_{n-1} p^n (1-p)^j$$

문제 3. 33. X 와 Y 가 독립인 음이항분포를 따르는 확률변수라고 하자. 이들의 성공 확률은 p 로 동일하며, 원하는 성공 횟수가 각각 a, b 이다. 조건부 확률 $P(Y = b + j | X + Y = a + b + k)$ 을 구하여라. 앞선 문제의 결과를 적극 활용하여라.

문제 3. 34. 아래를 보여라.

$$1) {}_{n+m}C_r = \sum_{k=0}^r {}_nC_k {}_mC_{r-k}$$

$$2) {}_{2n}C_n = \sum_{k=0}^n ({}_nC_k)^2$$

문제 3. 35. 본페로니 부등식을 보여라.

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) \geq \sum_{i=1}^n P(E_i) - (n-1)$$

Note : 사실 이 문제는 여기에 오기에 적절한 문제는 아니지만, 아이디어가 어느 정도 비슷하여 넣어 보았다.

문제 3. 36. 골초로 알려진 가수 A 는 항상 두 개의 라이터를 들고 다닌다. 하나는 왼쪽 주머니에, 한쪽에는 오른쪽 주머니에 넣어 다닌다. 그가 연초에 불을 붙이려 할 때, 그는 두 주머니 중 하나를 골라 라이터를 사용한다. 각 라이터는 N 번 사용 가능하다고 하자. 그 가수가 자신의 한 라이터가 다 된 것을 발견한 순간, 나머지 라이터가 X 번 불을 붙일 수 있을 것이라 하자. $P(X = k) = p_k$ 라 하자. 단, 다 된 것을 발견하는 것은 라이터를 $N + 1$ 번 쓰려는 순간이라고 생각하자. 즉 $X = 0$ 일 가능성성이 있다.

- 1) $\sum_{k=0}^N p_k = 1$ 임을 보여라.
- 2) X 의 기댓값을 구하여라.

문제 3. 37. 성공률이 p 인 베르누이 독립시행을 한없이 행할 때, X_n 을 n 번째 시행에서 성공이면 1, 실패면 0인 확률변수라고 하자.

- 1) $X = X_1 + X_2 + \dots$ 일 때, $\lim_{n \rightarrow \infty} E[X] = \infty$ 임을 보여라.
- 2) $Y_n = X_n X_{n+1}$ 이라고 정의하자. $Y = Y_1 + Y_2 + \dots$ 라고 할 때, $E[Y] = \infty$ 임을 보여라.

문제 3. 38. 위의 문제에서, $2n - 1$ 번째 시행에서 처음으로 성공의 누적 횟수가 실패의 누적 횟수보다 커지는 사건을 C_n 이라고 하자. Z 는 $P(Z = x) = P(C_x)$ 인 이산 확률변수라고 하자. Z 의 확률밀도함수를 구하여라.

1 연속확률분포

1.1 연속확률변수와 확률밀도함수

연속확률변수 : 이산적인 양이 아닌 연속적인 양을 취하는 확률변수

구간 $[a, b]$ 에서 정의된 함수 $f(x)$ 가 다음 성질을 만족할 때, 이를 **확률밀도함수**라고 한다. 이에 대응되는 확률변수 X 는 $[a, b]$ 사이의 모든 실수 값을 취한다.

- 1) $f(x) \geq 0$
- 2) $\int_a^b f(x)dx = 1$
- 3) $P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx \quad (a \leq \alpha \leq \beta \leq b)$

그리고 이렇게 확률밀도함수를 가지는 확률변수 X 는 **연속확률분포**를 따른다고 말한다.

즉 확률밀도함수는 확률질량함수와 같이 해당하는 구간에 어느 정도의 확률로 확률변수 X 의 값이 나타나는지를 대응하는 함수이다. 하나 조심할 점은, $P(X = x)$ 는 $f(x)$ 와 다른 값이며, 0에 수렴한다는 점이다. 즉, 임의의 한 값을 취할 확률은 0이다.

1.2 누적분포함수

누적분포함수를 연속확률변수에 대해서도 정의할 수 있다. X 가 특정한 값 x 보다 작을 확률인 $P(X \leq x)$ 를 **누적분포함수**라 하고,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x)dx$$

로서 계산한다. 따라서 아래 성질이 만족함을 보일 수 있다.

- 1) $0 \leq F(x) \leq 1$
- 2) $F(x)$ 는 단조증가함수이다.
- 3) $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$
- 4) $\frac{d}{dx} F(x) = f(x)$
- 5) $P(a \leq X \leq b) = F(b) - F(a)$

이때 가장 아래의 두 성질은 미적분학 기본정리에 의해 성립한다.

문제 4. 1. 연속확률변수 X 의 확률밀도함수가

$$f(x) = ae^{-x}, \quad 0 \leq x$$

로 정의될 때, a 의 값을 구하여라.

문제 4. 2. 연속확률변수 X 의 누적분포함수가

$$F(x) = \frac{x}{1+x}, \quad 0 \leq x$$

로 주어질 때, $f(x)$ 를 구하여라.

1.3 연속확률변수의 평균과 표준편차

연속확률변수의 평균과 분산, 표준편차는 아래와 같이 정의된다.

$$\begin{aligned} E[X] &= m = \int_a^b xf(x)dx \\ V(X) &= \int_a^b (x-m)^2 f(x)dx = \int_a^b x^2 f(x)dx - m^2 \\ \sigma(x) &= \sqrt{V(x)} \end{aligned}$$

놀랍게도, 연속확률분포는 이것으로 마친다. 아래는 연습문제이다.

2 연습문제

문제 4. 3. 연속확률변수 X 에 대하여, $E[X]$ 가 존재한다고 가정하자.

$$E[X] = \int_0^\infty P(X > x)dx - \int_0^\infty P(X < -x)dx$$

임을 보여라.

문제 4. 4. X 가 연속확률변수라고 하며, $a \leq X \leq b$ 이고 $E[X] = \mu$ 라고 하자.

- (1) $a \leq \mu \leq b$ 임을 보여라.
- (2) $V(X) \leq \frac{1}{4}(b-a)^2$ 임을 보여라.

문제 4. 5. 연속확률변수 X 에 대해서도, 마코프 부등식과 채비셰프 부등식이 성립함을 보여라.

문제 4. 6. 소방서가 수직선 위에 있다. 불의 위치가 $f(x)$ 라는 확률밀도함수를 따라 분포해 있으며, $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$ 를 만족한다고 하자. 그렇다면, 불로부터 소방서 사이의 거리의 기댓값을 최소화할 수 있는 소방서의 위치는 어디인가?

문제 4. 7. 연속확률변수에 대해서도 적률생성함수를 만들 수 있다. $\phi(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ 라고 정의할 때,

$$\phi'(0) = E[X]$$

$$\phi''(0) = E[X^2]$$

임을 보여라.

문제 4. 8. 연속확률변수 X 가 확률밀도함수

$$f(x) = \begin{cases} cx^3, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

를 가진다고 하자. c 의 값을 구하여라.

문제 4. 9. X_1, X_2, \dots, X_n 은 독립인 확률변수이며, 그들의 확률밀도함수는 모두

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

라고 하자. $M = \max(X_1, X_2, \dots, X_n)$ 이라고 정의할 때, M 의 확률밀도함수를 구하여라.

문제 4. 10. 확률변수 X 가 확률밀도함수

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

를 따른다고 하자. 확률변수 $Y = X^3$ 의 확률밀도함수를 구하여라.

문제 4. 11. X 의 확률밀도함수가

$$f(x) = \begin{cases} a + bx^2, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

로 주어진다. $E[X] = 0.6$ 일 때, a, b 를 구하여라.

문제 4. 12. 한 음반이 발표되어 차트아웃 될 때까지의 시간은 아래와 같은 확률밀도함수 $f(x)$ 를 가지는 확률변수 X 로 표현된다.

$$f(x) = a^2 x e^{-ax}, \quad x \geq 0$$

$E[X]$ 를 a 로 표현하고, a 가 2일 때의 값을 구하라.

문제 4. 13. X 가 연속확률변수이고 누적분포함수 F 를 가질 때, 그 중간값은

$$F(m) = 0.5$$

인 m 으로 정의된다.

$f(x) = e^{-x}, x \geq 0$ 의 중간값을 구하여라.

문제 4. 14. 확률변수 X, Y 가 누적분포함수 F_X 와 F_Y 를 가지고 $a, b > 0$ 일 때,

$$F_X(x) = F_Y\left(\frac{x-a}{b}\right)$$

이때

- 1) $E[X]$ 를 $E[Y]$ 로서 표시하라.
- 2) $V(X)$ 를 $V(Y)$ 로 표현하라.

문제 4. 15. X_1, X_2, \dots 는 독립이며 동일한 연속확률분포를 가지는 확률변수의 수열이다. $N \geq 2$ 를

$$X_1 \geq X_2 \geq \dots \geq X_{N-1} < X_N$$

인 N 이라 정의하자. 즉, 감소를 멈추는 첫 점이라고 생각하자.

- 1) $P(N \geq n)$ 의 값을 구하여라.
- 2) $E[N] = e$ 임을 보여라.

문제 4. 16. 문제를 제거한다.

문제 4. 17. 확률변수 $R \stackrel{\text{의}}{\sim}$

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}$$

이라는 확률밀도함수를 가질 때 Rayleigh 분포라고 부른다. $E[R]$ 을 σ 를 이용해 표현하여라.

단, $\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$ 임을 이용하여도 좋다.

1 여러가지 연속확률분포

1.1 균등분포

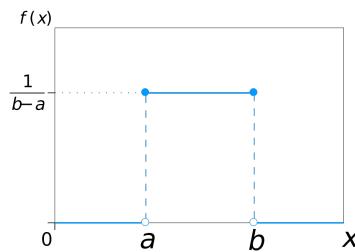
균등분포 : 연속확률변수가 어떤 실수 구간에서 일어날 확률이 구간의 길이에 비례하는 경우의 확률 분포

구간 $[a, b]$ 에서 정의된 연속확률변수 X 에 대하여, 확률변수는

$$f(x) = \frac{1}{b-a} (a \leq x \leq b)$$

로 나타내며, $X \sim U(a, b)$ 와 같이 나타낸다.

즉 아래와 같은 $f(x)$ 의 그래프를 가진다.



문제 5. 1. 연속확률변수 X 가 확률밀도함수를

$$f(x) = \frac{1}{b-a} (a \leq x \leq b)$$

로 가질 때, 누적분포함수 $F(x)$ 를 구하여라.

또한, 기댓값과 분산을 구하여 보면

$$E[X] = \int_a^b x f(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}$$

$$\begin{aligned} V(X) &= \int_a^b x^2 f(x) dx - \frac{a^2 + 2ab + b^2}{4} \\ &= \left[\frac{x^3}{3(b-a)} \right]_a^b - \frac{a^2 + 2ab + b^2}{4} \\ &= \frac{(b-a)^3}{12} \end{aligned}$$

로 나타내진다.

문제 5. 2. 규찬이와 규만이는 함께 콘서트장에 가려 하고 있다. 둘이 나오는 시각을 X, Y 라고 하면 X, Y 는 6과 7 사이의 균등분포이다. 즉, 6시부터 7시 사이에 둘이 집 밖으로 나오며, 그 분포는 균등분포 $U(6, 7)$ 을 따른다. 둘은 나온 이후 10분 동안 친구가 나오기를 기다리며, 10분 이후에도 나오지 않으면 그냥 먼저 출발해버린다. 규찬이와 규만이가 함께 콘서트장에 갈 확률은?

1.2 지수분포

지수분포는 앞서 배운 푸아송 분포와 매우 밀접한 관계를 맺고 있다. 앞서 푸아송 과정에 의하여, 시간 $(0, t)$ 동안 일어나는 사건의 발생 건수 N_t 는 $POI(\lambda t)$ 를 따름을 배웠었다. 푸아송 과정에서 어떤 사건이 처음으로 발생할 때까지 걸리는 시간의 확률분포 T 를 생각해보자. 누적분포함수 $F(t)$ 는

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= P(N_t \geq 1) \\ &= 1 - P(N_t = 0) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

이며, 이에 의해

$$f(t) = \frac{d}{dt} F(t) = \lambda e^{-\lambda t}$$

가 되다. 이와 같은 확률밀도함수를 갖는 확률분포를 **지수분포**라 하며, 기호로 $X \sim EXP(\lambda)$ 와 같이 나타낸다. 유념할 것은, 시간은 항상 양수이므로 $f(t)$ 는 t 가 양수일 때만 정의된다는 것이다.

지수분포의 평균과 분산을 구해 보면,

$$\begin{aligned} E[X] &= \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \\ V(X) &= \int_0^\infty (x - \frac{1}{\lambda})^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2} \end{aligned}$$

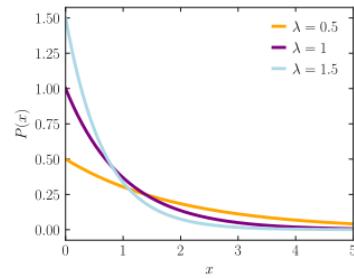
로 나타난다. 증명에서 많은 적분 계산이 생략되어 있다.

문제 5. 3. 위에서 빠진 적분 부분을 계산해 위의 결과가 맞음을 보여라.

기하분포에서 무기억성에 대해 다루었던 것처럼, 지수분포 역시 무기억성을 가진다. 지수분포가 푸아송 과정에서 사건과 사건 사이의 간격을 의미함을 고려하면, 푸아송 과정에서 단순히 구간의 길이가 발생 확률과 비례했으므로 지수분포는 지나간 시간과는 상관이 없다는 것을 알 수 있다. 이를 직접 계산해내면,

$$\begin{aligned} P(X > s + t | X > t) &= \frac{P(X > t \cap X > s + t)}{P(X > t)} \\ &= \frac{P(X > s + t)}{P(X > t)} \\ &= \frac{\int_{s+t}^{\infty} \lambda e^{-\lambda x} dx}{\int_t^{\infty} \lambda e^{-\lambda x} dx} \\ &= \frac{e^{-\lambda(s+t)}}{-e^{-\lambda t}} = e^{-\lambda s} = P(X > s) \end{aligned}$$

이다. 아래는 지수분포의 그래프이다.



문제 5. 4. 걸그룹이 1위를 하기 위해 걸리는 시간은 평균이 100인 지수분포를 따른다고 한다. 어느 걸그룹이 300일 동안 1위를 하지 못했을 때, 400일 안에 1위를 할 확률은 얼마인가?

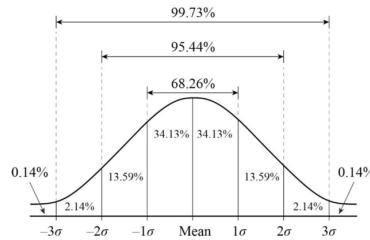
2 정규분포

2.1 정규분포의 소개

연속확률변수 X 가 $-\infty < x < \infty$ 의 값을 취하고 확률밀도함수 $f(x)$ 가

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

로 주어지는 확률분포를 **정규분포**라고 하며, 이때 평균은 m , 분산은 σ^2 이다. 이 경우 $X \sim N(m, \sigma^2)$ 으로 나타낸다. **정규분포곡선** : 확률분포함수 $f(x)$ 가 이루는 곡선



$f(x)$ 가 확률밀도함수임을 보여 보자. 이를 위해서는,

- 1) $0 \leq f(x)$
- 2) $\int_{-\infty}^{\infty} f(x)dx = 1$ 임을 보여야 한다. 그런데 1)은 매우 자명하므로, 2)를 보이면 되는데... 우리는 수업 전 이미 감마함수를 아주 살짝 이용하기로 했었다.

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \quad (y = \frac{x-m}{\sigma}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = 1 \end{aligned}$$

따라서, $f(x)$ 는 확률밀도함수로서 기능함을 알 수 있다.

문제 5. 5. 확률변수 X 가 $X \sim N(m, \sigma^2)$ 일 때, 그 확률밀도함수 $f(x)$ 가 m 을 기준으로 대칭임을 보여라. 이를 통하여, 평균이 0인 정규분포의 누적분포함수 $F(x)$ 에 대하여

$$F(x) = 1 - F(-x)$$

임을 보여라.

이제 가정한 바와 같이 평균과 분산이 각각 m , σ^2 임을 보이자. 여기서도, 감마함수를 아는 걸로 가정하고 진행하자.

$$\begin{aligned}
 E[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} xe^{-\frac{(x-m)^2}{2\sigma^2}} dx \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma y + m) e^{-\frac{y^2}{2}} dy \quad (y = \frac{x-m}{\sigma}) \\
 &= \frac{1}{\sqrt{2\pi}} \left(\sigma \int_{-\infty}^{\infty} ye^{-\frac{y^2}{2}} dy + m\sqrt{2\pi} \right) \\
 &= m + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ye^{-\frac{y^2}{2}} dy \\
 &= m \quad (ye^{-\frac{y^2}{2}} \text{ 는 기함수})
 \end{aligned}$$

$$\begin{aligned}
 V(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}} dx \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma y)^2 e^{-\frac{y^2}{2}} dy \quad (y = \frac{x-m}{\sigma}) \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}} dy \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \left([-ye^{-\frac{y^2}{2}}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \right) \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} (0 + \sqrt{2\pi}) = \sigma^2
 \end{aligned}$$

계산이 꽤 어렵지만, 치환적분에 매우 잘 적응된 상태라면 문제는 없다.

문제 5. 6. 시간이 된다면,

$$\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}$$

임을 보여라.

2.2 정규분포곡선의 성질

책에는 정규분포곡선으로부터 얻을 수 있는 성질 4개가 나와 있는데, 이는 사실 앞에서 동치인 명제들을 한 번씩은 보였거나, 그 자체가 정의인 경우가 많다. 4개는 아래와 같다.

1) $f(x) \geq 0$ 이며 $x = m$ 에 대해 대칭이고, 최댓값은 $\frac{1}{\sqrt{2\pi}\sigma}$ 이다.

2) x 축을 점근선으로 하며, 곡선과 x 축 사이의 넓이는 1이다.

3) m 이 일정할 때, σ 의 값이 작을수록 그래프의 폭은 좁고 높아지고, σ 가 일정하고 m 이 변하면 그래프의 모양은 그대로인 채 평균의 위치만 달라진다.

4) $P(\alpha \leq X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx$

사실 가장 중요한것은 아래처럼, 정규분포에서는 평균을 기준으로 표준편차에 의해 결정되는 구간이 특정 확률을 가진다는 점이다.

$$1) P(m - \sigma \leq X \leq m + \sigma) = 0.683$$

$$2) P(m - 2\sigma \leq X \leq m + 2\sigma) = 0.954$$

$$3) P(m - 3\sigma \leq X \leq m + 3\sigma) = 0.997$$

즉, 평균으로부터 2시그마정도 떨어진 구간에 약 95퍼센트가, 평균으로부터 3시그마 떨어진 구간에 약 99퍼센트가 있음을 의미한다.

문제 5. 7. 서울과학고 기초통계학 시험점수는 평균이 50, 표준편차가 10인 정규분포를 따른다고 한다. 시험에서 30점 이하인 학생의 비율은 대략 얼마인가?

2.3 표준정규분포

확률변수 X 가 정규분포 $N(m, \sigma^2)$ 를 따르고 그 누적분포함수를 $F(x)$ 라 할 때, $Z = \frac{X - m}{\sigma}$ 를 생각하자.

$$E[Z] = E\left[\frac{X - m}{\sigma}\right] = \frac{m - m}{\sigma} = 0$$

$$V(Z) = V\left(\frac{X - m}{\sigma}\right) = \frac{1}{\sigma^2}V(X) = 1$$

이고, $g(z)$ 라는 확률밀도함수와 누적분포함수 $G(z)$ 를 생각하게 된다면

$$\begin{aligned} G(z) &= P(Z \leq z) \\ &= P\left(\frac{X - m}{\sigma} \leq z\right) \\ &= P(X \leq m + z\sigma) = F(m + z\sigma) \end{aligned}$$

이며 양변을 z 에 대해 미분하면 표준정규분포의 확률밀도함수 $g(z)$ 는

$$\begin{aligned} g(z) &= \frac{d}{dz}G(z) \\ &= \sigma f(m + z\sigma) \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \end{aligned}$$

임을 알 수 있다. 즉 Z 는 평균은 0이고, 표준편차가 1인 정규분포 $N(0, 1)$ 을 따름을 알 수 있다.

표준정규분포 : $N(0, 1)$ 을 따르는 정규분포

정규화 : 정규분포 $N(m, \sigma^2)$ 을 $N(0, 1)$ 로 바꾸는 과정

표준정규분포표 : 표준정규분포에서 $P(0 \leq Z \leq x)$ 의 값을 표로 표시한 것

문제 5. 8. $N(10, 2)$ 를 따르는 정규분포 X 를 표준화하려면 Z 를 무엇으로 두어야 하는가? X 에 대해 표시하라.

2.4 이항분포의 정규분포 근사

확률변수 X 가 이항분포 $B(n, p)$ 를 따를 때, n 이 충분히 크면 X 는 근사적으로 $m = np$, $\sigma = \sqrt{np(1-p)}$ 인 정규분포 $N(np, np(1-p))$ 를 따른다. 이때, 일반적으로 $n \geq 30$ 이면 충분히 크다.

기초통계학 수준에서는, 이를 자세하게 다루지는 않는다. 근데 여기서 하나 유념할 것이 있다. 아래 문제를 보자.

문제 5. 9. 한 개의 주사위를 720회 던져서 1의 눈이 나오는 횟수를 X 라 할 때, $P(90 \leq X \leq 140)$ 의 값을 구하여라.

이 문제에서 $P(90 \leq X \leq 140)$ 을 $P(-3 \leq Z \leq 2)$ 로 조정하여 풀어야 할까? 그런데 문제는, 이항분포는 이산화률분포인 반면 정규분포는 연속화률분포라는 것이다. 따라서, 이산화률변수에서는 $P(X < x)$ 와 $P(X \leq x)$ 가 굉장히 큰 차이이다. $P(X = x)$ 가 0이 아니라면 무시할 수 없는 차이가 발생하기 때문이다. 반면, 연속화률변수에서는 $P(X = x)$ 는 항상 0이기에, 비율이 같다. 따라서 단순히 $-3 \leq Z \leq 2$ 으로 구간을 정한다면, $-3 < Z < 2$ 과도 확률이 같고 이는 $X = 90$ 과 $X = 140$ 은 제외하게 된다. 따라서 이런 문제를 방지하기 위해, 우리는 이 과정에서 **연속성 조정**을 수행한다. 즉, 90과 140을 안정적으로 표현할 수 있도록, $P(90 \leq X \leq 140) = P(89.5 \leq X \leq 140.5)$ 로 바꾸어 계산한다. 따라서, 계산하려면 $P(-3.05 \leq Z \leq -2.05)$ 를 구해야 한다.

문제 5. 10. 어느 가수의 곡이 성공할 확률은 0.3이라 한다. 이 가수가 25개의 곡을 냈을 때, 10곡 이상 히트했을 확률을 정규분포를 이용해 근사하시오.

3 연습문제

문제 5. 11. U 는 $[0, 1]$ 에서 정의된 균등확률분포를 가지는 확률변수이다. $a + (b - a)U$ 의 확률분포를 구하라.

문제 5. 12. 특정 전구는 일반적으로 수명이 평균이 2000이고 표준편차가 85인 정규분포를 따른다고 알려져 있다. 오직 5퍼센트만의 전구만이 L 보다 작은 수命을 가지게 하는 L 의 값을 구하시오. 단, $P(Z \geq z_\alpha) = \alpha$ 라고 정의한다. 또한, $z_{0.05} = 1.645$ 이다.

문제 5. 13. 기계를 고치는 데 필요한 시간은 $\lambda = 1$ 인 지수분포를 따른다고 한다.

- 1) 고치는 데 2시간 이상 걸릴 확률은?
- 2) 고치는 데 이미 2시간이 지났을 경우, 그 시점부터 3시간 안에 기계가 고쳐질 확률은?

문제 5. 14. 확률분포 X 는 만약 $\log X$ 가 정규분포를 이룬다면 로그정규분포라고 부른다. X 가 로그정규분포이고 $E[\log X] = \mu, V(\log X) = \sigma^2$ 라고 할 때, X 의 누적분포함수와 확률밀도함수를 구하여라.

문제 5. 15. 이차원 과녁에 총을 쏘 때, 수평 방향의 오차는 평균 0에 분산 4이며, 수직 방향의 오차도 동일한 평균과 동일한 분산을 가진다. 그리고 오차들은 모두 정규분포를 따른다고 한다. D 를 과녁으로부터 실제 총이 쏘인 부분 사이의 거리라고 둘 때, $E[D^2]$ 를 구하여라.

문제 5. 16. 신명호의 3점슛 성공률은 50퍼센트이며, 한 게임에서 그가 40번의 삼점슛을 시도하였다고 하자. 이때, 각 삼점슛의 성공 여부는 다른 삼점슛의 성공 여부와는 독립이다. 이때, 정규분포로의 근사를 통해 그가 3점슛으로 60점을 얻어낼 확률을 구하시오. 단, $P(0 \leq Z \leq \frac{1}{2\sqrt{10}}) = 0.0636$ 으로 둔다.

문제 5. 17. 주어진 정규분포 $N(m, \sigma^2)$ 의 확률밀도함수 $f(x)$ 에 대해서여, $\phi_{m,\sigma^2}(x) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ 라고 정의하자.

일반적으로, $\phi(x)$ 가 정해지면 그 분포 역시도 정해진다. 또한, X, Y 가 독립일 경우 $X + Y$ 의 $\phi(x)$ 는 X 의 $\phi(x)$ 와 Y 의 $\phi(x)$ 를 곱한 것임이 알려져 있다.

확률변수 X_1, X_2, \dots, X_n 이 정규분포 $N_1(m_1, \sigma_1^2)$ 부터 $N_n(m_n, \sigma_n^2)$ 을 따른다. 이때, $X_1 + X_2 + \dots + X_n$ 이 어떤 분포를 따르는지 이야기하고, 평균과 분산을 계산하라.

문제 5. 18. 나PD는 신서유기 프로그램에서 n 명의 멤버들에게 적어도 k 명이 지압마사지를 버티면 그동안 식사를 제공하는 k out of n 시스템을 도입하려 한다. 특히 $k = n$ 이면 직렬시스템, $k = 1$ 이면 병렬시스템이라 부른다. 각 멤버들이 지압마사지를 버티는 시간 T_i , $i = 1, 2, \dots, n$ 이 서로 독립이고 모두 $EXP(\lambda)$ 를 따를 때, 식사를 제공받는 시간 T_s 에 대하여 다음 물음에 답하라.

- (a) 직렬시스템의 T_s 의 기댓값을 구하라.
- (b) 병렬시스템의 T_s 의 기댓값을 구하라.

문제 5. 19. 서로 독립인 확률변수 U_1, U_2, \dots, U_n 이 모두 $U(0, 1)$ 을 따를 때, $X = \max(U_1, U_2, \dots, U_n), Y = \min(U_1, U_2, \dots, U_n)$ 이라고 정의하자. X 와 Y 의 확률밀도함수를 구하여라.

문제 5. 20. U_1, U_2, U_3 는 독립인 확률변수이며 모두 $U(0, 1)$ 을 따른다.

- 1) $U_1 + U_2$ 의 확률밀도함수를 구하라.
- 2) $U_1 + U_2 + U_3$ 의 확률밀도함수를 구하여라.

문제 5. 21. 임의의 점 P 가 수직선 위에 있으며, P 의 위치는 $\lambda = 1$ 인 지수분포를 따른다. 반면 점 U 는 구간 $[0, l]$ 에서 정의되는 균등분포를 따라 분포한다. P 와 U 로 수직선을 잘랐을 때, P 가 포함된 구간의 길이를 X 라 하자. $E[X]$ 의 값은? 단, P 와 U 의 위치는 독립이다.

문제 5. 22. 이 문제는 삭제하도록 한다.

문제 5. 23. $\{X_k\}$ 를 독립된 균등확률분포를 따르는 확률변수의 수열이라고 두자. a_k 가 양의 실수라면, $X_k \sim U(0, a_k)$ 를 따른다. $S_n = \sum_{k=1}^n X_k$ 라고 두며, $f_n(x)$ 와 $F_n(x)$ 를 그 확률밀도함수와 누적분포함수라고 부르자.

$$1) F_1(x) = \frac{\max(x, 0) - \max(x - a_1, 0)}{a_1} \text{임을 보여라.}$$

$$2) f_{n+1}(x) = \frac{F_n(x) - F_n(x - a_{n+1})}{a_{n+1}} \text{임을 보여라.}$$

3) $F_2(x)$ 를 구하여라.

4) $x^+ = \max(x, 0)$ 으로 정의하자. 만약 모든 a_k 의 값이 a 로 같다면,

$$F_n(x) = \frac{1}{n!a^n} \sum_{r=0}^n (-1)_n^r C_r [(x - ra)^+]^n$$

임을 보여라.

문제 5. 24. X_i 를 독립된 확률변수라고 두자. 이들이 각각 모수가 i 인 지수분포를 따를 때, $Z = \min(X_1, \dots, X_{10})$ 으로 정의하자.

- 1) $P(Z > 2)$ 의 값을 구하라.
- 2) $V(Z)$ 의 값을 구하여라.

문제 5. 25. $F(x) = \frac{e^x}{e^x + e^{-x}}$ 의 누적분포함수임을 보이고, 확률밀도함수를 찾아라.

1 모집단과 표본

1.1 모집단과 표본

전수조사 : 통계조사의 조사대상 전체에 대해 조사하는 것
표본조사 : 자료의 일부를 조사하여 전체의 분포를 추측하는 것
모집단 : 조사의 대상이 되는 전체
표본 : 통계의 자료로 선정된 것으로, 그 개수를 **표본의 크기**라 부름
임의추출 : 모집단 내에서 임의로 표본을 추출하는 과정, 기준 없이 정말 무작위로 표본을 추출해내는
단순임의표집법이 일반적임
임의표본 : 임의추출을 통해 만들어진 표본

1.2 표본평균의 분포

모평균 : 모집단의 평균
모분산 : 모집단의 분산
모표준편차 : 모집단의 표준편차

그런데 정해져 있는 모집단에서 표본을 뽑는다고 생각하면, 추출마다 뽑히는 표본은 다를 것이다. 따라서 이는 확률적으로 표본의 값이 달라진다고 해석할 수 있다. 이런 경우에는 우리는 각 표본을 X_i 라는 확률변수에 대응시킬 수 있다. 또한, 복원추출하는 경우 그 분포는 모집단의 분포를 따른다고 생각할 수 있다.

모집단에서 크기 n 인 표본을 추출하는 경우 추출된 n 개의 변량을 X_1, X_2, \dots, X_n 이라 할 때, 이들의 평균

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

을 **표본평균**이라고 하며, 이 역시 확률변수이다.

표본평균은 n 개의 유사한 분포를 따르는 확률변수의 평균이라고 생각할 수도 있다. 먼저, 복원추출하는 경우의 표본평균의 기댓값과 분산을 알아보자.

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n}E(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n}(E(X_1) + E(X_2) + \dots + E(X_n)) \\ &= m \end{aligned}$$

$$\begin{aligned} V(\bar{X}) &= \frac{1}{n^2}V(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n^2}(V(X_1) + V(X_2) + \dots + V(X_n)) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

즉, 평균은 동일하며 분산은 표본의 크기로 나눠짐을 알 수 있다. 이것이 자연스러운데, 왜냐하면 큰 수의 법칙에 의하여 표본평균이 평균에 점점 가까워지며 평균과 가까울 확률이 1에 수렴함을 보였기 때문이다. 즉, 분산은 당연히 작아질 수밖에 없다. 큰 수의 법칙은 이에 대한 직관을 어느 정도 제공한 것이라 볼 수 있다. 문제는 비복원추출에서는 뽑을 때마다 분포가 다르다는 것이다. 그런데, X_1 와 X_2 는 사실 분포가 같다.

그러나 X_2 의 분포가 X_1 에 의존하기 때문에, $V(X)$ 를 계산할 때 Cov 항을 생략할 수가 없다. 이미 사실 우리는 초기하분포를 다를 때 이와 비슷한 것을 수행해본 적이 있다. 여기서는 조건부 확률변수를 사용해야 계산이 가능한데, 우리는 배우지 않았기에 결과만 작성한다.

비복원추출하는 경우에는 모집단의 크기 N 에 대해

$$E(X) = m$$

$$V(X) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

이때, N 이 커짐에 따라, 즉 모집단의 크기가 큼에 따라 앞선 차례에서 뽑은 것이 뒷차례에 영향을 미치지 않게 되면 이는 복원추출과 다를 바가 없게 된다. 따라서, 그 경우에는 분산이 앞선 경우와 같이 $\frac{\sigma^2}{n}$ 으로 나타난다.

중심극한정리

모집단이 정규분포 $N(m, \sigma^2)$ 일 경우, 표본평균

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

은 정규분포 $N(m, \frac{\sigma^2}{n})$ 을 따른다.

또한, n 이 충분히 커진다면 모집단의 분포에 상관없이 $\bar{X} \sim N(m, \frac{\sigma^2}{n})$ 이다.

라고 언급이 되어 있지만, 사실 정확한 의미의 중심극한정리는 모집단이 정규분포라는 가정이 없이도, n 이 충분히 커지면 표본평균의 분포가 정규분포를 따른다는 것을 의미한다. 모집단이 정규분포일 경우에는 표본평균이 정규분포를 따름을 극한의 개념 없이도 증명할 수 있다. 사실 이는 5단원 연습문제에서 이미 다룬 적이 있다.

문제 6. 1. 확률변수 X 가 $POI(3)$ 을 따를 때, 크기가 16인 표본평균 \bar{X} 의 분산은 얼마인가?

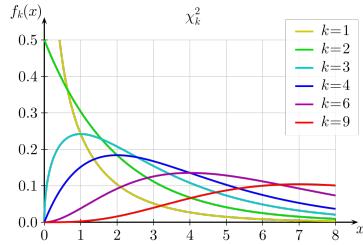
문제 6. 2. 전국 고등학교 학생들의 모의고사 점수는 평균이 60점이고, 표준편차가 15점이라고 한다. 고등학생 중 100명을 임의로 추출해 표본평균을 구했을 때, 그 값이 63점 이상일 확률을 구하여라.

2 표본분산의 분포

2.1 카이제곱분포

카이제곱분포 : 확률변수 Z_1, Z_2, \dots, Z_k 가 각각 표준정규분포 $N(0, 1)$ 을 따르고 서로 독립일 때, $Z_1^2 + Z_2^2 + \dots + Z_k^2$ 의 분포를 자유도 k 인 카이제곱분포라고 한다. 기호는 χ 를 사용하여 $\chi^2(k)$ 로 나타낸다.

그런데, 확률밀도함수는 감마함수를 포함한다. 이는 카이제곱분포가 사실은 감마함수를 기반으로 하는 감마분포의 일종이기 때문인데, 우리는 배우지 않았으므로 아예 생략한다. 카이제곱분포의 곡선은 자유도에 따라 결정되며, Z 들은 0 근처의 값을 주로 가지기에 $\chi^2(k)$ 는 오른쪽으로 꼬리가 긴 분포를 가진다.



이때, $0 \leq \alpha \leq 1$ 에 대하여, $\chi_{k,\alpha}^2$ 는 자유도가 k 인 카이제곱분포에서 $100(1-\alpha)$ 백분위수에 해당하는 값을 의미한다. 즉 그 값에서 무한대까지 확률밀도함수를 적분하면 그 값이 α 라는 것과 매한가지이다.

문제 6. 3. $\chi_{k,\alpha}^2 > \chi_{k,1-\alpha}^2$ 가 성립하는 α 의 범위를 구하여라.

카이제곱분포에 대하여 $X \sim \chi^2(k)$ 일 때, 평균과 분산은

$$E(X) = E(Z_1^2 + Z_2^2 + \dots + Z_k^2) = k(\sigma^2 + m) = k$$

$$V(X) = 2k$$

이다.

문제 6. 4. $V(X) = 2k$ 임을 보여라.

그 다음 카이제곱분포의 성질을 확인하여 보자.

- 1) $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$ 이고 V_1 과 V_2 가 서로 독립이면 $V_1 + V_2 \sim \chi^2(k_1 + k_2)$
- 2) $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$ 이고 $k_1 > k_2$ 이며 $V_1 = V_2 + V_3$ 이라 표시된다면, V_2 와 V_3 이 독립이면 $V_3 \sim \chi^2(k_1 - k_2)$ 이다.

이는 감마함수와 이전에 배웠던 적률생성함수를 응용하면 증명할 수 있다. 그러나 여기서는 다루지 않겠다. 하지만, 이것을 굳이 다루는 이유는 표본분산을 이해하는 데 필요하기 때문이다.

2.2 표본분산의 분포

표본분산이라는 것은 X 들을 하나의 자료로 보고, 그들의 펴진 정도를 파악하고자 하는 것이다. 즉 n 개의 자료에 대해서 편차의 제곱의 평균을 분산으로 정의하였던 것과 유사한 방법과 같이 이를 정의한다.

표본분산 : 표본으로 얻어지는 자료의 분산 S^2

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

이때 $n-1$ 으로 나누는 것이 굉장히 특이한데, 우리가 1단원에서 분산을 구할 때는 n 으로 나눴기 때문이다. 표본분산은 원래 모집단의 분산을 추측하기 위해서 사용하는 값이다. 따라서 표본분산의 기댓값은 원래 모집단의 분산 σ^2 과 같아야 할 것이다. 즉, 불편(unbiased)여야 하기에 $n-1$ 로 나누어야 한다는 것이다.

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m + m - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n [(X_i - m)^2 + 2(X_i - m)(m - \bar{X}) + (m - \bar{X})^2] \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m)^2 + \frac{2(m - \bar{X})}{n-1} \sum_{i=1}^n (X_i - m) + \frac{n}{n-1} (m - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m)^2 - \frac{n}{n-1} (m - \bar{X})^2 \end{aligned}$$

이므로,

$$E[S^2] = \frac{1}{n-1} n \sigma^2 - \frac{n}{n-1} \frac{\sigma^2}{n} = \sigma^2$$

이 될 것이다.

문제 6. 5. 위의 등식에서는 빠진 부분이 조금 있다. 사이사이를 채워넣어라.

모집단이 모분산 σ^2 인 정규분포를 따를 때 크기가 n 인 표본을 추출하면, $\frac{(n-1)S^2}{\sigma^2}$ 는 자유도가 $(n-1)$ 인 카이제곱분포를 따른다.

문제 6. 6. $\frac{X_i - m}{\sigma}$ 가 어떤 분포를 따르는지 말하라.

문제 6. 7. $\sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2$ 가 어떤 분포를 따르는지 말하라.

문제 6. 8. $\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$ 가 아래같이 표현됨을 보여라.

$$\sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 - \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2$$

문제 6. 9. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 임을 보여라.

3 연습문제

문제 6. 10. 서울과학고 기초통계학 시험의 점수는 평균이 500이고 표준편차가 100인 정규분포를 따른다.
만약 5명의 학생이 임의로 골라진다면,

- 1) 모든 학생의 점수가 600 이하일 확률은?
- 2) 그들 중 정확히 3명만 640점 이상일 확률은?
- 3) 5명의 평균점수가 600점 이하일 확률은?
- 4) 5명의 평균점수가 560점 이하일 확률은?

문제 6. 11. 방탄소년단의 연간 스트리밍 수는 평균이 40.14억이고 표준편차가 8.7억인 정규분포를 따른다고 하자.

- 1) 올해 신곡의 스트리밍 수가 42억을 넘을 확률은?
- 2) 차후 2년간 스트리밍 수가 84억을 넘을 확률은?
- 3) 차후 3년간 스트리밍 수가 126억을 넘을 확률은?

문제 6. 12. 양궁 국대 기보배는 엑스텐(과녁의 정중앙을 맞추는 것)을 노리고 있다. 이때, 수직 오차는 평균이 0이고 분산이 4인 정규분포를 따르며, 수평 오차는 수직 오차와는 독립이며 평균이 0이고 분산이 4인 정규분포를 따른다. D 는 화살이 맞은 점과 정중앙 사이의 거리라고 할 때, $E[D^4]$ 의 값은?

문제 6. 13. X 와 Y 는 서로 독립인 카이제곱분포를 따르는 확률변수이며, 각각의 자유도는 3과 6이다. $X+Y$ 의 값이 10보다 클 확률을 구하여라. $\chi^2_{0.35}(9) = 10$ 이다.

문제 6. 14. 10개의 주사위가 굴려져서, 나온 값의 총합이 30과 40 사이일 확률을 근사하여라.

문제 6. 15. 16개의 서로 독립인 균등분포 $U(0, 1)$ 을 따르는 확률변수들의 합이 10보다 클 확률을 근사하시오.

문제 6. 16. 50개의 숫자들의 가장 가까운 정수로부터의 차이가 -0.5 와 0.5 사이에서 균등분포를 이룰 때,
50개를 그냥 합했을 때와 반올림해서 합했을 때의 차이가 3 이상일 확률을 근사하라.

문제 6. 17. 케이티엔지는 그들의 담배에 포함된 니코틴의 양이 평균 $2.2mg$ 에 표준편차가 $0.3mg$ 인 정규
분포를 따르는 확률변수라고 주장하였다. 그러나, 100개의 담배를 모은 다음 얻은 표본평균은 $3.1mg$ 였다.
만약 담배공사의 주장이 옳을 때, 표본평균이 $3.1mg$ 이상으로 나올 확률은 얼마인지 근사하라.

문제 6. 18. 기초통계학 선생님께서는 오랜 경험을 통해 시험 점수가 평균 77점에, 표준편차 15점임을 알아냈다. 이 교사는 현재 두 반을 가르치고 있는데, 한 반은 25명이고 한 반은 64명이다.

- 1) 25명 분반에서 평균 시험 점수가 72와 82 사이일 확률을 근사하라.
- 2) 64명 분반에 대해서는 어떤가?
- 3) 25명 분반의 평균 점수가 64명 분반의 평균점수보다 높을 확률은?
- 4) 두 분반의 평균 점수가 순서 상관 없이 76점과 83점이었다. 두 분반 중 83점일 가능성성이 더 높은 분반은 어디인가?

문제 6. 19. 어떤 나라의 풋살리그는 두 리그로 나누어져서 운영된다. 한 팀은 총 60게임을 치루며, 32 게임은 같은 그룹에 있는 팀과, 28게임은 다른 그룹에 있는 팀과 한다고 가정해보자. 설곽FC는 같은 그룹에 속한 팀에게는 0.5의 확률로, 다른 그룹에 속한 팀에게는 0.7의 확률로 승리한다. X 를 해당 시즌의 총 승리 수라고 하자.

- 1) X 는 이항분포를 따르는가?
- 2) X_A 와 X_B 는 각각 같은 리그, 다른 리그와 경기한 경우의 승리 수라고 정의하자. X_A 와 X_B 의 분포는 어떠한가?
- 3) X_A , X_B 와 X 의 관계는?
- 4) 설곽FC가 40게임 이상 승리할 확률을 근사하라.

문제 6. 20. 대형 시계를 구성하는 부품 중 톱니바퀴A는 구동에 매우 중요하여 고장나면 바로 바꿔 주어야 한다. 이 톱니바퀴의 평균 수명은 100시간이며 표준편차가 30시간이다. 2000시간 동안 0.95 이상의 확률로 구동시킬 수 있으려면, 최소 몇 개의 톱니바퀴A가 준비되어 있어야 하는가?

문제 6. 21. X_1, X_2, X_3 은 평균이 120이고 분산이 9인 정규분포를 따르는 모집단에서 뽑힌 표본이다.

- 1) $P(X_1 + X_2 + X_3 > 35)$ 은?
- 2) $P(\bar{X} > 14)$ 는?
- 3) $P(\min(X_1, X_2, X_3) < 9)$ 은?
- 4) $P(\max(X_1, X_2, X_3) < 15)$ 은?

문제 6. 22. X_1, X_2, \dots, X_{25} 는 평균이 3이고 분산이 100인 정규분포에서 뽑은 표본이다. $P(0 < \bar{X} < 4, 56.2 < S^2 < 164)$ 는? 단, 표준정규분포 Z 에서 $P(Z > z_\alpha) = \alpha$, 자유도가 k 인 카이제곱분포를 따르는 확률변수 X 에 대해 $P(X > \chi^2_\alpha(k)) = \alpha$ 라고 정의한다.

단, 표본평균과 표본분산은 독립이라고 가정한다.

1 추측통계학과 추정

1.1 점추정

모평균의 점추정 : 모평균의 점추정값은 표본평균이다.

모비율의 점추정 : 모비율의 점추정값은 표본비율이다.

모분산의 점추정 : 모분산의 점추정값은 표본분산이다.

점추정이라 함은 우리가 모르는 모집단의 성질을 표시하기 위해, 표본으로부터 얻어내어 사용할 수 있는 값이다. 불확실성은 포함하지 않으며, 주로 해당 결과가 나올 확률이 가장 높은 값을 이용한다던가(maximum likelihood estimator), 기댓값이 모집단의 그것과 같은 값(unbiased estimator)을 사용한다. 그러나 이 책에서는 위의 박스에 나온 내용만 알고 있으면 상관없다. 보통, 점추정값은 구간추정을 하는 데 중요한 요소로서 작용한다.

2 구간추정

2.1 모분산을 알고 있을 때의 모평균 추정

모집단이 평균 m , 분산이 σ^2 인 정규분포를 따른다면 크기가 n 인 표본의 표본평균 \bar{X} 의 분포는 평균 m , 분산이 $\frac{\sigma^2}{n}$ 인 정규분포를 따른다는 것을 알고 있다. 이로부터 아래의 식들을 유도할 수 있다.

$$P(m - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq m + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

이로부터 우리는 σ 를 알고 있을 때 이런 방식으로 구한 구간

$$(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$$

중 95퍼센트가 모평균 m 을 포함하고 있음을 의미한다. 여기서, 해당 구간을 모평균의 **95% 신뢰구간**이라 부른다. 이때, 1.96은 $z_{0.025}$, 혹은 $Z(\frac{0.05}{2})$ 라고 표시할 수 있으며, 이들은 그동안 사용했던 대로 그 이상의 값을 가질 확률이 0.025인 점을 의미한다.

모집단이 정규분포이고 모분산 σ^2 을 알고 있을 때, 모평균 m 의 $100(1-\alpha)\%$ 신뢰구간은 아래와 같다.

$$\left[\bar{X} - Z(\frac{\alpha}{2}) \frac{\sigma}{\sqrt{n}}, \bar{X} + Z(\frac{\alpha}{2}) \frac{\sigma}{\sqrt{n}} \right]$$

이때, α 는 신뢰구간 100개를 만들면 평균적으로 100 α 개가 실제 모평균을 포함하지 않는다는 것을 의미하게 된다.

문제 7. 1. $Z(1-\alpha)$ 를 $Z(\alpha)$ 로써 표시하라.

문제 7. 2. $Z(\frac{\alpha}{2})$ 와 $Z(\alpha)$ 중 어느 것이 더 큰지 확인하라.

2.2 모분산을 모를 때의 모평균 추정

앞페이지에서의 방법은 모분산을 알 때의 방법인데, 실제로는 그 분산이 얼마나 되는지 알기란 쉽지 않다. 만약 n 이 충분히 커진다면 S^2 의 값이 σ^2 의 값과 충분히 가까워질 것이기에 문제가 없으나, n 이 작을 때는 이런 문제를 해결해야만 한다.

t분포 : 표준정규분포를 따르는 확률변수 Z 와 자유도가 r 인 카이제곱분포를 따르는 확률변수 U 에 대하여, Z 와 U 가 독립일 때 새로 정의된 확률변수

$$T = \frac{Z}{\sqrt{U/r}}$$

는 자유도 r 인 t분포를 따른다고 이야기하며, $T \sim t(r)$ 로 표시한다.

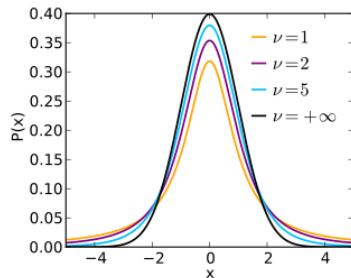
또한, 증명하기에는 어려우나

$$E(T) = 0 \quad (r \geq 2)$$

$$V(T) = \frac{r}{r-2} \quad (r \geq 3)$$

이며, t분포는 r 이 커질수록 기댓값이 0, 분산이 1에 가까워진다는 것을 알 수 있다. 이로부터 예상할 수 있듯이, t분포는

- 1) 평균이 0이고 0을 중심으로 좌우 대칭인 분포를 가진다.
- 2) 표준정규분포와 비슷한 형태의 분포이지만, 표준정규분포보다 봉우리는 낮고 꼬리가 두껍다.
- 3) 자유도가 커질수록 표준정규분포와 유사하다.



여기서도 동일하게, 자유도가 n 인 t분포에서 그것보다 큰 수들의 비율이 α 인 점을 $t_{n,\alpha}$ 라고 정의한다.

문제 7. 3. $t_{n,1-\alpha}$ 을 $t_{n,\alpha}$ 을 이용하여 표시하라.

$X_1, X_2, \dots, X_n \mid N(m, \sigma^2)$ 을 따르는 모집단으로부터 얻어진 n 개의 표본일 때, 확률변수

$$T = \frac{\bar{X} - m}{S/\sqrt{n}}$$

은 자유도 $n - 1$ 인 t 분포를 따른다.

증명은

$$Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$$

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

임을 고려하고, 이들이 독립이기에

$$T = \frac{\bar{X} - m}{S/\sqrt{n}} = \frac{\frac{\bar{X} - m}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{Z}{\sqrt{U/(n-1)}} \sim t(n-1)$$

로 정리할 수 있음을 알면 끝난다. 동일한 방식으로 구간추정을 해보면,

$$P(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}) = 1 - \alpha$$

$$P\left(-t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \leq \bar{X} - m \leq t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \leq m \leq \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

이므로 100개의 신뢰구간을

$$\left[\bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}\right]$$

위와 같은 형식으로 만들었을 때 평균적으로 α 개를 제외하고는 모평균을 포함하게 된다. 즉, $100(1 - \alpha)\%$ 신뢰구간이다.

문제 7. 4. 가수 장혜진의 앨범에서 임의추출된 9개의 곡 길이가 평균 240초이고, 분산이 100이라고 한다. 곡 길이의 표준편차를 모른다고 할 때,

- 1) 곡 길이 평균의 점추정값은?
- 2) 곡 길이 평균의 95% 신뢰구간을 구하여라.

2.3 모비율의 구간추정

모집단에서 특정 속성을 가지는 원소의 비율을 **모비율**이라 하는데, 모비율은 특정 사건이 일어나면 1, 일어나지 않으면 0인 확률변수 여러 개가 더해져서 만들어진 것이라고 생각하면 모평균과 다른 점이 거의 없게 된다. $X_1 + X_2 + \dots + X_n = X$ 라면 $X \sim B(n, p)$ 를 따를 것이며, 이때의 관측값 $\frac{X}{n} = \hat{p}$ 를 **표본비율**이라 부른다.

표본의 개수가 n 이고 특정 조건을 만족하는 표본의 수를 X 라 할 때, 표본비율 $\hat{p} = \frac{X}{n}$ 은 n 이 충분히 클 경우

$$N(p, \frac{p(1-p)}{n})$$

을 따르게 된다.

특히, 이를 표준화하면

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim Z$$

이다.

그 증명은 이항분포가 정규분포로 근사될 수 있다는 것을 생각하면 큰 문제가 없이 가능하다.

그런데, 우리가 모비율인 p 를 알고 있다면 추정값을 이미 알고 있는 것이므로 의미가 없다. 즉, 우리는 보통 p 를 모르는 상태에서 관측을 통해 이를 보아야 하는데, 정규분포에서 분산은 평균과 관련없던 것과 달리 여기서는 분산이 p 에 의존하는 값이다. 따라서, 여기서는 n 이 충분히 크고 \hat{p} 가 p 와 거의 같다는 가정 하에

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim Z$$

라고 본다.

표본의 크기가 충분히 큰 경우, 모비율 p 의 $100(1 - \alpha)\%$ 신뢰구간은

$$\left[\hat{p} - Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

임 역시 어렵지 않게 증명할 수 있다.

문제 7. 5. 동욱이는 자신이 하고 있는 게임에서 '기초통계학의 검'이라는 아이템을 뽑고 싶어한다. 동욱이가 100번 뽑았을 때, 기초통계학의 검은 30번 등장하였다. 기초통계학의 검 등장 확률의 95% 신뢰구간을 구하여라.

3 모분산의 추정

3.1 모분산의 구간추정

앞서, 모집단이 정규분포를 따르는 경우

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

임을 이야기했었다. 그렇다면,

$$P(\chi_{n-1,1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}^2) = 1 - \alpha$$

$$P\left(\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2}\right) = 1 - \alpha$$

이므로 σ^2 의 $100(1 - \alpha)\%$ 신뢰구간은

$$\left[\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} \right]$$

로 주어지게 된다.

문제 7. 6. 걸그룹 여자친구의 곡별 BPM은 정규분포를 따른다고 한다. 25개의 활동곡을 골라 BPM을 계산했더니, 분산값이 0.32였다. 모표준편차 σ 의 95% 신뢰구간을 구하여라.

4 연습문제

문제 7. 7. T 가 자유도가 8인 t 분포를 따른다고 할 때, 표를 이용하여

- 1) $P(T \geq 1)$
 - 2) $P(T \leq 2)$
 - 3) $P(-1 < T < 1)$
- 의 값을 구하여라.

문제 7. 8. 2013년 8월, 뉴욕타임즈는 오바마 대통령의 국정수행에 대한 긍정률이 50퍼센트라고 보도하였으며, 95퍼센트 신뢰도에서 오차는 $\pm 4\%$ 이하라고 주장하였다. 표본 크기의 최솟값을 구하여라.

문제 7. 9. 앞선 문제와 같이, 모비율의 구간추정에서 신뢰도가 $1 - \alpha$ 일 때 오차가 $\pm b\%$ 이하가 되기 위해 필요한 표본의 수를 결정하길 원한다. 표본비율에 대한 확신이 없이 계획을 수립하려 할 때, 표본 크기 n 은 최소 얼마여야 하는가?

문제 7. 10.

$$P(Z \leq z_\alpha) = \alpha$$

임을 이용하여, 표본평균이 \bar{X} 이고 모분산이 σ^2 이라 알려져 있을 때 95%의 신뢰도로 모평균을 포함하도록, 구간의 오른쪽 끝은 ∞ 인 구간을 만들고 싶다. 원하는 구간은 무엇이 되는가?

문제 7. 11. 위의 문제와 유사한 방법으로, 구간의 왼쪽 끝은 $-\infty$ 인 신뢰구간을 만들어라.

문제 7. 12. 문제 7.10, 문제 7.11과 유사한 방법으로 모분산을 모를 때의 한쪽 끝의 절댓값이 ∞ 인 모평균의 신뢰구간을 만들어라.

문제 7. 13. 문제 7.10, 문제 7.11과 유사한 방법으로, 모비율의 신뢰구간을 만들어라.

문제 7. 14. 문제 7.10, 문제 7.11과 유사한 방법으로, 모분산의 신뢰구간을 만들어라.

문제 7. 15. 모표준편차의 신뢰구간은 어떻게 구할지 이야기하고, 위의 문제와 같이 한 쪽으로는 가능한 데 까지 뻗어 있는 신뢰구간을 만들어라. 기존에 배운 신뢰구간을 양측신뢰구간, 한쪽이 뻗어 있는 신뢰구간을 단측신뢰구간이라 부르기로 하자.

문제 7. 16. 저울이 보여주는 값은 사실 원래 물체의 무게와 비교하여, 오차가 포함된 값이다. 오차는 정규 분포를 따르며, 평균은 0이고 표준편차가 $0.1mg$ 이라고 한다. 동일한 다섯 개의 물체를 저울에 올려본 결과, $3.142mg, 3.163mg, 3.155mg, 3.150mg, 3.141mg$ 이 나왔다. 실제 물체의 무게에 대한 95퍼센트 신뢰구간을 구하라.

문제 7. 17. 랜덤으로 뽑힌 학생 81명에 대해 기초통계학 점수의 표본평균이 74.6 이고 표본표준편차가 11.3 이었을 때, 전체 모평균에 대한 90퍼센트 신뢰구간을 구하여라.

문제 7. 18. 알지 못하는 평균 μ 와 분산 1을 가지는 정규분포를 따르는 모집단에서 표본 X_1, X_2, \dots, X_{n+1} 을 뽑아냈다. 처음 n 개의 표본평균을 \bar{X}_n 이라고 정의하자. 서로 독립인 정규분포는 더해도 독립이며, 평균과 분산은 둘의 합과 같다.

- 1) $X_{n+1} - X_n$ 의 분포는?
- 2) $\bar{X}_n = 4$ 일 때, X_{n+1} 값의 95퍼센트 신뢰구간을 구하여라.

문제 7. 19. U_1, U_2, \dots 는 서로 독립이고 $U(0, 1)$ 을 따르는 확률변수들이라고 하자. 그 다음, N 이라는 확률변수를

$$N = \min(n : U_1 + \dots + U_n > 1)$$

이라고 정의하자. 계산을 통해 $E[N]$ 의 값을 계산하여라.

추가하여, 계산이 맞는지 확인하기 위해 N 에서 표본을 100개 뽑아 95퍼센트 신뢰구간을 구하고자 한다. N 의 모표준편차는 알 수 없지만, N 은 근사적으로 정규분포를 따른다는 사실을 이용할 수는 있다. 어떤 방식을 사용해 신뢰구간을 구해야 하는지 이야기하여라. 결과적으로 표본평균은 \bar{N} , 표본분산은 S^2 으로둔다.

문제 7. 20. 어떤 과학자는 성인의 나트륨 섭취에 관하여 연구를 하고 있다. 남성 9명을 골라 설문한 결과, 그들의 하루 평균 나트륨 섭취량은 1560그램이었으며 표준편차는 33그램이었다. 이를 바탕으로, 성인 남성의 하루 평균 나트륨 섭취량의 95퍼센트 신뢰구간을 완성하여라.

문제 7. 21. 효범이는 동급생 81명에게 물어 자신의 노래점수를 평가해달라고 했다. 그 결과, 표본평균은 95점이었으며 표본분산은 5로 나타났다.

- 1) 모분산의 점추정량은?
- 2) 모분산에 대한 95퍼센트 신뢰구간을 만들어라. 단, $\chi^2_{80,0.025}$ 등을 계산할 필요는 없다.

1 가설검정

1.1 가설검정의 개념

가설 : 모집단의 분포에 대한 어떤 예상

가설 검정 : 모집단에서 추출된 표본을 이용하여 가설의 진위를 판정하는 과정

귀무가설(H_0) : 그동안 믿어왔던, 혹은 사실이라고 생각되는 보수적인 가설

대립가설(H_1) : 표본 자료를 통해 입증하고자 하는 가설

예를 들어, 항암제 투여의 효과가 암 예방에 도움이 된다는 사실을 보이고 싶을 때에는, 귀무가설을 항암제는 암 유병률에 영향을 미치지 않는다는 것으로 하고 대립가설을 항암제는 암 유병률에 영향을 미친다고 두면 된다.

문제 8. 1. 당근에 들어 있는 카로틴 성분이 눈에 좋다는 것을 보이고자 한다. 사람들을 모아 당근 섭취량에 따른 야맹증 발병률을 비교함으로써 이를 보이고자 한다. 귀무가설과 대립가설은 각각 무엇인가?

또한, 귀무가설과 대립가설은 글만이 아니라 기호로서 표현할 수도 있다. 예를 들어, 김종국이 운영하는 헬스장에 온 사람들의 근육량 평균을 μ 라고 하면, 성인 남성의 평균 근육량이 $20kg$ 이라 할 때 귀무가설은

$$H_0 : \mu \leq 20$$

이며, 우리는 김종국 헬스장의 사람들이 많은 근육량을 가진다는 것을 보이고 싶으니 대립가설이

$$H_1 : \mu > 20$$

이 된다.

귀무가설이 참이라고 가정했을 때, 표본이 나올 확률이 매우 적은 경우 우리는 귀무가설이 거짓이라 의심할 수 있다. 예를 들어, 김종국 헬스장의 사람들의 평균 근육량이 $50kg$ 와 같이 $20kg$ 와 비교해 매우 큰 값이라면, 우리는 귀무가설을 의심하고 대립가설이 맞는 것이라 생각하게 될 것이다. 이처럼, 표본으로부터 얻어내는 평가를 위한 값을 **검정통계량**이라 부르며, 검정통계량이 **기각역** A 에 포함되면 귀무가설을 기각하여 대립가설을 선택하고, **채택역**에 포함되면 귀무가설을 기각하지 못한다.

그러나 확률적으로 이를 판단하는 것이기에, 실제와는 다른 가설을 채택할 가능성도 있다. 표본이 우연찮게 엄청난 헬스인들만을 포함하여, 다른 사람들을 고려하지 못했을 가능성이 있는 것이다. 따라서 우리는 두 종류의 오류를 생각한다.

검정결과	실제상황	귀무가설의 상황	대립가설의 상황
	귀무가설을 기각 안 함	옳은 결정	제2종 오류
귀무가설을 기각함	제1종 오류	옳은 결정	

위와 같이 제 1종 오류는 귀무가설이 참임에도 기각하고 대립가설을 선택하여 발생한 오류이며, 제 2종 오류는 귀무가설이 거짓임에도 기각하지 않고 채택하여 발생한 오류라고 이야기할 수 있다. 가설 검정에서 제 1종의 오류를 범할 확률의 최대 한계를 유의수준이라 부르며, α 로 표시한다. 유의수준이 작으면 작을수록, 기각의 가능성성이 줄어드므로 기각역은 줄어들며 유의수준이 클 때는 기각되었던 가설도 작아짐에 따라 기각되지 않을 수 있다. 즉, 더 극단적인 표본이 나와도 귀무가설이 맞을 가능성을 더 높게 평가해준다는 것이다.

가설검정의 순서

- 1) 귀무가설 H_0 과 대립가설 H_1 을 선택한다.
- 2) 검정통계량을 선택한다.
- 3) 유의수준 α 를 정한다.
- 4) 기각역을 구한다.
- 5) 주어진 표본 자료를 이용하여 검정을 하고 결론을 유도한다.

이를 다르게 보면, 우리가 앞서 배운 구간추정과도 연관지을 수 있다. 예를 들어, $\alpha = 0.05$ 일 때 표본평균의 값이 귀무가설에서 정한 모평균의 값보다 멀리 떨어져 있다면 우리는 그 귀무가설을 기각한다. 그런데, 다르게 생각하면 표본평균을 기반으로 95% 신뢰구간을 만들었을 때, 그것이 모평균을 포함하지 않는다면 둘이 그만큼 차이날 확률이 95퍼센트 이하라는 말이므로 기각할 충분한 근거가 된다. 즉, 가설검정은 구간추정을 통해 얻어낸 구간이 원하는 것과 겹치는지를 확인하는 것과 동치임을 알 수 있다.

마지막으로, 가설검정에는 두 가지 종류가 있다.

양측검정 : 귀무가설이 $H_0 : \mu = \mu_0$, 대립가설이 $H_1 : \mu \neq \mu_0$ 와 같은 형태로 표시되며, 집단의 어떤 대푯값이 정해진 값과 다른지를 중심으로 판정한다.

단측검정 : 귀무가설이 $H_0 : \mu \leq \mu_0$ 고 대립가설이 $H_1 : \mu > \mu_0$ 이거나, 귀무가설이 $H_0 : \mu \geq \mu_0$ 이고 대립가설이 $H_1 : \mu < \mu_0$ 와 같이 주어져 원하는 모집단의 성질이 특정 값보다 크거나 작은 것을 보이고자 하는 검정이다.

2 단일 표본의 가설검정

2.1 모평균의 가설검정

정규분포 $N(m, \sigma^2)$ 을 따르는 모집단에서 추출된 크기 n 의 표본의 표본평균은 $N(m, \frac{\sigma^2}{n})$ 을 따른다. 표준편차 σ 의 값을 있다고 가정할 때, 평균이 m_0 가 아님을 보이고 싶다. 그렇다면,

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0$$

으로 식을 세울 수 있으며, 이로부터 기각역은 표본평균이 m_0 으로부터 충분히 먼 곳으로 정해짐을 알 수 있다. 유의수준 α 에서, 정규분포의 대칭성에 의하여

$$P\left(-Z(\frac{\alpha}{2}) \leq Z \leq Z(\frac{\alpha}{2})\right) = 1 - \alpha$$

이고 Z 를 \bar{X} 가 표준화된 것이라 생각하면, 모평균이 m_0 이라는 가정 하에

$$P\left(m_0 - \frac{\sigma}{\sqrt{n}}Z(\frac{\alpha}{2}) \leq m \leq m_0 + \frac{\sigma}{\sqrt{n}}Z(\frac{\alpha}{2})\right) = 1 - \alpha$$

이며, 곧 m 이

$$\left[m_0 - \frac{\sigma}{\sqrt{n}}Z(\frac{\alpha}{2}), m_0 + \frac{\sigma}{\sqrt{n}}Z(\frac{\alpha}{2})\right]$$

여기에 속하지 못한다면 기각할 만큼 충분히 표본평균이 가정된 모평균보다 멀다는 것을 의미한다. 반대로, 여기에 속한다면 모평균이 m_0 이라는 귀무가설을 기각하기는 어렵다.

단측검정 역시 수행할 수 있는데, 그 경우에는

$$H_0 : m \leq m_0$$

$$H_1 : m > m_0$$

와 같이 부등호가 포함되는 형식이며, 우리가 보이고자 하는 것은 m 이 m_0 보다 크다는 사실이다. 그러면, 우리는 m 이 m_0 에 비해 비정상적으로 크게 나오면 귀무가설을 기각하고 대립가설을 채택할 수 있다. 즉, 채택역은 m 이 $-\infty$ 가 되는 점 역시 포함하므로, 왼쪽으로 길게 뻗어 있다. 그러면

$$P(-\infty < Z \leq Z(\alpha)) = 1 - \alpha$$

라는 사실을 이용한다면

$$P(-\infty < m < m_0 + \frac{\sigma}{\sqrt{n}}Z(\alpha)) = 1 - \alpha$$

이다. 따라서 채택역은

$$\left(-\infty, m_0 + \frac{\sigma}{\sqrt{n}}Z(\alpha)\right]$$

이 될 것이다.

문제 8. 2.

$$H_0 : m \geq m_0$$

일 때 H_1 을 쓰고, 채택역과 기각역을 써라.

이를 정리하면,

- 1) $H_0 : m \leq m_0, H_1 : m > m_0$ 일 때 $m > m_0 + \frac{\sigma}{\sqrt{n}}Z(\alpha)$ 이면 H_0 기각
- 2) $H_0 : m \geq m_0, H_1 : m < m_0$ 일 때 $m < m_0 - \frac{\sigma}{\sqrt{n}}Z(\alpha)$ 이면 H_0 기각
- 3) $H_0 : m = m_0, H_1 : m \neq m_0$ 일 때 $m < m_0 - \frac{\sigma}{\sqrt{n}}Z(\frac{\alpha}{2})$ 이거나, $m > m_0 + \frac{\sigma}{\sqrt{n}}Z(\frac{\alpha}{2})$ 면 H_0 기각

문제 8. 3. 작년의 자영업자 월세는 모평균 200만원, 모표준편차 20만원인 정규분포로 나타났다. 올해 자영업자 100명에게 월세를 조사한 결과, 평균이 205만원이었다.

- 1) 올해 평균 월세가 올랐는지 유의수준 0.01에서 검정하라.
- 2) 올해 평균 월세가 변했는지 유의수준 0.01에서 검정하라.

2.2 모분산의 가설 검정

우리는 모집단이 정규분포를 따를 때

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

이 $\chi^2(n-1)$ 을 따른다는 사실을 배웠었다. 따라서,

$$P(\chi_{n-1,1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}^2) = 1 - \alpha$$

$$P(0 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,\alpha}^2) = 1 - \alpha$$

$$P(\chi_{n-1,1-\alpha}^2 \leq \frac{(n-1)S^2}{\sigma^2} < \infty) = 1 - \alpha$$

임을 이용하면 모분산에 대해서도 가설검정의 기각역과 채택역을 만들어낼 수 있다.

- 1) $H_0 : \sigma^2 \leq \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2$ 일 때, $S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha}^2}{n-1}$ 이면 H_0 기각
- 2) $H_0 : \sigma^2 \geq \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2$ 일 때, $S^2 < \frac{\sigma_0^2 \chi_{n-1,1-\alpha}^2}{n-1}$ 이면 H_0 기각
- 3) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$ 일 때, $S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha/2}^2}{n-1}$ 이거나 $S^2 < \frac{\sigma_0^2 \chi_{n-1,1-\alpha/2}^2}{n-1}$ 면 H_0 기각

문제 8. 4. 위의 박스를 증명해 보아라.

3 두 독립표본의 비교

3.1 두 독립표본의 평균비교 - 모집단의 분산이 알려진 경우

여기서는 표본을 두 그룹으로 나눈 후 각각 다른 처리를 하고, 그 산출을 비교하여 처리의 효율이나 특성을 분석하는 방법 등을 배울 것이다. 이때, 각 모집단에서 확률 표본이 독립적으로 관측된 경우 이를 **독립표본**이라 하며, 이처럼 두 독립표본에 각각 다른 처리를 한 경우 **대응표본**이라 부른다. 두 모집단의 평균을 보기 위한 가설은

$$H_0 : m_1 - m_2 = D_0$$

$$H_1 : m_1 - m_2 \neq D_0$$

와 같다. 이때, m_1 은 표본 1의 모평균, m_2 는 표본 2의 모평균, D_0 는 예상되는 모평균의 차이이다.

\bar{X}_1 과 \bar{X}_2 에 대하여, 이들이 각각 $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$ 를 따르는 서로 독립인 정규분포에서 추출된 크기가 n_1, n_2 인 표본의 표본평균이라 하면

$$\bar{X}_1 - \bar{X}_2 \sim N(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

이다.

이때, 여기서는 $m_1 - m_2$ 의 값이 D_0 와 너무 차이나면 기각을 하게 될 것인데, 두 그룹의 표본평균의 차이의 기댓값이 $m_1 - m_2$ 가 됨을 알 수 있다. 따라서, 그 분포를 가지고 검정하는 것은 꽤나 자연스럽게 느껴진다.

문제 8. 5. 위의 박스를 증명하여라. 서로 독립인 정규분포의 합과 차는 역시 정규분포임을 알고 있다고 가정하자.

그러면, 이러한 가정이 모두 합쳐질 경우

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Z$$

임을 알 수 있고, 이로부터 $m_1 - m_2$ 의 $100(1 - \alpha)\%$ 구간추정은

$$\left[(\bar{X}_1 - \bar{X}_2) - Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

이며 우리가 모든 값을 알고 있기에 이를 구해낼 수 있다. 앞서 가설검정은 가정의 모평균이 표본평균으로 써 계산한 구간에 들어오는지를 확인한 것이라는 언급을 해본 적이 있다. 따라서 이를 바탕으로 가설의 종류에 따라 가설검정을 하는 방법을 알아보면,

- 1) $H_0 : m_1 - m_2 \leq D_0, H_1 : m_1 - m_2 > D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이면 H_0 기각
- 2) $H_0 : m_1 - m_2 \geq D_0, H_1 : m_1 - m_2 < D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이면 H_0 기각
- 3) $H_0 : m_1 - m_2 = D_0, H_1 : m_1 - m_2 \neq D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha/2) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이거나, $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha/2) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이면 H_0 기각

임을 알 수 있다.

문제 8. 6. 위를 증명하여라.

3.2 두 독립표본의 평균비교 - 모집단의 분산이 알려져 있지 않은 경우

우리는 앞서 가설검정을 할때 모분산을 알고 있기에 검정통계량의 분모에 올 값을 계산해낼 수 있었다. 그러나 지금처럼 이를 알지 못하는 경우에는, 모분산의 점추정량인 표본분산을 이용하는 편이 좋을 것이다. 통계학자들은 검정통계량의 분모로서 사용할수 있는 값에 대한 **합동추정량**을 생각해냈다. 단, 여기서는 모분산이 서로 같다는 가정을 해야 한다. 모분산이 다르다면, 더 복잡한 공식을 사용하여야 하는데 여기서는 배우지 않는다. 문제를 풀 때는 등분산을 항상 가정하도록 하자.

공통분산 S_p^2 는 아래와 같이 계산된다.

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

또한, 이를 이용하여 계산하게 된다면

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

에서 σ 를 S_p 로 대체하려면, 양변에서

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

를 나눠주면 된다. 따라서 좌변은 표준정규분포를 $\sqrt{\frac{\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}}$ 로 나눠준 것이기에,

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

이다.

이를 바탕으로 또다시 가설검정을 시행한 결과는 아래와 같다.

- 1) $H_0 : m_1 - m_2 \leq D_0, H_1 : m_1 - m_2 > D_0$ 인 경우 $(\bar{X}_1 - \bar{X}_2) > D_0 + t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 면 H_0 기각
- 2) $H_0 : m_1 - m_2 \geq D_0, H_1 : m_1 - m_2 < D_0$ 인 경우 $(\bar{X}_1 - \bar{X}_2) < D_0 - t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 면 H_0 기각
- 3) $H_0 : m_1 - m_2 = D_0, H_1 : m_1 - m_2 \neq D_0$ 인 경우 $(\bar{X}_1 - \bar{X}_2) > D_0 + t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
이거나 $(\bar{X}_1 - \bar{X}_2) < D_0 - t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 면 H_0 기각

문제 8. 7. 위 박스를 증명하여라.

3.3 두 독립표본의 평균비교 - 두 독립표본의 크기가 모두 큰 경우

이 경우에는 중심극한정리에 의해서 정규분포 가정이 필요없고, $n - 1$ 과 n 의 차이가 미미해지기 때문에, S^2 이 σ^2 에 대한 꽤 정확한 추정값이 된다. 따라서

$$\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0, 1)$$

이라는 가정 하에 분석을 수행할 수 있다.

두 모집단의 분산이 알려져 있지 않더라도 표본의 크기가 충분히 큰 경우, 모평균의 차 $m_1 - m_2$ 의 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같다.

$$\left[(\bar{X}_1 - \bar{X}_2) - Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

이를 바탕으로 가설검정을 수행하면

- 1) $H_0 : m_1 - m_2 \leq D_0, H_1 : m_1 - m_2 > D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이면 H_0 기각
- 2) $H_0 : m_1 - m_2 \geq D_0, H_1 : m_1 - m_2 < D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이면 H_0 기각
- 3) $H_0 : m_1 - m_2 = D_0, H_1 : m_1 - m_2 \neq D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha/2) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이거나, $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha/2) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이면 H_0 기각

임을 알 수 있다.

문제 8. 8. 위를 증명하여라.

4 두 모분산의 비교

4.1 F분포와 분산

두 모분산이 같은지 확인하기 전, F분포에 대해 공부하고 가자.

V_1, V_2 를 각각 자유도 k_1, k_2 인 카이제곱분포를 따르는 서로 독립인 확률변수들이라 할 때, $F = \frac{V_1/k_1}{V_2/k_2}$ 의 분포를 문자자유도 k_1 , 분모자유도 k_2 인 **F분포**라 부르며, $F(k_1, k_2)$ 와 같이 나타낸다.

두 모집단이 각각 정규분포를 따를 경우, 크기가 n_1 인 표본과 크기가 n_2 인 표본을 각각에서 뽑으면

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$$

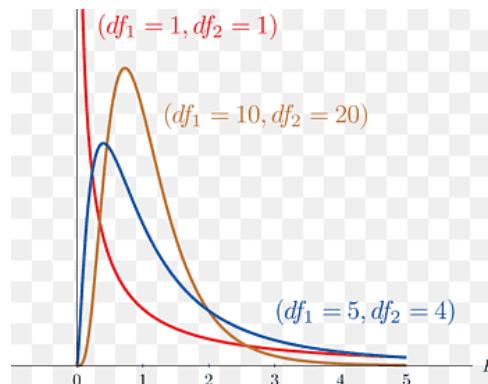
$$\frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

임을 배웠다. 그렇다면, 아래 박스가 성립할 것이다.

두 모집단이 각각 정규분포를 따를 경우, 통계량

$$F = \frac{(S_1^2/\sigma_1^2)}{(S_2^2/\sigma_2^2)}$$

는 $F(n_1 - 1, n_2 - 1)$ 을 따른다.



이때, 자유도가 각각 df_1, df_2 인 F분포의 $100(1 - \alpha)$ 백분위수를 $F_{df_1, df_2, \alpha}$ 라고 정의하며, 이는 기존의 방식과 매우 유사하다.

문제 8. 9. $F_{df_1, df_2, 1-\alpha}$ 를 $F_{df_2, df_1, \alpha}$ 로 표현하라.

그렇다면, 우리는 가설검정을 이용하여 아래 박스를 이야기할 수 있다. 여기서는, 귀무가설이 $\sigma_1^2 = \sigma_2^2$ 인 경우만 다루기로 한다.

$$H_0 : \sigma_1^2 = \sigma_2^2, H_1 : \sigma_1^2 \neq \sigma_2^2 \text{일 때},$$

$$\frac{\text{큰 표본분산}}{\text{작은 표본분산}} > F_{n_1-1, n_2-1, \alpha/2}$$

이면 H_0 을 기각한다.

여기서, 기존의 양측검정과 다르게 부등호가 $>$ 만 있는데, 이는 애초에 검정통계량에서 큰 표본분산을 분모 위로 올렸기 때문에 가능하게 되는 것이다.

문제 8. 10. 위의 박스를 보여라.

5 연습문제

문제 8. 11. T_n 의 자유도가 n 인 t 분포를 따를 때, T_n^2 은 분자의 자유도가 1, 분모의 자유도가 n 인 F 분포임을 보여라.

문제 8. 12. 두 개의 독립표본이 있으며, 그들은 모두 정규분포를 따르는 모집단으로부터 왔다. 그리고 그 모집단은 모두 같은 표준편차 σ 를 가지고 있다고 한다. 자료가

16, 17, 19, 20, 18

과

3, 4, 8

일 때, σ 의 점추정량을 구하라.

문제 8. 13. 만약 두 표본이 서로 독립적이지 않은 상태에서 둘의 평균을 비교하고 싶으면 어떻게 할까? 예를 들어, 학생들의 영어 말하기 점수와 듣기 점수 비교를 통해 말하기 선생님과 듣기 선생님 중 어느 분이 더 잘 가르치시는지 알고 싶다. 따라서 한 반 26명의 학생들에게 수업 이후 시험을 보게 해 성취도를 평가하였다. 근데 문제는 각 학생의 말하기 점수와 듣기 점수는 독립이 아니라는 점이다. 공부를 잘하는 학생이 둘 다 잘 봤을 확률이 높다. 이런 경우에는 대응비교를 통해 수행해낸다.

- 1) 말하기 점수의 모평균을 m_1 , 듣기 점수의 모평균을 m_2 라고 하며, 말하기 점수의 표본평균이 \bar{X}_1 , 듣기 점수의 표본평균이 \bar{X}_2 라고 하자. 앞서 배운 것들을 이용하여, $\bar{X}_1 - \bar{X}_2$ 의 평균을 구하여라.
- 2) 모분산을 모를 때의 단일표본 가설검정을 이용하여, 적절한 가설을 세우고 신뢰수준이 α 일 때의 채택역을 구하여라.

문제 8. 14. 생산 라인의 두 기계로부터 물품이 나오는데, 각 기계에서 생산되는 물품의 무게는 독립표본이라고 하자. 첫째 기계에서는 36개의 물품에 대해 표본평균이 120그램, 표본분산이 4였다. 두번째 기계에서는 64개의 물품에 대해 표본평균이 130그램, 표본분산이 5였다. 첫째 기계에서 생산되는 물품의 무게는 $N(\mu_1, \sigma^2)$ 을 따르며 둘째 기계에서 생산되는 물품의 무게는 $N(\mu_2, \sigma^2)$ 을 따른다고 가정하자. $\mu_1 - \mu_2$ 의 95% 신뢰구간을 구하여라.

문제 8. 15. 문제 8.14를 모분산이 각각 4, 5인 것으로 바꾸어서 풀어보아라.

문제 8. 16. 시장에서 유통되는 모든 담배는 현재 평균 니코틴 함유량이 최소 $1.6mg$ 이다. 그런데 어떤 담배회사가 담뱃잎을 새롭게 가공하는 법을 개발해내 그 방법으로 담배를 생산할 시 담배의 평균 니코틴 함유량이 $1.6mg$ 보다 작아질 수 있다고 주장하였다. 담배공사는 20개의 해당 담배를 얻어내 분석하였다. 일반적으로, 담배의 니코틴 함유량 모분산은 0.8임이 알려져 있다. 표본평균이 $1.54mg$ 이었을 때, 유의수준 5%에서 이 회사의 주장이 맞다고 이야기할 수 있는가?

문제 8. 17. 병원 환자들의 혈중 콜레스테롤 농도를 줄일 수 있는 약을 새롭게 개발하였다. 50명의 환자들에 대해 이 약을 1달 동안 투여한 이후, 혈중 콜레스테롤 농도의 감소를 확인하였다. 표본평균은 $14.8mg/ml$, 모표준편차는 $6.4mg/ml$ 로 나타났다. 신약의 효과가 있다고 이야기할 수 있는가? 유의수준 0.05에서 검정하라.

문제 8. 18. 빼빼로 기계는 빼빼로에 묻혀질 초콜릿의 길이를 조절한다. 만약 초콜릿 길이의 표준편차 σ 가 0.15cm 보다 작다면, 이 기계는 효과적이라고 판정할 수 있다. 20개의 빼빼로를 뽑아 표본분산을 확인한 결과, 0.025제곱센치미터 였다. 우리는 이 기계가 비효과적이라고 이야기할 수 있는가? 유의수준 0.05 에서 검정하라.

문제 8. 19. 어떤 효소의 활성을 막는 비가역성 억제제에는 *suicide inhibitor*와 *transition state analog*가 있다. 어떤 억제제를 사용해도 효소의 활성에서 분산 차이가 없다는 것을 밝히기 위해, 첫째 억제제에 대해서는 10번, 둘째 억제제에 대해서는 12번의 실험을 수행하였다. 그 결과, 표본분산은 각각 0.14 와 0.28 로 나타났다. 유의수준 0.05 에서 우리는 둘이 다른 분산을 가진다고 이야기할 수 있는가?

문제 8. 20. 실제 모평균은 10인데, 우리는 귀무가설에서 모평균이 15라고 가정하고 크기가 100인 표본을 뽑아 유의수준 0.05에서 양측검정을 진행하고 있다. 모집단이 정규분포를 따르며 모분산이 20이라는 사실은 우리가 이미 알고 있었을 때, 제 2종의 오류를 저지를 확률은?

문제 8. 21. 어떤 제약사가 실험을 진행하고 있다. 이 제약사는 자신들의 약을 투여했을 때 근육량이 늘어난다고 주장한다. 일반적으로 성인 남성의 근육량은 20kg, 표준편차는 5kg이라 하자. 해당 제약사는 자신들의 제품을 복용하면 근육량이 5kg 정도 증가할 것이라고 예상하고 있다. 만약 그들의 예상이 맞을 때, 유의수준이 0.05인 가설검정 결과 제 2종의 오류를 저지를 확률이 20퍼센트 이하가 되려면 표본의 크기는 최소 얼마여야 하는가?

문제 8. 22. 특정 트랜지스터가 버틸 수 있는 전류의 세기가 최소 $210A$ 는 된다고 믿어진다. 표본을 구해 확인한 결과, 그들의 평균 한계전류는 $200A$ 였으며, 표본표준편차가 35 였다. 유의수준 0.05 에서 귀무가설을 기각할 수 있는지

- 1) 표본의 크기가 25 일 때
- 2) 표본의 크기가 64 일 때 판정하라.

문제 8. 23. 어떤 교수는 장어가 한우에 비해 더욱 비싸다고 주장하고 있다. 16개의 장어집에 대해 평균 가격은 72700 원이었으며, 표본표준편차는 2400 원이었다. 16개 고깃집에서 한우는 표본평균이 71400 원, 표본표준편차가 2200 원이었다. 교수의 주장이 옳은가? 유의수준 0.05 에서 검정하라.

문제 8. 24. 모비율에 대한 신뢰구간을 구하였듯이, 가설 $H_0 : p = p_0, H_1 : p \neq p_0$ 에 대하여 표본비율 \hat{p} 일 때의 가설검정 방법을 개발하라.

문제 8. 25. 모비율에 대한 단측검정을 개발하라.

추가문제1. 1. '애쉬'가 본인의 기지로부터 상대 기지까지 매를 보낼 때까지 걸리는 시간이 평균 30초, 표준편차 5초인 정규분포를 근사적으로 따른다고 하자.

1) 애쉬의 매가 전투 시작 8분 25초 경에 출발하였을 때, 전투 시작 9분 전까지 도착할 수 있을 확률은 얼마인가?

2) 애쉬의 매가 전투 시작 9분 전까지 상대 기지에 도착할 확률을 97.5% 이상으로 만들기 위해서는 적어도 전투 시작 몇 분 후에 매를 날려야 하는가?

추가문제1. 2. 서울과학고에서, 학점에 따라 수업에 대한 기초통계학 수업 만족도를 조사하였다. 조사 결과, 4점대 이하의 학생 25명에 대해서는 만족도 평균이 31, 표본분산이 27이었으며, 4점대 이상의 학생 30명에 대해서는 만족도 평균이 29, 표본분산이 29였다.

1) 4점대 이하 그룹과 4점대 이상 그룹의 분산은 동일하다고 볼 수 있는가? 적절한 가설을 쓰고 유의수준 5%에서 이를 검정하시오.

2) 4점대 이하 그룹과 4점대 이상 그룹의 수업 만족도에는 차이가 존재한다고 볼 수 있는가? 둘의 분산이 같다고 가정하고, 유의수준 5%에서 이를 검정하시오.

추가문제1. 3. 농부 '나서스' 씨는 총 n 개의 당근을 농사지어 납품하면 하루 일과를 마무리할 수 있다. 당근이 수확되는 시간은 서로에게 독립적인 낮 12시와 1시 사이의 균등분포를 가진다. 나서스 씨가 첫 당근을 수확한 이후 마지막 당근을 수확하기까지 걸리는 시간을 확률분포 X 라고 하자. X 의 기댓값을 구하여라.

추가문제1. 4. 임업자 '티모'씨는 10000개 가량의 버섯을 심는데, 한 버섯에서 발견되는 독포자의 개수가 평균 0.5인 푸아송 분포를 따른다는 것을 알아냈다. 확률변수 X 를 독포자가 없는 버섯의 개수라 할 때, X 가 $6000 - m$ 과 $6000 + m$ 사이에 있을 확률이 근사적으로 $2/3$ 이 되는 m 의 값을 구하여라. 단, 계산의 편의를 위하여 $e^{-0.5} = 0.6$ 이라고 하자.

추가문제1. 5. 1부터 6까지의 눈을 가진 주사위가 있다. 주사위를 던졌을 때 나오는 첫 번째 눈의 수를 X_1 , 주사위의 밑면에 있는 네 모서리 중 하나를 무작위로 골라 그 모서리를 기준으로 1번 회전시킬 때 위에 온 눈의 수를 X_2 라고 하자. 예를 들면, $X_1 = 1$ 일 경우 X_2 는 2, 3, 4, 5가 될 수 있다. 즉, 이 주사위는 마주보는 면의 눈의 합이 7인 주사위이다. 다음 물음에 답하시오.

- 1) 확률변수 X_1 과 X_2 의 결합확률질량함수를 구하고, 이를 이용하여 X_1 과 X_2 가 종속임을 설명하시오.
- 2) $Y = X_1 - X_2$ 일 때, $E[Y]$ 의 값을 구하시오.

추가문제1. 6. 자연수 $1, 2, 3, \dots, n$ 의 값을 가지는 이산확률변수 X 에서, 확률질량함수 F 에 대하여

$$\frac{F(k) - F(m)}{k - m} = a(k + m + 1)$$

이 성립한다고 한다. 단, 위의 식에서 m 과 k 는 모두 양의 정수이며, k 는 m 보다 크다. X 의 기댓값이 9일 때, a 와 n 의 값을 구하시오.

추가문제1. 7. '젠향타' 씨는 자신의 구슬 주머니에서 구슬을 꺼내려 한다. 구슬 주머니에는 보라색 구슬이 M 개, 노란색 구슬이 N 개, 그리고 푸른색 구슬이 $7 - M - N$ 개 포함되어 있다. 단, 구슬의 개수는 음이 아닌 정수개다.

1) $M + N = 4$ 일 때 복원추출을 3회 반복하여 나온 보라색 구슬과 노란색 구슬 개수의 합을 확률변수 X 라 하자. X 가 어떤 확률분포를 따르는지 이야기하고, $E[X^2]$ 의 값을 구하라.

2) $M = 2, N = 3$ 일 때 비복원추출을 2회 하여 나온 보라색 구슬의 수를 확률변수 Y , 노란색 구슬의 수를 확률변수 Z 에 대응시키자. $Cov(Y, Z)$ 의 값을 구하여라.

3) 비복원추출을 7회 하려고 한다. 이때, 짹수 번째 추출에서는 보라색 구슬이 뽑히면, 홀수 번째 추출에서는 노란색 구슬이 뽑하면 옆에 있는 판에 우물 정자로 표시한다. 시행 이후 기록된 횟수가 확률변수 W 라고 할 때, $E[W]$ 의 값을 구하여라.

추가문제1. 8. '엘리스' 씨가 기르는 거미가 낳는 알은 평균이 100인 푸아송 분포를 따르며, 알의 부화 확률은 0.3이다. 또한, 각 알의 부화는 서로의 부화에 독립적이다.

- 1) 엘리스 씨의 거미가 낳는 알의 개수가 확률변수 X 일 때, X 의 분산을 구하여라.
- 2) 엘리스 씨가 100개의 알을 낳았을 때, 부화한 알이 25개에서 30개 사이일 확률을 근사하라.
- 3) 부화한 새끼 거미의 수가 푸아송 분포를 따름을 보이고, 그 평균을 구하라.

추가문제1. 9. 자연수 n 에 대하여, 실수에서 정의된 확률변수 X_n 의 누적분포함수가 아래와 같다.

$$F_n(x) = b_n \frac{e^{a_n x}}{e^{a_n x} + e^{-a_n x}}$$

- 1) a_n 이 양수여야 함을 보이고, b_n 의 값을 구하여라.
- 2) $f_n(0)$ 의 값이 n 일 때, a_n 을 n 으로 표현하라.
- 3) a_n 이 $1/n$ 이라고 하자. 이때, $(0, 0.5)$ 에서의 $F_n(x)$ 의 접선이 지나는 점 중 y 좌표가 1보다 작으면서 x 좌표가 양의 정수인 점이 20개 이상이 되는 n 의 최솟값을 구하여라.

추가문제1. 10. 대장장이 '오른' 씨가 5분 동안 만드는 무기의 수는 2,3,5개 중 하나이다. 만드는 무기의 수를 확률변수 X 라고 할 때, 다음 물음에 답하여라.

- 1) $E[X] = 3, V(X) = 2$ 일 때, X 의 확률질량함수를 구하여라.
- 2) 오른 씨는 자신이 평균적으로 3개의 무기를 만들어 선물한다고 주장하고 있다. 이를 확인하기 위해, 오른 씨가 5분 동안 만드는 무기의 수를 100번 관찰하였다. 그 결과 표본평균이 2.8개로 나타났다. 표본분산은 앞선 문제와 같이 2였다. 귀무가설과 대립가설이 무엇인지 쓰고, 유의수준 5%에서 오른 씨의 주장이 맞는지 확인하여라.
- 3) 오른 씨가 만든 무기의 가격은 항상 만 원 단위이다. 오른 씨가 만들어 본인이 쓰는 무기 5개와 다른 사람에게 선물한 무기 5개의 평균 가격을 비교해 보고자 했다. 오른 씨가 사용하는 무기 5개의 가격은 각각 7, 9, 5, 10, 8만원이었고, 선물한 무기의 가격은 4, 5, 9, 6, 4만원이었다. 둘의 분산이 같다고 할 때, 가격의 차이가 있는지 유의수준 5%에서 검정하여라.
- 4) 만약 10개의 무기를 무작위로 5개씩 나누어 사용하는 무기와 선물한 무기로 분류했다면, 3)에서 구한 평균 가격의 차이보다 크거나 같은 평균 가격이 나올 확률은 얼마인가?

추가문제2. 1. 케인 씨는 농사를 위해 낫을 고르고 있다. 다르킨 사에서 나온 낫 패키지에 들어 있는 낫의 개수는 1, 2, 4개 중 하나이며, 그암 사에서 나온 낫 패키지에 들어 있는 낫의 개수는 1, 3개 중 하나이다. 케인 씨는 다르킨 사와 그암 사에서 낫 패키지를 하나씩 구매했다. 다르킨 사 패키지에 든 낫의 개수를 이산확률변수 X , 그암 사 패키지에 든 낫의 개수를 이산확률변수 Y 에 대응시키자. X 와 Y 의 결합확률질량함수는

$$P(X = x, Y = y) = \frac{xy + x + y + 2}{a} \quad (x = 1, 2, 4, y = 1, 3)$$

이다.

- 1) 양의 실수 a 의 값을 구하시오.
- 2) X 와 Y 의 주변확률질량함수를 구하시오.
- 3) $E(XY)$ 의 값을 구하시오.
- 4) X 와 Y 는 독립된 이산확률변수인가?

추가문제2. 2. 훌륭한 연주자 소나 씨는 3종류의 악곡을 연주한다. 악곡1을 연주할 확률은 p_1 , 악곡2를 연주할 확률은 p_2 , 악곡3을 연주할 확률은 p_3 이다. 소나 씨는 연주마다 독립적인 곡 선택을 한다. 소나 씨가 총 n 번의 연주를 할 때, 악곡1의 연주 횟수를 N_1 , 악곡2의 연주 횟수를 N_2 라는 이산확률변수에 대응시키자.

- 1) N_1 과 N_2 는 각각 어떤 확률분포를 따르는가?
- 2) $Cov(N_1, N_2)$ 를 구하여라.
- 3) N_1 과 N_2 는 서로 독립인 확률변수인가?
- 4) $Corr(N_1, N_2)$ 를 구하고, 이를 통해 둘의 상관관계를 설명하여라.

추가문제2. 3. 마오카이 씨는 묘목을 심을 장소를 모색하고 있다. 마오카이 씨는 묘목 5개를 가지고 있으며, 심을 장소는 우물 근처 혹은 수풀 근처이다. 마오카이 씨의 집 주변에는 총 160개의 우물이 있고, 수풀은 140개 있다. 마오카이 씨가 심은 묘목 중 우물 근처에 심은 묘목의 개수를 확률변수 X 라고 하자. 묘목을 심은 우물이나 수풀 근처에는 다시 묘목을 심지 못한다.

1) X 가 무슨 분포를 따르는지 말하고, 기호로 나타내라. 예를 들어 X 가 $n = 50$ 이고 $p = 0.7$ 인 이항분포를 따른다면, $X \sim B(50, 0.7)$ 로 써라.

2) 우물 근처에 심은 묘목 개수의 기댓값과 분산을 구하여라.

3) 이항분포로의 근사를 통해, 우물 근처에 1개의 묘목만이 심어질 확률을 근사하여라.

추가문제2. 4. 횟집을 운영하는 일라오이 씨가 차리는 문어의 무게는 $1 \sim a$ 킬로그램이며, 이를 연속확률변수 X 에 대응시킬 때 X 의 확률밀도함수 $f(x)$ 는

$$f(x) = \frac{1}{6a+2}x^3 \quad 1 \leq x \leq a$$

라고 한다. 단, a 는 양의 실수이다.

- 1) a 의 값을 구하여라.
- 2) 일라오이 씨가 차리는 문어 무게의 기댓값을 구하여라.
- 3) 일라오이 씨가 차리는 문어의 무게가 x 일 때, 그 가격은 $10000x^2$ 원이다. 일라오이 씨네 가게에서 문어를 먹었을 때, 내야 하는 가격의 기댓값을 구하여라.
- 4) 문어 가격을 확률변수 Y 에 대응시키자. 즉, $Y = 10000X^2$ 이다. Y 의 확률밀도함수를 구하고, 40000 원 이하의 가격을 지불하고 문어를 먹을 확률을 구하라.

추가문제2. 5. 재봉사 그웬 씨가 인형 하나를 만드는 데 걸리는 시간은 지수분포를 따른다. 또한, 그웬 씨가 1시간 안에 만들 확률은 $1 - e^{-\frac{3}{5}}$ 이라고 한다.

- 1) 인형 하나를 만드는 데 2시간 이상이 걸릴 확률은?
- 2) 그웬 씨가 60시간 동안 만드는 인형의 개수를 확률변수 X 라 하자. $E(X)$ 의 값은?
- 3) X 의 확률질량함수를 쓰고, 이것이 어떤 분포를 따르는지 이야기하라.

추가문제2. 6. 확률변수 A 는 정규분포를 따르며, $A \sim N(2, 4)$ 라고 표현할 수 있다. 이때, 이차방정식 $-2x^2 + Ax + 6$ 의 근 중 더 큰 것을 확률변수 B 라고 하자. $P(\sqrt{3} \leq B \leq 3)$ 를 구하여라.

추가문제2. 7. 음이 아닌 정수 n, m 에 대하여 $[0, a]$ 에서 정의된 연속확률변수 $X_{n,m}$ 의 확률밀도함수 $f_{n,m}(x)$ 가

$$f(x) = \begin{cases} nx^{17} + mx^{11} & 0 \leq x \leq a \\ 0 & otherwise \end{cases}$$

라고 한다.

- 1) $a = 1$ 일 때, 가능한 (n, m) 의 쌍을 모두 구하시오.
- 2) $n = 18, m = 24$ 일 때, a^6 의 값을 구하여라.
- 3) $n = 18, m = 24$ 일 때, $\frac{E(X)}{a}$ 의 값을 구하여라.
- 4) $X_{18,24}$ 의 값이 $\frac{20}{21}a$ 보다 클 확률이 0.98보다 작음을 보여라. 단,

$$\frac{18}{19}(\sqrt{5} - 2) + \frac{12}{13}(3 - \sqrt{5}) = 0.93$$

이라고 둔다.

추가문제2. 8. 헤카림 씨는 경마에서 말을 선택하기 위해 간단한 놀이를 진행하고 있다. 경마에는 1번 말부터 9번 말까지가 참여하며, 헤카림 씨는 아래의 과정을 거쳐서 말을 선택한다.

먼저, 헤카림 씨는 익명의 점쟁이으로부터 말의 상태에 대한 정보를 얻는다. 점쟁이는 3번, 4번, 7번 말을 제외한 6마리의 말 중 하나를 임의로 선택해 알려준다. 즉, 6마리의 말이 선택될 확률은 모두 같다.

그 다음, 점쟁이가 못 미더운 헤카림 씨는 점쟁이가 알려준 번호를 디지털 숫자로 변환한 다음 그것을 뒤집어 새로운 숫자를 얻는다. 예를 들어, 1은 뒤집어도 1이며, 6은 뒤집으면 9가 된다.

마지막으로 얻은 숫자에 대해, 헤카림 씨는 정보 상인에게 물어 자신이 앞서 구한 번호의 말을 택하면 돈을 팔 가능성이 있을지 물어본다. 그러나 이 정보 상인은 가짜 상인으로, 동전 던지기를 통해 앞면이 나오면 가능성이 높다고 말하고, 뒷면이 나오면 가능성이 적다고 말한다. 만약 헤카림씨가 가능성이 높다는 대답을 들으면, 헤카림씨는 앞선 과정에서 얻은 숫자에 배팅한다. 반면, 가능성이 적다는 대답을 들으면 3, 4, 7 중 하나를 동일한 확률로 임의로 선택한다.

결론적으로 헤카림 씨가 선택한 말의 번호를 X 라 하자.

1) X 의 확률질량함수를 구하여라.

2) 만약 정보 상인이 진짜 상인이어서, 모든 말의 우승 확률을 알고 있다고 하자. 우승 확률은 말의 번호에 비례한다. 즉, 9번 말이 우승할 확률은 1번 말이 우승할 확률의 9배이다. 우승하는 말의 번호를 Y 라 할 때, $E(Y)$ 의 값을 구하여라.

3) 진짜 정보 상인을 헤카림씨가 방문했다고 하자. 정보 상인은 헤카림씨가 제시한 말의 번호가 5보다 크면 가능성이 높다고 이야기해주고, 5보다 작으면 가능성이 적다고 이야기해준다. 이 경우에 헤카림씨가 선택한 말의 번호 Z 의 확률질량함수를 구하여라.

4) 점쟁이가 점지한 말의 번호를 W 라고 하자. 말의 번호를 각각 W, X, Z 로 결정했을 때, 돈을 팔 확률이 가장 높은 방식은? 단, 배팅한 말이 1등하였을 때에만 돈을 한다.

추가문제2. 9. 양조업자 그라가스 씨가 한 시간 동안 만드는 술통의 개수는 1, 2, 4개 중 하나이다. 그 개수를 확률변수 X 에 대응시킬 때, $P(X = 1) = 0.3, P(X = 2) = 0.4, P(X = 4) = 0.3$ 이다.

1) 관찰을 통해 그라가스 씨가 한 시간 동안 평균적으로 만드는 술통의 개수를 확인하고자 했다. 100시간 동안 관찰했을 때, 만드는 술통의 표본평균을 \bar{X} 이라 하자. 중심극한정리를 통해 \bar{X} 의 분포를 근사하라.

2) 2시간 관찰할 때의 표본분산 S^2 의 확률질량함수를 구하고, $E(S^2)$ 의 값을 구하라.

추가문제2. 10. 판테온 씨가 하루 동안 굽는 빵의 개수는 표준편차가 20개인 정규분포를 따른다고 한다. n 일 동안 판테온씨의 빵집을 보면서, 하루에 굽는 빵의 개수를 확인하였다. 이를 이용하여 신뢰도 95%로 추정한 빵의 개수 평균의 신뢰구간이 [100, 110]일 때, n 의 값을 구하여라. 단, $P(Z < 2) = 0.975$ 로 계산하자.

추가문제2. 11. 시비르 씨가 하루 동안 판매하는 피자의 개수는 정규분포를 따른다고 한다. 9일 동안 시비르 씨의 피자가게를 보면서, 하루에 파는 피자의 개수를 확인하였다. 판매한 피자 수의 표본평균은 100판, 표본분산은 27이었다고 하자. 모집단의 평균을 95%의 신뢰도로 추정하라.

추가문제2. 12. 악세서리 판매업자 모데카이저씨가 자신들의 고객들 100명에게 조사해본 결과, 수은 머리띠를 구매하는 고객은 36명이었다. 실제로 수은 머리띠를 구매한 사람들의 모비율을 p 라고 할 때, p 의 95% 신뢰구간을 구하여라.

추가문제2. 13. 양봉업자 신지드 씨는 400개의 꿀통에서 꿀을 얻어낸다. 한 꿀통에서 얻은 꿀에 포함된 당분은 일반적으로 평균이 $30g$ 이고, 표준편차가 $4g$ 인 정규분포를 따른다고 알려져 있다. 당분이 적절한 당분양인 $30g$ 보다 너무 많거나 적은 꿀은 상품성이 없다고 생각되어 폐기된다.

1) 먼저 4개의 꿀통에서만 꿀을 얻어냈고, 그 결과 표본평균이 $33g$ 으로 나타났다. 신지드 씨의 꿀통에 들어있는 꿀들의 평균 당분 양이 $30g$ 인지 아닌지 가설 검정을 통해 판단하시오. 단, 유의수준은 0.05 로 한다.

2) 심층 조사 결과 알려진 것과 같이 당분의 양 X 는 $N(30, 16)$ 을 따른다고 한다. 신지드 씨는 400개의 꿀통을 4개씩 나누어 100개의 그룹으로 만들었다. 그 다음, 4개의 꿀통을 혼합하여 혼합벌꿀을 얻었다. 만약 혼합벌꿀에 포함된 벌꿀의 양이 $125g$ 이상이거나, $115g$ 이하라면 그 혼합벌꿀은 폐기한다. 폐기된 혼합벌꿀이 5개 이상일 확률의 근삿값을 구하여라.

추가문제2. 14. 알리스타라는 이름을 가진 젖소가 있다. 이 젖소로부터 얻은 우유에 포함된 리터당 유지방의 양은 평균이 $15g$ 인 정규분포를 따른다고 한다. 또한, 저지방 우유를 만들기 위해서는 유지방 양의 분산이 $0.005g$ 이하여야 한다. 최근에 생산된 우유 37리터를 분석해본 결과, 표본분산은 S^2 이었다. 이때, 유의수준 0.05에서 저지방 우유를 만들 수 있다는 결론이 만들기 위해서는 S^2 의 값이 c 보다는 작아야 한다. c 의 값을 구하라.

추가문제2. 15. 양계업자 클레드 씨는 큰 닭과 작은 닭을 나누어 기른다. 큰 닭 10마리가 낳는 알의 개수는 표본평균 5개, 모분산 4였으며 작은 닭 8마리가 낳는 알의 개수는 표본평균 4개, 모분산 1이었다. 낳는 알의 개수는 정규분포를 따른다고 가정하자.

- 1) 닭의 크기에 따라 낳는 알의 개수가 다른지를 유의수준 0.05에서 가설검정하여라.
- 2) 모분산이 아니라 표본분산일 때로 바꾸어 1)을 해결하라.

추가문제2. 16. 초원에 사는 니달리 씨는 암사자 100마리와 수사자 150마리를 관리하고 있다. 일주일에 암사자가 먹이를 구하는 데 사용하는 시간은 표본평균 10시간, 표본분산 4임을 확인하였고 수사자의 경우에는 표본평균 6시간, 표본분산 9였다. 성별에 따른 평균 사냥 시간의 차에 대한 95% 신뢰구간을 구하여라.

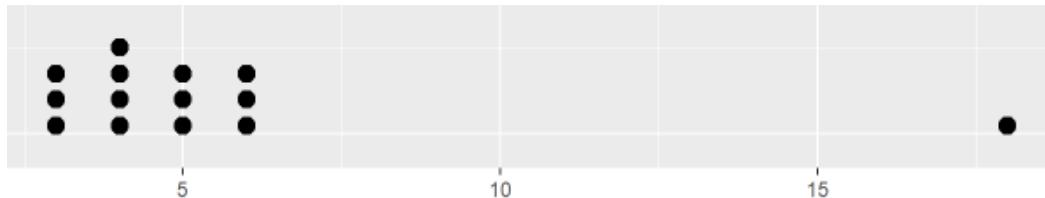
추가문제2. 17. 자이라 씨가 기르는 장미의 가시 수는 정규분포를 따르며, 4송이 장미에 대한 표준편차는 5개였다. 선인장의 가시 수 역시 정규분포를 따랐으나, 선인장 5개의 표준편차가 6개였다. 두 식물의 가시 개수 분산이 같은지를 유의수준 0.01에서 검정하여라.

문제 1. 1. 서울과학고등학교 학생 중 아이돌을 좋아하는 학생이 300명 있다고 하자. 그 중 40명을 뽑아 무슨 아이돌을 좋아하는지 확인하였다. 모집단의 크기와 표본의 크기는 각각 얼마인가?

모집단의 크기는 300, 표본의 크기는 40이다.

문제 1. 2. 중학교 때 배운 '도수분포표'는 기술통계학과 추측통계학 중 어느 분야에 속하겠는가?

기술통계학에 속한다.



문제 1. 3. 위의 점도표는 서울과학고등학교 학생들 중 몇 명을 뽑아 그들의 하루 핸드폰 사용 시간을 기록한 것이다. 표본의 크기는?

표본의 크기는 14이다.

1		34
2		399
3		27
4		12889
5		1
6		
7		
8		
9		3

문제 1. 4. 위 줄기와 앞 그림은 걸그룹 우주소녀 팬클럽 우정에게 물어 조사한 우주소녀 앨범 구매 개수이다. 이 자료에서 이상점은?

이상점은 93이다.

계급	도수
0곡~10곡	2
10곡~20곡	2
20곡~30곡	16
30곡~40곡	10
40곡~50곡	6
계	36

문제 1. 5. 위의 도수분포표는 서울과학고등학교 학생들을 대상으로 90년대 노래를 몇 곡 아는지를 조사하여 나타낸 도수분포표이다. 90년대 노래를 40곡 이상 50곡 미만 아는 학생들의 상대도수는 얼마인가?

$6/36 \approx 0.1667$ 그 상대도수이다.

문제 1. 6. 히스토그램에서 그려진 막대들의 넓이 합은 (계급의 나비) \times (도수의 총합)임을 보여라. 단, 계급의 나비는 모든 계급에 대해 같다고 하자.

막대들의 넓이 합 $\sum_{i=1}^n S_i$ 를 생각하자. 이때 계급의 개수를 n , 각 막대의 넓이를 S_i 로 둔 것이다. $S_i = (\text{계급의 나비}) \times (f_i)$ 이며 이때 f_i 는 i 번째 계급의 도수를 의미한다. 즉

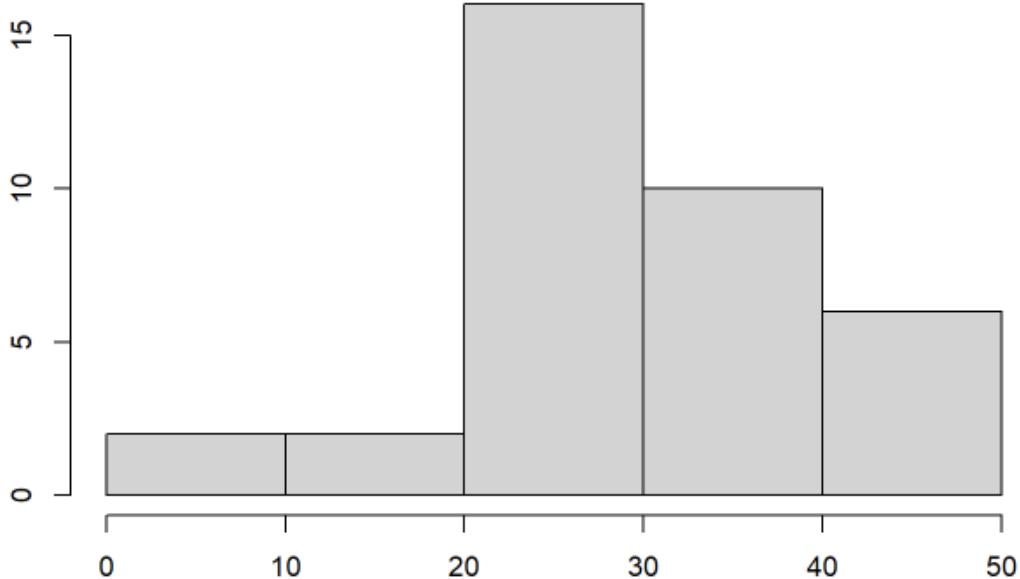
$$S = \sum_{i=1}^n S_i = \sum_{i=1}^n (\text{계급의 나비}) \times (f_i) = (\text{계급의 나비}) \sum_{i=1}^n (f_i) = (\text{계급의 나비}) \times (\text{도수의 총합})$$

임을 알 수 있다.

문제 1. 7. $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$ 임을 보여라.

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \end{aligned}$$

임을 알 수 있다.



문제 1. 8. 3쪽의 히스토그램에서 제1사분위수와 제3사분위수는 각각 얼마인가?

제 1사분위수는 25곡, 제3 사분위수는 35곡이다. 계급으로 주어져 있으므로 계급값을 대신 사용하기로 한다.

문제 1. 9. 함수 $f : \mathbb{R} \rightarrow \mathbb{R}$ 이 있다. 주어진 자료 x_1, x_2, \dots, x_n 에 대하여, 처리한 자료 $f(x_1), f(x_2), \dots, f(x_n)$ 을 생각하자. $f(x) := ax + b$ 일 때의 $\overline{f(x)}$ 를 a, b, \bar{x} 를 이용하여 표시하라.

$$\overline{f(x)} = \sum_{i=1}^n (ax_i + b) = a \sum_{i=1}^n x_i + b = a\bar{x} + b$$

문제 1. 10. 위의 문제에서 f 가 단조함수라고 생각하자. 훌수 개의 자료 $f(x_1), f(x_2), \dots, f(x_n)$ 의 중앙값을 함수 f 와 \hat{x} 를 이용하여 표시하라.

단조함수라면 f 를 취하여도 크기 순서는 유지된다. 따라서 이들의 중앙값은 $f(\bar{x})$ 라고 이야기할 수 있다.

문제 1. 11. 위의 문제에서 $f(x_1), f(x_2), \dots, f(x_n)$ 의 최빈값을 함수 f 와 원래 자료의 유일한 최빈값 x_0 를 이용하여 표시하라.

f 가 단조함수이므로 크기가 유지되고, 이 상황에서 최빈값은 f 를 씌우기 전 최빈값에 f 를 씌운 것이나 매한가지다. 즉 $f(x_0)$ 이다.

문제 1. 12. 함수 $f(k) = \frac{(x_1 - k)^2 + (x_2 - k)^2 + \dots + (x_n - k)^2}{n}$ 의 최소가 되는 k 의 값을 구하여라.

위의 함수는 k 에 대해 정리하면

$$f(k) = k^2 - 2\bar{x}k + \frac{\sum_{i=1}^n x_i^2}{n}$$

으로 k 에 대한 이차함수이기에, 이것이 최소가 되는 k 의 점은 축의 x 좌표인

$$\bar{x}$$

이다.

문제 1. 13. 함수 $g(k) = \frac{|x_1 - k| + |x_2 - k| + \dots + |x_n - k|}{n}$ 의 최소가 되는 k 의 값을 구하여라.

이렇게 절댓값들의 합으로 주어지는 함수는 그래프를 그려볼 시 x_1, x_2, \dots, x_n 들의 중앙값일 때 k 가 최소가 됨을 확인할 수 있다. 즉, k 는 x_1, x_2, \dots, x_n 의 중앙값.

문제 1. 14. 문제 1.9.와 문제 1.7.을 이용하여 가평균 x_0 의 주어졌을 때 분산을 구해 보아라.

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\ &= \sum_{i=1}^n ((x_i - x_0)^2 + 2x_0x_i - x_0^2) - n\bar{x}^2 \\ &= \sum_{i=1}^n (x_i - x_0)^2 + 2nx_0\bar{x} - nx_0^2 - n\bar{x}^2 \\ &= \sum_{i=1}^n (x_i - x_0)^2 - n(x_0 - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - x_0)^2 - n(\bar{u})^2 \end{aligned}$$

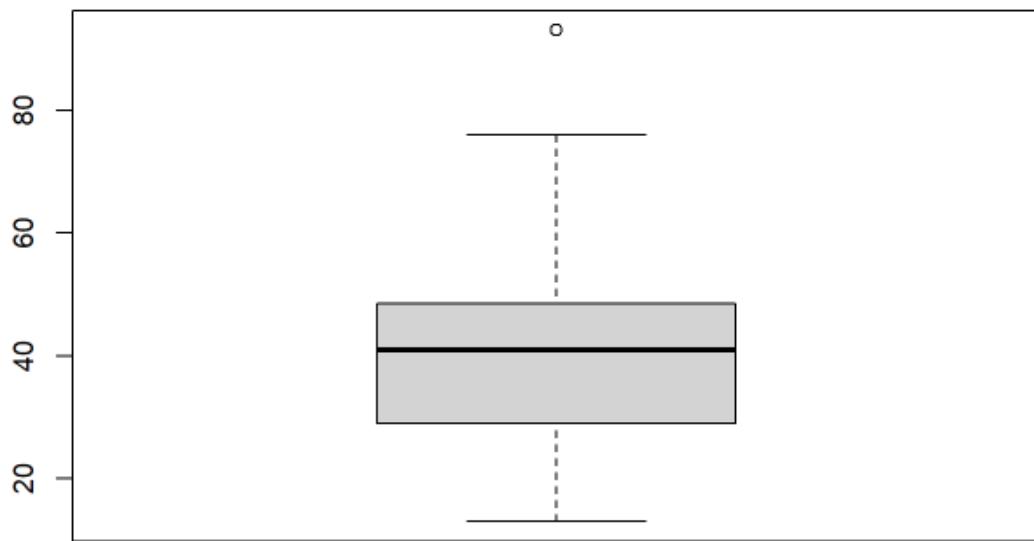
이므로 분산은

$$\frac{\sum_{i=1}^n (u_i)^2}{n} - (\bar{u})^2$$

로, 제평평제 공식을 그대로 적용하면 됨을 알 수 있다.

문제 1. 15. 위의 상자 그림에서 UIF를 넘어서는 점은 몇 개인가?

1개 있다.



1 연습문제

문제 1. 16. 자료 x_1, x_2, \dots, x_n 의 평균이 m , 표준편차가 σ 이다. 자료 $3x_1^2 - 2, 3x_2^2 - 2, \dots, 3x_n^2 - 2$ 의 평균을 구하여라.

이들의 평균 A 를 구하여 보자.

$$\begin{aligned} A &= \frac{\sum_{i=1}^n (3x_i^2 - 2)}{n} \\ &= 3 \frac{\sum_{i=1}^n x_i^2}{n} - 2 \\ &= 3(\sigma^2 + m^2) - 2 \end{aligned}$$

로 표현할 수 있다.

문제 1. 17. 평균과 표준편차를 계산할 때, 계산량을 줄이기 위하여 점화식을 이용하는 경우가 있다. 아래 식들이 성립함을 보여라.

$$\bar{x}_{j+1} = \bar{x}_j + \frac{x_{j+1} - \bar{x}_j}{j+1}$$

$$\sigma_{j+1}^2 = \frac{j}{j+1} \sigma_j^2 + j(\bar{x}_{j+1} - \bar{x}_j)^2$$

먼저, 평균에 대한 식부터 먼저 해보자.

$$\begin{aligned}\bar{x}_{j+1} &= \frac{\sum_{i=1}^{j+1} x_i}{j+1} \\ &= \frac{\sum_{i=1}^j x_i + x_{j+1}}{j+1} \\ &= \frac{j\bar{x}_j + x_{j+1}}{j+1} \\ &= \bar{x}_j + \frac{x_{j+1} - \bar{x}_j}{j+1}\end{aligned}$$

그 다음, 표준편차에 대한 식을 구성하여 보자.

$$\begin{aligned}\sigma_{j+1}^2 &= \frac{\sum_{i=1}^{j+1} x_i^2}{j+1} - \bar{x}_{j+1}^2 \\ &= \frac{\sum_{i=1}^j x_i^2 + x_{j+1}^2}{j+1} - \bar{x}_{j+1}^2 \\ &= \frac{j(\sigma_j^2 + \bar{x}_j^2) + x_{j+1}^2}{j+1} - \bar{x}_{j+1}^2 \\ &= \frac{j}{j+1} \sigma_j^2 + \frac{1}{j+1} (j\bar{x}_j^2 + x_{j+1}^2 - (j+1)\bar{x}_{j+1}^2) \\ &= \frac{j}{j+1} \sigma_j^2 + \frac{1}{j+1} (j\bar{x}_j^2 + ((j+1)\bar{x}_{j+1} - j\bar{x}_j)^2 - (j+1)\bar{x}_{j+1}^2) \\ &= \frac{j}{j+1} \sigma_j^2 + \frac{1}{j+1} ((j^2 + j)\bar{x}_j^2 + (j^2 + j)\bar{x}_{j+1}^2 - 2j(j+1)\bar{x}_j\bar{x}_{j+1}) \\ &= \frac{j}{j+1} \sigma_j^2 + j(\bar{x}_{j+1} - \bar{x}_j)^2\end{aligned}$$

위와 같이 계산하면 증명이 완료된다.

문제 2. 1. 주사위를 던져 나오는 수를 확률변수 X 라고 하자. X 의 확률분포를 구하여라.

$$P(X = i) = \frac{1}{6}, i = 1, 2, \dots, 6$$

문제 2. 2. X 를 주사위를 던져 나오는 수로 정의하자. X 가 짝수일 확률을 구하여라.

주사위 1, 2, 3, 4, 5, 6에서 짝수인 수는 2, 4, 6이다. 따라서 확률은

$$\frac{3}{6} = \frac{1}{2}$$

문제 2. 3. 단어 DOGS에서 임의로 한 알파벳을 뽑아 그 알파벳의 알파벳상 순서를 확률변수 X 라 정의하자. 예를 들어, D를 뽑았다면 $X = 4$ 이다. X 의 확률분포를 구하라.

$$P(X = 1) = \frac{1}{4}, P(X = 2) = \frac{1}{4}, P(X = 15) = \frac{1}{4}, P(X = 19) = \frac{1}{4}$$

문제 2. 4. 문제 2.3.에서 다른 확률변수 X 에 대하여, 누적분포함수의 그래프를 그려라.

$x = 4, 7, 15, 19$ 일 때 0.25씩 증가하는 계단형 그래프가 그려진다.

문제 2. 5. 두 개의 주사위를 던졌을 때 두 주사위 눈의 합을 확률변수 Y 라고 하자. $E[Y]$ 의 값은?

한 주사위를 던졌을 때의 기댓값은 3.5이다. 두 개의 주사위를 던지므로 둘을 합쳐 7이 $E[Y]$ 가 된다.

문제 2. 6. 이산확률변수의 분산도 제평균제를 만족함을 보여라. 이로부터, $E[X^2]$ 를 $V(X)$ 와 $E(X)$ 를 이용해 표시해 보아라.

$$\begin{aligned} V(X) &= \sum (x - m)^2 p(x) \\ &= \sum x^2 p(x) - 2m \sum x p(x) + m^2 p(x) \\ &= E[X^2] - E[X]^2 \end{aligned}$$

로 이산확률변수에 대해서도 제평균제가 만족한다. 이를 잘 활용하면

$$E[X^2] = V(X) + E[X]$$

임을 재확인할 수 있다.

문제 2. 7. 주사위를 던져 나오는 눈을 X 라는 확률변수라 하자. X 의 분산을 구하여라.

주사위를 던져 나오는 눈의 평균은 3.5이고, 이로부터 분산의 정의를 이용하여 계산하면

$$V(X) = (2.5)^2 \frac{2}{6} + (1.5)^2 \frac{2}{6} + (0.5)^2 \frac{2}{6} = \frac{35}{12}$$

임을 알 수 있다.

- 1) $E[aX + b] = aE[X] + b$ (기댓값의 선형성)
- 2) $V(aX + b) = a^2 V(X)$
- 3) $\sigma(aX) = |a|\sigma(X)$

문제 2. 8. 위 박스의 증명을 하여라.

여기서는 이산확률변수에 대해서만 증명하였다. \sum 을 \int 로만 바꾸면, 연속확률변수에 대해서도 동일한 방식으로 증명이 가능하다.

1)

$$\begin{aligned} E[aX + b] &= \sum (ax + b)p(x) \\ &= a \sum xp(x) + b \sum p(x) \\ &= aE[X] + b \end{aligned}$$

2)

$$\begin{aligned} V(aX + b) &= \sum (aX + b - am - b)^2 p(x) \\ &= \sum a^2(X - m)^2 p(x) \\ &= a^2V(X) \end{aligned}$$

3)

$$\begin{aligned} \sigma(aX) &= \sqrt{V(aX)} \\ &= \sqrt{a^2V(X)} \\ &= |a|\sigma(X) \end{aligned}$$

문제 2. 9. 90년대 CD 5장과 최신가요 CD 4장이 들어있는 음반에서 임의로 3장의 CD를 뽑고자 한다. 뽑힌 90년대 CD의 수를 확률변수 X , 최신가요 CD의 수를 확률변수 Y 라고 할 때, X 와 Y 의 결합확률질량함수를 구하여라.

$$\begin{aligned} P(X = 3, Y = 0) &= \frac{5C_3}{9C_3} = \frac{5}{42} \\ P(X = 2, Y = 1) &= \frac{5C_{24}C_1}{9C_3} = \frac{10}{21} \\ P(X = 1, Y = 2) &= \frac{5C_{14}C_2}{9C_3} = \frac{5}{14} \\ P(X = 0, Y = 3) &= \frac{4C_3}{9C_3} = \frac{1}{21} \end{aligned}$$

문제 2. 10. n 번째 주사위에서 나오는 수를 확률변수 X_n 이라 두자. $X_1 + X_2$ 와 X_1 은 독립인가?

$P(X_1 = 1, X_1 + X_2 = 3) = \frac{1}{36}$ 인 반면 $P(X_1 = 1) \cdot P(X_1 + X_2 = 3) = \frac{1}{108}$ 로 독립이 아니다.

문제 2. 11. 세 번째 식, 즉 $E[aX + bY] = aE[X] + bE[Y]$ 을 보여라.

$$\begin{aligned} E[aX + bY] &= \sum_x \sum_y (ax + by)p(x, y) \\ &= \sum_x \sum_y axp(x, y) + \sum_y \sum_x byp(x, y)) \\ &= \sum_x axp_x(x) + \sum_y byp_y(y) \\ &= a \sum_x xp_x(x) + b \sum_y yp_y(y) \\ &= aE[X] + bE[Y] \end{aligned}$$

문제 2. 12. $Cov(aX + b, cY + d) = acCov(X, Y)$ 임을 보여라.

$$\begin{aligned}
Cov(aX + b, cY + d) &= E[(aX + b)(cY + d)] - E[aX + b]E[cY + d] \\
&= E[acXY + bcY + adX + bd] - (aE[X] + b)(cE[Y] + d) \\
&= acE[XY] + bcE[Y] + adE[X] + bd - acE[X]E[Y] - bcE[Y] - adE[X] - bd \\
&= ac(E[XY] - E[X]E[Y]) \\
&= acCov(X, Y)
\end{aligned}$$

가 성립한다.

문제 2. 13. $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$ 임을 보여라.

$$\begin{aligned}
Cov(X_1 + X_2, Y) &= E[(X_1 + X_2)Y] - E[X_1 + X_2]E[Y] \\
&= E[X_1Y] + E[X_2Y] - E[X_1]E[Y] - E[X_2]E[Y] \\
&= Cov(X_1, Y) + Cov(X_2, Y)
\end{aligned}$$

이 성립한다.

문제 2. 14. 아래 식을 보여라.

$$Cov\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m Cov(X_i, Y_j)$$

$$\begin{aligned}
Cov\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) &= Cov\left(\sum_{i=1}^{n-1} X_i, \sum_{j=1}^m Y_j\right) + Cov\left(X_n, \sum_{j=1}^m Y_j\right) \\
&= \dots \\
&= \sum_{i=1}^n Cov(X_i, \sum_{j=1}^m Y_j) \\
&= \sum_{i=1}^n Cov\left(\sum_{j=1}^m Y_j, X_i\right) \\
&= \sum_{i=1}^n \sum_{j=1}^m Cov(Y_j, X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^m Cov(X_i, Y_j)
\end{aligned}$$

임을 확인할 수 있다.

문제 2. 15. 아래 식을 보여라.

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n Cov(X_i, X_j)$$

$$\begin{aligned}
V\left(\sum_{i=1}^n X_i\right) &= Cov\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right) \\
&= \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j) \\
&= \sum_{i=1}^n V(X_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n Cov(X_i, X_j)
\end{aligned}$$

임을 확인할 수 있다.

문제 2. 16. 다섯 개의 트랜지스터가 테스트를 앞두고 있다. 무작위로 그들을 배열하고, 그들이 불량품인지 확인하려 한다. 다섯 개 중 3개의 트랜지스터가 불량이라고 할 때, N_1 을 첫 불량품이 밝혀질 때까지 수행되는 테스트의 횟수, N_2 를 그 이후 두 번째 불량품이 밝혀질 때까지 수행되는 테스트의 횟수라고 하자. N_1 과 N_2 의 결합확률질량함수를 구하여라.

$$\begin{aligned}
P(N_1 = 1, N_2 = 1) &= \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10} \\
P(N_1 = 1, N_2 = 2) &= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5} \\
P(N_1 = 1, N_2 = 3) &= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = \frac{1}{10} \\
P(N_1 = 2, N_2 = 1) &= \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{5} \\
P(N_1 = 2, N_2 = 2) &= \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{10} \\
P(N_1 = 3, N_2 = 1) &= \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}
\end{aligned}$$

임을 단순 계산으로부터 알 수 있다.

문제 2. 17. 매일 밤 기상학자들은 다음 날 비가 올 확률을 제시한다. 그들의 예측을 평가하기 위하여, 아래와 같이 점수를 매기기로 했다. 만약 기상학자가 p 의 확률로 비가 내린다고 했다면, 비가 내일 경우 $1 - (1-p)^2$ 점을, 비가 내리지 않을 경우 $1 - p^2$ 점을 받는다. 이를 수행하여 가장 높은 점수를 받은 사람을 최고의 기상학자로 생각하기로 했다. 그들이 그들의 점수 기댓값을 최대한으로 만들기 위해서 어떤 전략을 취해야 할까? 단, 그들이 실제로 생각하는 비가 올 확률은 $p*$ 이라고 두자.

기상학자는 비가 내릴 확률이 $p*$ 이라고 굳게 믿고 있는 상황이다. 만약 그가 p 의 확률로 비가 내린다고 했을 경우, 얻을 점수의 기댓값은

$$(1 - p^2)(1 - p*) + (1 - (1 - p)^2)(p*)$$

점이다. 이를 정리해 보면,

$$\begin{aligned}
(1 - p^2)(1 - p*) + (1 - (1 - p)^2)(p*) &= 1 - p^2 - p* + p^2 p* + 2pp* - p^2 p* \\
&= -p^2 + 2p*p + 1 - p* \\
&= -(p - p*)^2 + 1 - p* + p*^2
\end{aligned}$$

이다. $p*$ 은 고정된 값이므로, $p = p*$ 일 때 이 값은 최대가 될 것이다. 따라서 자신이 믿는 그대로 강우 확률을 설명하면 된다.

문제 2. 18. 보험회사는 교통사고가 일어났을 때 대물보험 가입자에게 A 원을 보상한다. 만약 일년 동안 가입자가 교통사고를 일으킬 확률을 p 라 예상하고 있었다면, 보험사의 기대수익이 A 의 10퍼센트가 되려면 연 보험료가 얼마인가?

기대수익은 보험료에서 대물보상금과 사고 발생 비율을 곱한 것을 빼면 된다. 즉,

$$0.1A = (\text{보험료}) - Ap$$

인 것이다. 따라서, 적정한 연 보험료는 $A(0.1 + p)$ 이다.

문제 2. 19. 명반을 찾는 TV쇼에서 148명을 데려다 놓고 락밴드 A, B, C, D 에 대한 선호도 조사를 진행했다. 각각 40, 33, 25, 50명이 A, B, C, D 를 선호한다고 밝혔다. 148명 중 한 명의 방청객이 임의로 선택되었을 때, 확률변수 X 를 해당 사람이 좋아하는 밴드의 선호인 수라고 정의하자. 반대로, 밴드 중 하나를 임의로 선택한 다음 그 밴드의 선호인 수를 확률변수 Y 라고 두자. $E[X]$ 와 $E[Y]$ 를 비교하여라.

$$E[X] = 40 \times \frac{40}{148} + 33 \times \frac{33}{148} + 25 \times \frac{25}{148} + 50 \times \frac{50}{148} = 39.284$$

$$E[Y] = \frac{1}{4}(40 + 33 + 25 + 50) = 37$$

로, $E[X]$ 의 값이 더욱 크다. 이는 X 의 경우 선호인 수가 많은 밴드가 선택될 확률이 더 높기 때문이다.

문제 2. 20. $p_i = P\{X = i\}$ 이며 $p_1 + p_2 + p_3 = 1$ 이다. 만약 $E[X] = 2$ 라면, $V(x)$ 를 최소화/최대화시키는 p_1, p_2, p_3 의 값을 각각 무엇인가?

먼저, $p_1 + p_2 + p_3 = 1$ 인 것과 $p_1 + 2p_2 + 3p_3 = 2$ 인 것으로부터 p_2, p_3 을 p_1 으로 표시할 수 있다. 둘째 식에서 첫째 식을 빼면 $p_2 + 2p_3 = 1$ 이며, $p_2 = 1 - 2p_3$ 이므로, $p_1 + 1 - p_3 = 1$ 이다. 즉 $p_3 = p_1$ 이고, $p_2 = 1 - 2p_1$ 이다. $V(X) = p_1 + 0 + p_3 = 2p_1$ 임을 알고 있기에, 최대화되는 경우에는 0.5, 0, 0.5이며 최소화될 경우에는 0, 1, 0이다.

문제 2. 21. X, Y 의 분산을 각각 σ_X^2, σ_Y^2 이라 둔 상태에서, $V\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \geq 0$ 임을 이용하여 $\text{Corr}(X, Y) \geq -1$ 임을 보여라. 동일한 방식으로, $\text{Corr}(X, Y) \leq 1$ 임도 보여라.

$$\begin{aligned} 0 &\leq V\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= \text{Cov}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}, \frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= \frac{\text{Cov}(X, X)}{\sigma_X^2} + 2 \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} + \frac{\text{Cov}(Y, Y)}{\sigma_Y^2} \\ &= \frac{V(X)}{\sigma_X^2} + 2 \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} + \frac{V(Y)}{\sigma_Y^2} \\ &= 2 + 2\text{Corr}(X, Y) \end{aligned}$$

이다. 즉 이를 정리하면 $\text{Corr}(X, Y) \geq -1$ 임을 확인할 수 있다.

$$\begin{aligned} 0 &\leq V\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) \\ &= 2 - 2\text{Corr}(X, Y) \end{aligned}$$

이므로, $\text{Corr}(X, Y) \leq 1$ 임 역시 확인가능하다.

문제 2. 22. n 번의 독립시행에 대하여, 각 결과는 1, 2, 3중에 하나로 p_1, p_2, p_3 의 확률로 등장한다. 만약 N_i 를 i 라는 결과가 나온 시행 횟수라고 둔다면, $Cov(N_1, N_2) = -np_1p_2$ 임을 보여라. 또한, $Corr(N_1, N_2)$ 의 부호를 결정하고 이를 직관적으로 설명하여라.

i 번째 시행에 대하여, X_i 는 1이 나오면 1, 나머지 경우에는 0이며 Y_i 는 2가 나오면 1, 나머지 경우에는 0인 확률변수라 하자. 그러면

$$N_1 = \sum_{i=1}^n X_i, N_2 = \sum_{i=1}^n Y_i$$

라고 할 수 있다. 그러면

$$\begin{aligned} Cov(N_1, N_2) &= Cov\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, Y_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n (E[X_i Y_j] - E[X_i]E[Y_j]) \\ &= \sum_{i=1}^n (E[X_i Y_i] - E[X_i]E[Y_i]) \quad (i \neq j \rightarrow Cov(X_i, Y_j) = 0) \\ &= \sum_{i=1}^n (-p_1 p_2) = -np_1 p_2 \end{aligned}$$

가 성립함을 알 수 있다. 또한, 분산은 항상 양수이므로 공분산을 표준편차의 곱으로 나눈 $Corr(N_1, N_2)$ 역시 음수이다. 이는 일반적으로 1이 많이 나올수록 2가 적게 나올 수밖에 없다는 예상과 일치한다.

문제 2. 23. 만약 확률변수 X_1, X_2 가 동일한 확률질량분포함수를 가지고 있다면, $Cov(X_1 - X_2, X_1 + X_2) = 0$ 임을 보여라.

$Cov(X_1 - X_2, X_1 + X_2) = Cov(X_1, X_1) + Cov(X_1, X_2) - Cov(X_2, X_1) - Cov(X_2, X_2) = \sigma^2 - \sigma^2 = 0$ 이다.

문제 2. 24. 함수 $\phi(t)$ 를

$$\phi(t) = E[e^{tX}] = \sum_x e^{tx} P(X=x)$$

라고 정의하자.

$$\phi'(0) = E[X]$$

$$\phi''(0) = E[X^2]$$

임을 보여라. 수학적 귀납법을 이용하여, $n \leq 1$ 인 자연수에 대해

$$\phi^n(0) = E[X^n]$$

임을 보여라.

$$\phi(t) = \sum_x e^{tx} P(X=x)$$

이므로

$$\phi'(t) = \frac{d}{dt} \sum_x e^{tx} P(X=x) = \sum_x x e^{tx} P(X=x)$$

○] 고, $\phi'(0) = \sum_x xP(X=x) = E[X]$ 가 된다.

$$\phi''(t) = \frac{d}{dt} \sum_x xe^{tx} P(X=x) = \sum_x x^2 e^{tx} P(X=x)$$

이므로 $\phi''(0) = \sum_x x^2 P(X=x) = E[X^2]$ 이다. 또한, 수학적 귀납법을 이용하면

$$\phi^{(n)}(t) = \frac{d}{dt} \sum_x x^{n-1} e^{tx} P(X=x) = \sum_x x^n e^{tx} P(X=x)$$

이므로 $\phi^{(n)}(0) = \sum_x x^n P(X=x) = E[X^n]$ 이 됨을 확인할 수 있다.

문제 2. 25. 위에서 정의한 함수 $\phi(t)$ 를 적률생성함수라 한다. 두 확률변수 X 와 Y 의 적률생성함수를 $\phi_X(t)$, $\phi_Y(t)$ 라고 하자. 만약 X 와 Y 가 독립이라면, $X+Y$ 의 적률생성함수 $\phi_{X+Y}(t)$ 는

$$\phi_{X+Y}(t) = \phi_X(t) + \phi_Y(t)$$

임을 보여라.

두 확률변수 X 와 Y 가 독립이라면 e^{tX} 와 e^{tY} 도 독립이다. 따라서, $E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY})$ 이다. 즉 이를 정리하면

$$\phi_{X+Y}(t) = E(e^{t(X+Y)}) = \phi_X(t)\phi_Y(t)$$

임을 쉽게 확인가능하다.

문제 2. 26. 확률변수 X 와 Y 가 독립이라면, 확률변수 $h(X)$ 와 $g(Y)$ 도 독립임을 보여라.

$$\begin{aligned} P(h(X) \in A, g(Y) \in B) &= P(X \in h^{-1}(A), Y \in g^{-1}(B)) \\ &= P(X \in h^{-1}(A))P(Y \in g^{-1}(B)) \quad (X, Y \text{는 독립}) \\ &= P(h(X) \in A)P(g(Y) \in B) \end{aligned}$$

가 성립하기에, 둘은 독립이다.

문제 2. 27. 모두 같은 확률분포를 가지며 서로 독립인 확률변수 X_i 에 대하여, $Cov\left(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n X_i}{n} - X_1\right)$ 의 값을 구하여라. 단, $n \geq 2$ 라고 한다.

$$\begin{aligned} Cov\left(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n X_i}{n} - X_1\right) &= Cov\left(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n X_i}{n}\right) - Cov\left(\frac{\sum_{i=1}^n X_i}{n}, X_1\right) \\ &= Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) - \frac{1}{n} \sum_{i=1}^n Cov(X_i, X_1) \\ &= \frac{1}{n^2} \sum_{i=1}^n Var(X_i) - \frac{1}{n} Cov(X_1, X_1) \\ &= \frac{n}{n^2} Var(X_1) - \frac{1}{n} Var(X_1) = 0 \quad (X_1, \dots, X_n \text{는 같은 확률분포를 가짐}) \end{aligned}$$

문제 2. 28. 최근 EBS에서 문해력에 관한 다큐를 촬영하기 위해, 25명의 초등학생을 모아두고 수업 중 모르는 단어를 모두 보고하라고 이야기했다. 선행연구 결과, 학생들은 평균 40개의 단어를 몰랐으며, 모르는 단어의 표준편차는 20개였다.

25명의 모르는 단어 수를 모두 합친 것이 1100개를 넘을 확률이 10/11 이하임을 증명하여라.

아이들이 모르는 단어의 평균 개수를 X 라는 확률변수로 두자. 그러면, X 의 기댓값은 25명의 학생이 평균 40개의 단어를 모르므로 1000이다. 그렇다면, 마코프의 부등식에 의해,

$$P(X \geq 1100) \leq \frac{E(X)}{1100} = \frac{10}{11}$$

임을 확인할 수 있다.

문제 2. 29. 서울과학고의 기초통계학 학점 부여 기준을 보면, 30퍼센트의 학생은 A , 30퍼센트의 학생은 B , 20퍼센트의 학생은 C , 그리고 20퍼센트의 학생은 D 나 F 를 받게 되어 있다. 수학 선생님께서는 과목별로 독립적으로 학점을 부여하신다. 진섭이는 이번 학기에 수학 과목을 세 개 듣는다. 확률변수 X 를 진섭이가 받은 A 의 수라고 하자.

1) X 의 확률질량함수를 구하여라.

2) X 의 누적분포함수를 구하여라.

1) 과목을 3개 듣고 있으므로, X 의 값은 0, 1, 2, 3 중 하나이다.

$$P(X = 0) = (0.7)^3 = 0.343$$

$$P(X = 1) = 3 \times (0.7)^2(0.3) = 0.441$$

$$P(X = 2) = 3 \times (0.7)(0.3)^2 = 0.189$$

$$P(X = 3) = (0.3)^3 = 0.027$$

2) 누적분포함수 $F(x)$ 는 아래와 같이 표현할 수 있다.

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.343 & 0 \leq x < 1 \\ 0.784 & 1 \leq x < 2 \\ 0.973 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

문제 2. 30. 확률변수 X 가 확률질량함수를 $P(X = n) = \frac{1}{2^n}$ 으로 가진다. 단, $n \geq 1$ 이다. $E[X]$ 의 값을?

$$\begin{aligned} E(X) &= \sum_{n=1}^{\infty} nP(X = n) \\ &= \sum_{n=1}^{\infty} \frac{n}{2^n} \\ &= (x \sum_{n=1}^{\infty} nx^{n-1})|_{x=\frac{1}{2}} \\ &= (x \frac{d}{dx} (\sum_{n=1}^{\infty} x^n))|_{x=\frac{1}{2}} \\ &= (x \frac{d}{dx} (\frac{x}{1-x}))|_{x=\frac{1}{2}} \\ &= x \cdot \frac{(1-x)+x}{(1-x)^2}|_{x=\frac{1}{2}} \\ &= \frac{x}{(1-x)^2}|_{x=\frac{1}{2}} = 2 \end{aligned}$$

문제 2. 31. 확률변수 Y 가 확률질량함수를 $P(Y = n) = P(Y = \frac{1}{n}) = \frac{1}{2^{n+1}}$ 로 가진다. 단 $n \geq 2$ 일 때며, $n = 1$ 이라면 확률은 0.5이다. $E[Y]$ 의 값을 구하여라.

$$\begin{aligned}
E(Y) &= \sum_{n=2}^{\infty} nP(Y = n) + \sum_{n=2}^{\infty} \frac{1}{n} P(Y = \frac{1}{n}) + 0.5 \\
&= \sum_{n=2}^{\infty} n \frac{1}{2^{n+1}} + \sum_{n=2}^{\infty} \frac{1}{n} \frac{1}{2^{n+1}} + 0.5 \\
&= \frac{1}{2} \left(\sum_{n=1}^{\infty} n \frac{1}{2^n} - 0.5 \right) + \frac{1}{2} \left(\sum_{n=2}^{\infty} \frac{1}{n} \frac{1}{2^n} \right) + 0.5 \\
&= \frac{5}{4} + \frac{1}{2} \left(\sum_{n=2}^{\infty} \frac{1}{n} x^n \right) |_{x=\frac{1}{2}} \\
&= \frac{5}{4} + \frac{1}{2} \left(\int_0^x \left(\sum_{n=2}^{\infty} t^{n-1} \right) dt \right) |_{x=\frac{1}{2}} \\
&= \frac{5}{4} + \frac{1}{2} \left(\int_0^x \frac{t}{1-t} dt \right) |_{x=\frac{1}{2}} \\
&= \frac{5}{4} + \frac{1}{2} \left(\int_0^x -1 + \frac{1}{1-t} dt \right) |_{x=\frac{1}{2}} \\
&= \frac{5}{4} + \frac{1}{2} (-x - \ln(1-x)) |_{x=\frac{1}{2}} \\
&= 1 + \frac{1}{2} \ln 2
\end{aligned}$$

문제 2. 32. 이산확률변수 X 에 대하여, $E[X^2] < \infty$ 라면 $E[X] < \infty$ 임을 보여라.

$$V(X) = E(X^2) - (E(X))^2 \geq 0$$

이므로,

$$E(X) \leq \sqrt{E(X^2)} < \infty$$

임이 자명하다.

문제 2. 33. 유리수의 집합 \mathbb{Q} 는 셀 수 있는 집합이라고 알려져 있다. 또한, 두 실수 사이에는 항상 유리수가 존재한다. 이제 구간 $(0, 1)$ 에 있는 모든 유리수를 r_1, r_2, \dots , 와 같은 식으로 나열한 이후에 확률변수 X 를 확률질량함수 $P(X = r_n) = \frac{1}{2^n}$ 이라 정의하자. 그렇다면, $(0, 1)$ 의 어떤 열린 부분구간에서도 X 의 누적분포함수가 상수함수가 아님을 보여라.

만약 $(0, 1)$ 의 어떤 열린 부분구간 (a, b) 가 존재하여 그 범위에서 누적분포함수가 상수함수라고 하자. 그러면 $a < c < d < b$ 인 실수 c, d 가 존재하여 누적분포함수 $F(x)$ 에 대해 $F(c) = F(d)$ 다. 그러나 c, d 사이에는 유리수 p 가 존재하며, p 는 유리수이므로 나열된 수열 중 p 와 같은 것이 있다. 즉, 어떤 자연수 q 에 대하여 $p = r_q$ 이며 $P(X = p) = P(X = r_q) = \frac{1}{2^q}$ 이다. 그런데,

$$0 = F(d) - F(c) = P(c < X \leq d) \leq P(X = r_q) = \frac{1}{2^q} > 0$$

이 성립하며 이는 모순이다. 따라서, 그런 a, b 는 존재하지 않는다. 따라서 $(0, 1)$ 의 어떤 열린 부분구간에서도 X 의 누적분포함수는 상수함수가 될 수 없다.

문제 2. 34. 음이 아닌 정수를 값으로 갖는 이산확률변수 X 를 생각하자. X 의 기댓값은 유한하다고 가정하자.

- 1) $E[X] = \sum_{k=0}^{\infty} (1 - F_X(k))$ 임을 보여라. 단, $F_X(k)$ 는 X 의 누적분포함수이다.
 2) $E[X^2] - E[X] = 2 \sum_{k=1}^{\infty} kP(X > k)$ 임을 보여라.

1)

$$\begin{aligned}
 E(X) &= \sum_{k=0}^{\infty} kP(X = k) \\
 &= \sum_{k=1}^{\infty} (kP(X > k-1) - kP(X > k)) \\
 &= \sum_{k=1}^{\infty} k(1 - F_X(k-1)) - \sum_{k=1}^{\infty} k(1 - F_X(k)) \\
 &= \sum_{k=1}^{\infty} (k-1)(1 - F_X(k-1)) + \sum_{k=1}^{\infty} (1 - F_X(k-1)) - \sum_{k=1}^{\infty} k(1 - F_X(k)) \\
 &= \sum_{k=0}^{\infty} k(1 - F_X(k)) + \sum_{k=0}^{\infty} (1 - F_X(k)) - \sum_{k=1}^{\infty} k(1 - F_X(k)) \\
 &= \sum_{k=1}^{\infty} k(1 - F_X(k)) - \sum_{k=1}^{\infty} k(1 - F_X(k)) + \sum_{k=0}^{\infty} (1 - F_X(k)) = \sum_{k=0}^{\infty} (1 - F_X(k))
 \end{aligned}$$

임을 계산을 통해 확인할 수 있다.

2)

$$\begin{aligned}
 E(X^2) - E(X) &= E(X^2 - X) \\
 &= E(X(X-1)) \\
 &= \sum_{k=2}^{\infty} k(k-1)P(X = k) \\
 &= \sum_{k=2}^{\infty} k(k-1)(P(X > k-1) - P(X > k)) \\
 &= \sum_{k=2}^{\infty} k(k-1)P(X > k-1) - \sum_{k=2}^{\infty} k(k-1)P(X > k) \\
 &= \sum_{k=1}^{\infty} (k+1)kP(X > k) - \sum_{k=2}^{\infty} k(k-1)P(X > k) \\
 &= 2P(X > 1) + \sum_{k=2}^{\infty} 2kP(X > k) \\
 &= 2 \sum_{k=1}^{\infty} kP(X > k)
 \end{aligned}$$

임도 계산으로 확인가능하다.

문제 3. 1. 이항정리를 이용하여 이항분포에서 모든 가능한 값이 나올 확률의 합이 1임을 보여라.

Note : 주어진 확률질량함수에 대해서, 그들의 합이 1임은 항상 당연하다.

$B(n, p)$ 를 따르는 확률변수 X 에서 $k = 0, 1, \dots, n$ 에 대하여

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

이 성립한다.

$$\begin{aligned} \sum_{k=0}^n P(X = k) &= \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} \\ &= (p + 1 - p)^n = 1^n = 1 \end{aligned}$$

임을 계산으로부터 확인할 수 있다.

문제 3. 2. 이항분포의 평균과 분산은 책에 나온 것처럼 콤비네이션을 잘 조작하여 증명해낼 수도 있다. 혹은, 2장 연습문제에서 다룬 적률생성함수를 구함으로써 알아낼 수도 있다. 이 연습문제에서는 콤비네이션을 조작하여 기댓값과 분산이 각각 np 임과 $np(1-p)$ 임을 보여라.

$$\begin{aligned} E(X) &= \sum_{k=0}^n k P(X = k) \\ &= \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n \frac{n!}{(n-k)!(k-1)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n np \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{s=0}^{n-1} \binom{n-1}{s} p^s (1-p)^{n-1-s} \\ &= np \end{aligned}$$

$$\begin{aligned} E(X^2 - X) &= \sum_{k=0}^n k(k-1) P(X = k) \\ &= \sum_{k=2}^n k(k-1) k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=2}^n \frac{n!}{(n-k)!(k-2)!} p^k (1-p)^{n-k} \\ &= \sum_{k=2}^n n(n-1)p^2 \frac{(n-2)!}{(n-k)!(k-2)!} p^{k-2} (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{s=0}^{n-2} \binom{n-2}{s} p^s (1-p)^{n-2-s} \\ &= n(n-1)p^2 \end{aligned}$$

이므로

$$V(X) = E(X^2) - E(X)^2 = E(X^2 - X) + E(X) - E(X)^2 = n(n-1)p^2 + np - n^2p = np(1-p)$$

문제 3. 3. 책에서는 미분법을 이용한 증명도 소개하고 있다. 미분법을 이용하여 증명하여라.

$n \geq 2$ 일 때는 간단하게 증명이 가능하므로, 그것보다 $n > 1$ 일 때를 가정하고 풀자.

$$\begin{aligned} E(X) &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= (1-p)^n \sum_{k=0}^n k \binom{n}{k} \left(\frac{p}{1-p}\right)^k \\ &= p(1-p)^{n-1} \frac{d}{dx} \left(\sum_{k=0}^n \binom{n}{k} x^k \right) |_{x=p/(1-p)} \\ &= p(1-p)^{n-1} \frac{d}{dx} (1+x)^n |_{x=p/(1-p)} \\ &= p(1-p)^{n-1} n \left(\frac{1}{1-p}\right)^{n-1} \\ &= np \end{aligned}$$

$$\begin{aligned} E(X^2 - X) &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} \\ &= (1-p)^n \sum_{k=2}^n k(k-1) \binom{n}{k} \left(\frac{p}{1-p}\right)^k \\ &= p^2(1-p)^{n-2} \frac{d^2}{dx^2} \left(\sum_{k=2}^n \binom{n}{k} x^k \right) |_{x=p/(1-p)} \\ &= n(n-1)p^2 \end{aligned}$$

이다. 위의 과정을 거치면 동일하게

$$V(X) = np(1-p)$$

임을 증명할 수 있다.

문제 3. 4. 이항분포 $B(n, p)$ 의 표준편차 $\sigma \geq \sqrt{np(1-p)}$ 임을 보여라. 이항분포에서 n 만 알고 p 를 모를 때, 최대의 표준편차는 얼마만큼인가?

해당 이항분포의 분산은 $np(1-p)$ 다. 따라서, 그 표준편차는 $\sqrt{np(1-p)}$ 이다. n 만 알고 p 를 모른다면, $p(1-p) \leq 0.25$ 임을 고려하면 표준편차의 최댓값은 $\frac{1}{2}\sqrt{n}$ 일 것이다.

문제 3. 5. 체비셰프 부등식을 이용한 큰 수의 법칙의 증명에서, $\varepsilon = 1$ 이고 $\sigma = 1$ 일 때 확률

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right)$$

이 0.01보다 작음이 보장되려면 n 은 무엇보다 커야 하는가?

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} = \frac{1}{n}$$

이다. 따라서 보장되려면, n 은 100보다 커야 한다.

문제 3. 6. p 의 값에 관계없이, 기하분포의 확률질량함수는 확률질량함수의 조건을 만족시킴을 보여라.

$$\begin{aligned} \sum_{k=1}^{\infty} P(X = k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p \\ &= p \sum_{k=1}^{\infty} (1-p)^{k-1} \\ &= p \cdot \frac{1}{1 - (1-p)} \\ &= 1 \end{aligned}$$

이 성립한다. 또한, $P(X = k)$ 는 모두 0과 1 사이의 값이다. 따라서, 기하분포의 확률질량함수는 원하는 조건을 만족시킨다.

문제 3. 7. 이대호 선수가 안타를 치기 위해 필요한 타석의 수 X 는 $GEO(0.3)$ 을 따른다고 한다. 오늘 이대호 선수가 앞선 두 타석에서 안타를 치지 못했고, 세번째 타석에 들어섰다. 이대호 선수가 네 번째 타석까지 소화한다고 할 때, 오늘 안에 안타를 칠 확률은?

이대호 선수가 앞선 두 타석에서 어떤 결과를 얻든지 간에, 남은 두 타석에서 안타를 칠 확률은 $P(X \leq 2)$ 와 같다.

$$P(X \leq 2) = P(X = 1) + P(X = 2) = 0.3 + 0.7 \cdot 0.3 = 0.51$$

이므로, 3할타자 이대호는 0.51의 확률로 오늘 안에 안타를 친다.

문제 3. 8. 초기하분포의 확률질량함수가 확률질량함수의 조건을 만족시킴을 보여라.

먼저, 초기하분포에서 $P(X = k)$ 는 항상 0보다는 큼을 확인할 수 있다. 만약

$$\sum_{k=0}^n P(X = k) = 1$$

임을 증명한다면, $P(X = k)$ 가 1보다 작은 것 역시 자명하므로 확률질량함수의 두 조건을 충족시킴을 알 수 있게 된다. 따라서, 위의 시그마 식을 보여 보자.

$$\begin{aligned} \sum_{k=0}^n P(X = k) &= \sum_{k=0}^n \frac{{}_M C_{kN-M} C_{n-k}}{{}_N C_n} \\ &= \frac{1}{{}_N C_n} \sum_{k=0}^n {}_M C_{kN-M} C_{n-k} \\ &= 1 \end{aligned}$$

임을 확인할 수 있다. 이때, 마지막 등식은 $(1+x)^N$ 에서 x^n 의 계수와 $(1+x)^M(1+x)^{N-M}$ 에서 x^n 의 계수가 같은 것으로부터 유도해낼 수 있다.

문제 3. 9. X 의 분포가 $HYP(N, M, n)$ 일 때, i 번째 시행에서 원하는 공이 나오면 1, 아니면 0을 가지는 확률변수 X_i 를 생각하자. 그러면 $X_1 + X_2 + \cdots + X_n = X$ 라고 생각할 수도 있을 것이다. 모든 i 에 대하여, $P(X_i = 1) = \frac{M}{N}$ 임을 보여 보아라.

즉, 몇 번째로 공을 뽑는지에 상관 없이 원하는 공을 뽑을 확률은 초기 상태에서 원하는 공을 뽑을 확률과 같다는 것이다. 먼저, $i = 1$ 일 때는 $P(X_i = 1) = M/N$ 임이 자명하다. $i > 1$ 일 때를 생각하여 보자. 1번재부터 $i - 1$ 번재까지에 대하여 원하는 공이 소진된 개수 Y 는 $HYP(N, M, i - 1)$ 을 따른다. 소진된 공의 개수가 x 개라면, i 번재 공이 원하는 공일 확률은 $(M - x)/(N - i + 1)$ 이다. 따라서,

$$P(X_i = 1) = \sum_{x=0}^{i-1} \frac{M-x}{N-i+1} P(Y=x) = \sum_{x=0}^{i-1} \frac{M-x}{N-i+1} \frac{MC_{xN-M}C_{i-1-x}}{NC_{i-1}}$$

임을 확인 가능하다.

$$\begin{aligned} \sum_{x=0}^{i-1} \frac{M-x}{N-i+1} \frac{MC_{xN-M}C_{i-1-x}}{NC_{i-1}} &= \sum_{x=0}^{i-1} \frac{M}{N} \frac{M-1}{N-1} \frac{C_{xN-M}C_{i-1-x}}{C_{i-1}} \\ &= \frac{M}{N} \sum_{x=0}^{i-1} P(Z=x) \quad \text{where } Z \sim HYP(N-1, M-1, i-1) \\ &= \frac{M}{N} \end{aligned}$$

문제 3. 10. 푸아송분포의 확률질량함수가 그 정의에 잘 맞음을 보여라.

$$\begin{aligned} \sum_{k=0}^{\infty} P(X=k) &= \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \exp(\lambda) = 1 \end{aligned}$$

이며, $P(X=k)$ 는 항상 0과 1 사이에 존재하는 수이다.

문제 3. 11. 책에 나온 대로 이항분포를 푸아송 근사 해보아라. 즉,

$$P(X=x) =_n C_x p^x (1-p)^{n-x}$$

가

$$P(X=x) = \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

로 변형된 이후, n 이 무한으로 감에 따라

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

가 됨을 이용해 이것이 $\frac{e^{-\lambda} \lambda^x}{x!}$ 로 근사됨을 보이라는 것이다.

위의 과정으로부터,

$$P(X=x) = \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}$$

임은 알고 있다. 그러면 n 을 충분히 크게 만들게 된다면,

$$\lim_{n \rightarrow \infty} P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda} 1^{-x} = e^{-\lambda} \frac{\lambda^x}{x!}$$

임을 확인할 수 있다.

$X \sim POI(\lambda)$ 일 때,

$$E[X] = \lambda, V(X) = \lambda$$

문제 3. 12. 위 박스를 증명하라.

$$\begin{aligned}
 E(X) &= \sum_{k=0}^{\infty} k P(X = k) \\
 &= \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\
 &= \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{k - 1!} \\
 &= \lambda \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \\
 &= \lambda
 \end{aligned}$$

$$\begin{aligned}
 E(X^2 - X) &= \sum_{k=0}^{\infty} k(k-1)P(X = k) \\
 &= \sum_{k=2}^{\infty} e^{-\lambda} \frac{\lambda^k}{(k-2)!} \\
 &= \sum_{k=0}^{\infty} \lambda^2 e^{-\lambda} \frac{\lambda^k}{k!} \\
 &= \lambda^2
 \end{aligned}$$

이므로,

$$V(X) = E(X^2 - X) + E(X) - E(X)^2 = \lambda$$

임을 확인할 수 있다.

문제 3. 13. 인기 예능 신서유기에서는 n 명의 멤버를 데리고 90년대 노래 퀴즈를 진행한다. 각각이 정답을 맞출 확률은 p 이며, 각자의 결과에 독립적이다. 나PD는 최소 반 이상이 정답을 맞췄을 때 식사를 제공하기로 했다.

- 1) p 가 어느 값 이상이어야 5명의 멤버가 도전했을 때 3명의 멤버가 도전했을 때보다 더 성공 확률이 높을 것인가?
- 2) p 가 어느 값 이상이어야 $2k+1$ 명의 멤버가 도전했을 때 $2k-1$ 명의 멤버가 도전했을 때보다 효과적일 것인가?

1)

$$10p^3(1-p)^2 + 5p^4(1-p) + p^5 > 3p^2(1-p) + p^3$$

$$10p^3(1-p)^2 + 5p^4(1-p) + p^5 - p^3 - 3p^2(1-p) > 0$$

$$10p(1-p) + 5p^2 - p(1+p) - 3 = -6p^2 + 9p - 3 = 3(2p-1)(-p+1) > 0$$

$$p > \frac{1}{2}$$

임을 확인할 수 있다.

- 2) $2k+1$ 명의 멤버가 도전했을 때, $k+2$ 명 이상이 성공할 확률을 x_k , $k+1$ 명 성공했을 확률을 y_k , k 명 성공했을 확률을 z_k 라고 하자.

$2k+1$ 명의 멤버가 도전할 때, 먼저 $2k-1$ 명이 시도하여 결과가 나오고 나머지 두 명을 남겨둔 상태라고 생각하자. 만약 x_{k-1} 의 확률로 $k+1$ 명 이상 성공했을 경우에는 남은 두 명의 결과와 관계 없이 성공한다.

만약 y_{k-1} 의 확률로 k 명만 성공했을 경우에는, 나머지 두 번의 시행에서 적어도 한 번은 성공해야 하며 그 확률은 $1 - (1-p)^2$ 이다. 만약 z_{k-1} 의 확률로 $k-1$ 명 성공했을 때는, 나머지 두 번의 시행에서 모두 성공해야 하며 그 확률은 p^2 이다. 그 외의 경우에는 성공이 불가능하다. 따라서,

$$\begin{aligned} x_k + y_k &= x_{k-1} + (2p - p^2)y_{k-1} + p^2z_{k-1} \\ &= x_{k-1} + (2p - p^2)\binom{2k-1}{k}p^k(1-p)^{k-1} + p^2\binom{2k-1}{k-1}p^{k-1}(1-p)^k \\ &= x_{k-1} + (2p^2 - p^3 + p^2 - p^3)\binom{2k-1}{k}p^{k-1}(1-p)^{k-1} \\ &= (x_{k-1} + y_{k-1}) + (3p^2 - 2p^3)\binom{2k-1}{k}p^{k-1}(1-p)^{k-1} - p\binom{2k-1}{k}p^{k-1}(1-p)^{k-1} \\ &= (x_{k-1} + y_{k-1}) + (-2p^2 + 3p - 1)\binom{2k-1}{k}p^k(1-p)^{k-1} \end{aligned}$$

임을 확인할 수 있다. $x_k + y_k$ 가 성공 확률이므로, $(-2p^2 + 3p - 1)\binom{2k-1}{k}p^k(1-p)^{k-1}$ 가 양수여야 문제의 조건을 만족한다. 따라서, 1)에서 구한 것처럼 그 범위는 $p > 1/2$ 이다.

문제 3. 14. 나PD는 다음 게임으로 n 명의 멤버들을 방에 불러 두고 그들의 모자를 모두 모아 섞었다. 그 다음, 연장자 순으로 무작위의 모자를 고르고 가져갔다. X 를 자신의 모자를 되찾은 멤버의 수라고 하자. $E[X] = 1$ 임을 보여라.

$$X_i = \begin{cases} 1 & i\text{번째 사람이 자신의 모자를 되찾은 경우} \\ 0 & 그렇지 못한 경우 \end{cases}$$

라고 두면,

$X = X_1 + X_2 + \dots + X_n$ 이다. i 번째 사람의 모자를 N 명의 사람들에게 모두 동일한 확률로 가져갈 수 있으므로 $P(X_i = 1) = \frac{1}{N}$ 이다. 이로부터 $E[X_i] = \frac{1}{N}$ 임을 알 수 있다. 따라서 $E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = 1$.

문제 3. 15. $X \sim POI(\lambda_1)$, $Y \sim POI(\lambda_2)$ 이다. 그리고 이 두 확률변수는 독립이다. $X + Y$ 가 따르는 분포를 구하여라.

$$P(X + Y = z) = \sum_{x=0}^z \frac{e^{-\lambda_1} \lambda_1^x}{x!} \frac{e^{-\lambda_2} \lambda_2^{z-x}}{(z-x)!} = \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \sum_{x=0}^z {}_z C_x \lambda_1^x \lambda_2^{z-x} = \frac{e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^z}{z!}$$

따라서 $X + Y \sim POI(\lambda_1 + \lambda_2)$ 임을 확인할 수 있다.

문제 3. 16. 가수 신대철은 자신의 기타 연주회에 N 명의 대중을 초청하는데, N 은 공교롭게도 평균이 λ 인 푸아송분포를 따른다. 그런데, 초청된 관객은 p 의 확률로 신대철보다는 그의 아버지인 신중현을 더 좋아하고, $1-p$ 의 확률로 신대철을 더욱 좋아한다. 물론, 각 관객의 선호는 독립적이다. 각각의 수를 N_1, N_2 라고 하자. 즉, $N_1 + N_2 = N$ 이라고 하자. 그렇다면, N_1 과 N_2 역시도 푸아송 분포를 따름을 보이고, 그 기댓값을 구하여라.

$$P(N_1 = n, N_2 = m) = P(N_1 = n, N_2 = m | N = m+n)P(N = m+n)$$

첫째 확률은 이항분포로서, 둘째 확률은 푸아송 분포로서 계산한다.

$$P(N_1 = n, N_2 = m) = \frac{(m+n)!}{m!n!} p^n (1-p)^m \frac{e^{-\lambda} \lambda^{m+n}}{(m+n)!} = \frac{p^n (1-p)^m e^{-\lambda} \lambda^{m+n}}{m!n!}$$

$$P(N_1 = n) = \sum_{m=0}^{\infty} \frac{p^n(1-p)^m e^{-\lambda} \lambda^{m+n}}{m!n!} = e^{-\lambda} \frac{(\lambda p)^n}{n!} \sum_{m=0}^{\infty} \frac{\{\lambda(1-p)\}^m}{m!} = e^{-\lambda p} \frac{(\lambda p)^n}{n!}$$

$$P(N_2 = m) = \sum_{n=0}^{\infty} \frac{p^n(1-p)^m e^{-\lambda} \lambda^{m+n}}{m!n!} = e^{-\lambda} \frac{\{\lambda(1-p)\}^m}{m!} \sum_{n=0}^{\infty} \frac{(\lambda p)^n}{n!} = e^{-\lambda(1-p)} \frac{\{\lambda(1-p)\}^m}{m!}$$

따라서 $N_1 \sim POI(\lambda p)$, $N_2 \sim POI(\lambda(1-p))$, $E[N_1] = \lambda p$, $E[N_2] = \lambda(1-p)$ 임을 확인 가능하다.

문제 3. 17. 문제 3.9.에서 우리는 초기하 분포를 n 개의 작은 확률변수로 나누어서 분석할 수 있음을 배웠다.
그를 가정하고, $i \neq j$ 일 때

$$E[X_i X_j] = \frac{(M)(M-1)}{(N)(N-1)}$$

임을 보여라. 이를 통해

$$Cov(X_i, X_j) = \frac{-M(N-M)}{N^2(N-1)}$$

임을 보여라.

$$E[X_i X_j] = P(X_i = 1, X_j = 1)$$

이 확률은 N 개의 공 중에서 원하는 공 M 개가 있을 때, 두 번 연속으로 원하는 공을 뽑는 확률로 계산할 수 있으므로 $\frac{M}{N} \frac{M-1}{N-1} = \frac{M(M-1)}{N(N-1)}$ 이다.

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = \frac{M(M-1)}{N(N-1)} - \frac{M^2}{N^2} = \frac{-M(N-M)}{N^2(N-1)}$$

문제 3. 18. 이어서, $V(X)$ 가 우리가 아는 것처럼 $\frac{N-n}{N-1} \times n \left(\frac{M}{N} \right) \left(1 - \frac{M}{N} \right)$ 이 됨을 $E[X_i], V(X_i), Cov(X_i, X_j)$ 를 이용하여 보여 보아라.

$$V(X) = V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) + \sum_{i \neq j} Cov(X_i, X_j) = n\left(\frac{M}{N} - \frac{M^2}{N^2}\right) + n(n-1)\frac{-M(N-M)}{N^2(N-1)}$$

$$= n\frac{M}{N}\left(1 - \frac{M}{N}\right) - n(n-1)\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{1}{N-1} = \frac{N-n}{N-1} \times n\frac{M}{N}\left(1 - \frac{M}{N}\right)$$

임을 확인할 수 있다.

문제 3. 19. X 가 이항분포 $B(n, p)$ 를 따른다고 하자. 아래를 보여라.

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

2) k 가 0에서 n 으로 감에 따라, $P(X = k)$ 의 값은 증가하다가 감소한다. 이때, 이 값이 최대가 되는, 즉 증가세에서 감소세로 돌아서는 k 의 값은 $(n+1)p$ 이하의 자연수 중 최대의 자연수이다.

1)

$$P(X = k+1) =_n C_{k+1} p^{k+1} (1-p)^{n-k-1} = \frac{n!}{(k+1)!(n-k-1)!} p^{k+1} (1-p)^{n-k-1}$$

$$= \frac{p}{1-p} \frac{n-k}{k+1} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{p}{1-p} \frac{n-k}{k+1} P(X = k)$$

2) $\frac{p}{1-p} \frac{n-k}{k+1}$ 이 1보다 크거나 작으나에 따라 각각 $P(X = k)$ 가 증가세나 감소세가 결정된다. 즉, $np - pk > k+1 - pk - p$ 인 시점까지는 증가한다. 이를 정리하여 보면 $(n+1)p - 1 > k$ 이다. 따라서, 증가세는 $(n+1)p - 1$ 보다 k 가 작을 때 지속되며, 그것이 멈추는 최초의 점은 $[(n+1)p]$ 이다. 이때, $[x]$ 는 greatest integer function이다.

문제 3. 20. 솔리데어 게임은 52개의 카드로 이루어진 텍을 잘 섞어 만든 카드 더미를 통해 이루어진다. 첫 카드를 뒤집기 전, 에이스라고 말한다. 두번째 카드를 뒤집기 전, 2라고 말한다. 같은 방법으로 계속 하되, 깅까지 모두 말한 이후, 즉 14번째 카드를 뒤집을 때는 다시 에이스부터 시작하여 반복하는 식으로 진행한다. 당신은 당신이 말한 카드와 실제 카드가 일치할 때 패배한다. 푸아송분포를 이용해 52개 카드를 모두 말할 동안 패배하지 않을 확률을 근사하라.

푸아송 분포로 근사하기 위해 i 번째 카드를 뒤집는 사건들이 모두 독립이라고 가정하자. 이 i 개의 사건들 중에서 말한 카드와 실제 카드가 일치하는 가짓수를 X 라고 하면, $X \sim B(52, \frac{1}{13}) \approx POI(4)$ 이다. 즉, $P(X = 0) \approx e^{-4}$.

문제 3. 21. X 는 평균 λ 의 푸아송분포를 따른다. $P(X = i)$ 는 i 가 증가함에 따라 처음엔 증가했다가 그 이후 감소함을 보여라. 또한, 최댓값은 i 가 λ 이하의 자연수 중 가장 큰 것일 때 도달함을 보여라.

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!} = \frac{i+1}{\lambda} e^{-\lambda} \frac{\lambda^{i+1}}{(i+1)!} = \frac{i+1}{\lambda} P(X = i+1), \frac{P(X = i+1)}{P(X = i)} = \frac{\lambda}{i+1}$$

$i > \lambda - 1$ 일 때 $\frac{P(X=i+1)}{P(X=i)} < 1$ 이므로 $P(X = i)$ 는 감소하고, $i < \lambda - 1$ 일 때 $\frac{P(X=i+1)}{P(X=i)} > 1$ 이므로 $P(X = i)$ 는 증가한다. 따라서 $P(X = i)$ 의 최댓값은 $i > \lambda - 1$ 가 되는 최초 시점, 즉 $i = [\lambda]$ 일 때 도달한다.

문제 3. 22. 배구선수 김연경의 리시브효율은 p 이며, 연달아서 서브를 계속 받고 있다. 또한, 각 서브리 시브의 성공확률인 p 는 이전/이후 서브 시도의 영향을 받지 않는다. 확률변수 X 를 성공까지 필요한 시도 횟수라고 하자. 즉, X 가 k 라는 말은 $k - 1$ 번의 리시브 실패 후 k 번째 서브를 잘 받아냈다는 것을 의미한다.

1) X 는 어떤 확률분포를 따르는가? 2) X 의 평균과 분산은 각각 얼마인가?

1)

$$P(X = k) = p(1-p)^{k-1} \Rightarrow X \sim GEO(p)$$

. 즉, X 는 확률이 p 인 기하분포를 따른다.

2)

$$E[X] = \frac{1}{p}, V[X] = \frac{1-p}{p^2}$$

문제 3. 23. 평소 연습량이 많은 김연경 선수는 r 번의 성공을 얻기 전까지 리시브 훈련을 계속한다. 이처럼, 특정 성공 횟수를 얻기 위한 시행의 수를 Y 라고 할 때, Y 는 음이항분포를 따른다고 한다.

1) $k = r, r+1, \dots$ 에 대하여 $P(Y = k)$ 을 구하여라.

(Hint : 기하분포의 무기억성을 잘 기억해 보아라. $r-1$ 번째 사건 발생 이후 r 번째 사건이 발생하기 위해 필요한 시간과, 1번째 사건이 발생하기 위해 필요한 시간은 같은가? 다른가?)

2) $E[Y] = r/p$ 임을 보여라.

1) k 번째 실험에서 r 번의 성공을 얻기 위해서는 $k-1$ 번째 실험까지 $r-1$ 번의 성공이 있어야 한다. $k-1$ 번째 실험까지 $r-1$ 번의 성공이 있을 확률은 $\binom{k-1}{r-1} p^{r-1} (1-p)^{k-r}$ 이고 k 번째 실험에서 성공할 확률은 p 이므로

$$P(Y = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

2)

$$\begin{aligned} E[Y] &= \sum_{k=r}^{\infty} k \binom{k-1}{r-1} p^r (1-p)^{k-r} = \sum_{i=0}^{\infty} (i+r) \frac{(i+r-1)!}{i!(r-1)!} p^r (1-p)^i \\ &= \sum_{i=0}^{\infty} r p^r \frac{(i+r)!}{i!r!} (1-p)^i = r p^r \{1 - (1-p)\}^{-r-1} = \frac{r}{p} \end{aligned}$$

위의 방법 이외에도, 힌트에 나온 것처럼 $i-1$ 번째 성공부터 i 번째 성공까지 필요한 시행의 수가 기하분포임을 이용하여 Y 를 분리해 풀어낼 수도 있다.

문제 3. 24. $X \not\sim B(n, p)$ 를 따른다고 하자.

1) Y 를 $B(n - 1, p)$ 를 따른다는 이항확률변수라고 하자. $E[X^k] = npE[(Y + 1)^{k-1}]$ 임을 보여라.

2) X 의 기댓값과 분산을 구하되, 꼭 1)을 이용하여라.

1)

$$\begin{aligned} E[X^k] &= \sum_{i=0}^n i^k \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} = \sum_{i=1}^n i^k \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \\ &= \sum_{j=0}^{n-1} (j+1)^k \frac{n!}{(j+1)!(n-j-1)!} p^{j+1} (1-p)^{n-j-1} \\ &= np \sum_{j=0}^{n-1} (j+1)^{k-1} \frac{(n-1)!}{j!(n-1-j)!} p^j (1-p)^{n-1-j} = npE[(Y + 1)^{k-1}] \end{aligned}$$

2)

$$E[X] = npE[(Y + 1)^0] = np$$

$$V[X] = E[X^2] - E[X]^2 = npE[Y + 1] - n^2p^2 = np\{(n-1)p + 1\} - n^2p^2 = np(1-p)$$

문제 3. 25. $X \not\sim HYP(G + B, G, n)$ 을 따른다고 하자. 그러면 $G + B$ 가 충분히 커질 때, $B(n, \frac{G}{G+B})$ 로 근사 가능함을 보여라.

$$\begin{aligned} P(X = k) &= \frac{G!}{k!(G-k)!} \frac{B!}{(n-k)!(B-n+k)!} \frac{n!(G+B-n)!}{(G+B)!} \\ &= \frac{n!}{k!(n-k)!} \frac{G!/(G-k)!}{(G+B)!/(G+B-k)!} \frac{B!/(B-n+k)!}{(G+B-k)!/(G+B-n)!} \\ &= \frac{n!}{k!(n-k)!} \frac{G(G-1)...(G-k+1)}{(G+B)(G+B-1)...(G+B-k+1)} \frac{B(B-1)...(B-n+k+1)}{(G+B-k)(G+B-k-1)...(G+B-n+1)} \\ G + B &\text{가 충분히 크면 } 0 \leq i \leq k-1, 0 \leq j \leq n-k-1 \text{에 대해 } \frac{G-i}{G+B-i} \approx \frac{G}{G+B}, \frac{B-j}{G+B-k-j} \approx \frac{B}{G+B} \text{이다.} \\ \text{따라서 } P(X = k) &\approx (\frac{G}{G+B})^k (\frac{B}{G+B})^{n-k}, X \approx B(n, \frac{G}{G+B}) \text{ 임을 확인해줄 수 있다.} \end{aligned}$$

문제 3. 26. 두 확률변수 X, Y 에 대하여 항상 $Y > X$ 이고, 결합확률질량함수가

$$P(X = i, Y = j) = {}_j C_i e^{-2\lambda} \frac{\lambda^j}{j!}$$

이라고 정의하자.

1) Y 의 주변확률질량함수를 구하여라.

2) $E[X]$ 의 값을 구하여라.

3) $Z = Y - X$ 의 확률질량함수를 구하여라.

1)

$$P(Y = j) = \sum_{i=0}^{j-1} {}_j C_i e^{-2\lambda} \frac{\lambda^j}{j!} = e^{-2\lambda} \frac{\lambda^j}{j!} \sum_{i=0}^{j-1} {}_j C_i = e^{-2\lambda} (2^j - 1) \frac{\lambda^j}{j!}$$

2)

$$P(X = i) = \sum_{j=i+1}^{\infty} {}_j C_i e^{-2\lambda} \frac{\lambda^j}{j!} = \frac{e^{-2\lambda}}{i!} \sum_{j=i+1}^{\infty} \frac{\lambda^j}{(j-i)!} = \frac{e^{-2\lambda}}{i!} \lambda^i \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} = \frac{e^{-2\lambda}}{i!} \lambda^i (e^{\lambda} - 1) = \frac{e^{-\lambda}}{i!} \lambda^i (1 - e^{-\lambda})$$

$$E[X] = \lambda(1 - e^{-\lambda})$$

3)

$$P(Z = k) = \sum_{i=0}^{\infty} P(X = i, Y = i+k) = \sum_{i=0}^{\infty} {}_{i+k} C_i e^{-2\lambda} \frac{\lambda^{i+k}}{(i+k)!} = e^{-2\lambda} \frac{\lambda^k}{k!} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \frac{\lambda^k}{k!}$$

문제 3. 27. 호수에 총 b 마리의 물고기가 있다. 처음에 그물로 a 마리를 잡은 후에, 표식을 단 후 다시 호수에 방생하였다. 그 다음에 충분한 시간이 지나고 표식을 단 물고기를 m 마리 잡기 위해 다시 잡아야 할 물고기의 총 개체수를 X 라 하자. (단, $a << m \leq X << b$)

1) X 의 확률질량함수를 구하여라.

2) $E[X]$ 가 $m \frac{b+1}{a+1}$ 임을 보여라.

3) $V(X)$ 을 구하여라.

1)

$$\begin{aligned} P(X = n) &= \frac{\binom{a}{m-1} \binom{b-a}{n-m}}{\binom{b}{n-1}} \times \frac{a-m+1}{b-n+1} \\ &= \frac{\frac{a!}{(m-1)!(a-m+1)!} \times \frac{(b-a)!}{(n-m)!(b-a-n+m)!}}{\frac{b!}{(n-1)!(b-n+1)!}} \times \frac{a-m+1}{b-n+1} \\ &= \frac{\binom{n-1}{m-1} \binom{b-n}{a-m}}{\binom{b}{a}}, \quad n = m, \dots, b-a+m \end{aligned}$$

2)

Sol) (a)에서 $\frac{1}{\binom{b}{a}} \sum_{n=m}^{b-a+m} \binom{n-1}{m-1} \binom{b-n}{a-m} = 1$ 이다. 이를 $(b+1)$ 마리가 있는 호

수에서 $(a+1)$ 마리가 표식이 있고 표식이 있는 $(m+1)$ 마리 물고기를 잡는 경우

에 대응시켜 생각하면 $\frac{1}{\binom{b+1}{a+1}} \sum_{n=m+1}^{b-a+m+1} \binom{n-1}{m} \binom{b-n+1}{a-m} = 1$

$\Leftrightarrow \frac{1}{\binom{b+1}{a+1}} \sum_{n=m}^{b-a+m} \binom{n}{m} \binom{b-n}{a-m} = 1$ 이다. 따라서

$$\begin{aligned} E(X) &= \frac{m}{\binom{b}{a}} \sum_{n=m}^{b-a+m} \frac{n}{m} \binom{n-1}{m-1} \binom{b-n}{a-m} = \frac{m \times \binom{b+1}{a+1}}{\binom{b}{a}} \frac{1}{\binom{b+1}{a+1}} \sum_{n=m}^{b-a+m} \binom{n}{m} \binom{b-n}{a-m} \\ &= m \left(\frac{b+1}{a+1} \right) \end{aligned}$$

3)

$$\begin{aligned}
& \frac{1}{\binom{b+2}{a+2}} \sum_{n=m+2}^{b-a+m+2} \binom{n-1}{m+1} \binom{b-n+2}{a-m} = 1 \Leftrightarrow \frac{1}{\binom{b+2}{a+2}} \sum_{n=m}^{b-a+m} \binom{n+1}{m+1} \binom{b-n}{a-m} = 1 \\
& \Rightarrow E(X(X+1)) = \frac{1}{\binom{b}{a}} \sum_{n=m}^{b-a+m} n(n+1) \binom{n-1}{m-1} \binom{b-n}{a-m} \\
& = m(m+1) \frac{\binom{b+2}{a+2}}{\binom{b}{a}} \frac{1}{\binom{b+2}{a+2}} \sum_{n=m}^{b-a+m} \frac{n(n+1)}{m(m+1)} \binom{n-1}{m-1} \binom{b-n}{a-m} \\
& = m(m+1) \frac{(b+2)(b+1)}{(a+2)(a+1)} \frac{1}{\binom{b+2}{a+2}} \sum_{n=m}^{b-a+m} \binom{n+1}{m+1} \binom{b-n}{a-m} \\
& = m(m+1) \frac{(b+2)(b+1)}{(a+2)(a+1)} \\
E(X(X-1)) &= E(X(X+1)) - 2E(X) = m\left(\frac{b+1}{a+1}\right)[(m+1)\frac{b+2}{a+2} - 2]
\end{aligned}$$

이므로,

$$\begin{aligned}
V(X) &= E(X(X-1)) + E(X) - E(X)^2 \\
&= m\left(\frac{b+1}{a+1}\right) \left[(m+1)\frac{b+2}{a+2} - 2 \right] + m\left(\frac{b+1}{a+1}\right) - m^2 \left(\frac{b+1}{a+1}\right)^2 \\
&= m\left(\frac{b+1}{a+1}\right) \left[(m+1)\frac{b+2}{a+2} - m\frac{b+1}{a+1} - 1 \right]
\end{aligned}$$

임을 알 수 있다.

문제 3. 28. 멘델이 유전법칙을 발견하기 위해 완두콩을 이용하였는데, 완두콩에는 열성 형질인 노란색과 우성 형질인 초록색이 있다. 멘델은 자신의 법칙을 증명하기 위해 우성 중 순종의 비율이 $1/3$ 이 됨을 보이고자 했다. 멘델은 우성 표현형을 가진 2세대 완두콩 600개를 임의로 뽑은 후 자가수분을 통해 2세대 완두콩 각각에게서 10개의 3세대 완두콩을 만들고 10개 모두 우성인 경우 2세대 부모 완두콩이 순종이라 간주하였다. 실제로 600개 중 200개가 순종이었다고 하자.

1) X 를 우성으로 판단되는 2세대 부모 완두콩의 수라고 하자. X 의 확률질량함수를 구하여라.2) $P(X \leq 201)$ 을 구하여라.

1) 우선, 실제로 600개 중 200개가 순종이므로 $P(X = i)$ 는 $i \geq 200$ 일 때만 의미있다. 또한, 나머지 400 개의 잡종 완두콩 중에서 순종으로 판단되는 완두콩의 개수를 Y 라고 하면 $Y \sim B(400, (\frac{3}{4})^{10})$ 이다. 즉,

$$P(X = i) = P(Y = i - 200) = {}_{400}C_{i-200} \left(\frac{3}{4}\right)^{10(i-200)} \left(1 - \left(\frac{3}{4}\right)^{10}\right)^{600-i}$$

2)

$$P(X \leq 201) = P(X = 200) + P(X = 201) = \left(1 - \left(\frac{3}{4}\right)^{10}\right)^{400} + 400 \left(\frac{3}{4}\right)^{10} \left(1 - \left(\frac{3}{4}\right)^{10}\right)^{399}$$

문제 3. 29. 동전을 던져 앞면이면 H , 뒷면이면 T 라고 표시해 n 번의 시행을 통해 문자열 w 를 얻자. 예를 들어, $w = HHTTT$ 이다. 그리고, X_n 을 해당하는 문자열에서 나온 앞면의 개수라고 하자.

1) X_5 의 확률질량함수를 구하여라.

2) X_n 의 확률질량함수를 구하여라. $P(X_n = k) = p_n(k)$ 라고 정의할 때, $p_n(k)$ 의 n 에 대한 재귀식을 구하여라.

1) X_5 는 $B(5, \frac{1}{2})$ 를 따르는 확률변수이다. 따라서,

$$P(X_5 = i) = \binom{5}{i} \left(\frac{1}{2}\right)^5 \quad i = 0, 1, 2, 3, 4, 5$$

이다.

2) X_n 은 단순히 동전을 n 번 던져서 나온 앞면의 개수로 생각할 수 있으므로 $X_n \sim B(n, \frac{1}{2})$ 이다. 즉, $P(X_n = k) = {}_n C_k (\frac{1}{2})^n = {}_{n-1} C_k (\frac{1}{2})^{n-1} \frac{n}{2(n-k)}$ 이고 이로부터 $p_n(k) = \frac{n}{2(n-k)} p_{n-1}(k)$ 임을 알 수 있다.

문제 3. 30. X 와 Y 가 독립이며, 동일한 성공 확률 p 를 가진 기하분포를 따른다. $P(X = k | X + Y = n)$ 은 무엇인가?

$$P(X + Y = n) = \sum_{i=1}^{n-1} P(X = i)P(Y = n - i) = \sum_{i=1}^{n-1} p(1-p)^{i-1}p(1-p)^{n-i-1} = (n-1)p^2(1-p)^{n-2}$$

$$P(X = k | X + Y = n) = \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} = \frac{p(1-p)^{k-1}p(1-p)^{n-k-1}}{(n-1)p^2(1-p)^{n-2}} = \frac{1}{n-1}$$

문제 3. 31. 확률변수 X, Y 가 유한한 기댓값을 가지고 있다.

$$E[\max(X, Y)] = E[X] + E[Y] - E[\min(X, Y)]$$

임을 보여라.

$$E[\min(X, Y)] + E[\max(X, Y)] = E[\min(X, Y) + \max(X, Y)] = E[X + Y] = E[X] + E[Y]$$

$$\Rightarrow E[\max(X, Y)] = E[X] + E[Y] - E[\min(X, Y)]$$

문제 3. 32. 다음을 보여라.

1)

$$\sum_{j=0}^k {}_{a+k-j-1} C_{k-j} {}_{b+j-1} C_j = {}_{a+b+k-1} C_k$$

2)

$$\sum_{r=0}^{m-1} {}_{n+m-1} C_{n+r} p^{n+r} (1-p)^{m-1-r} = \sum_{k=n}^{n+m-1} {}_{k-1} C_{n-1} p^n (1-p)^{k-n} = \sum_{j=0}^{m-1} {}_{n-1+j} C_{n-1} p^n (1-p)^j$$

1) $a+b+k$ 번의 시도를 한 순간 성공 확률이 p 인 것을 $a+b$ 번 성공할 확률은 ${}_{a+b+k-1} C_{a+b-1} p^{a+b-1} (1-p)^k \times p$ 로, 이를 잘 정리하면

$$p^{a+b} (1-p)^k {}_{a+b+k-1} C_k$$

이다. 그러면 a 번 성공하는 시점은 a 번 시도했을 때부터 $a+k$ 번 시도했을 때이며, $a+k-j$ 번 시도를 한 순간 a 번 성공했을 확률은

$${}_{a+k-j-1} C_{a-1} p^{a-1} (1-p)^{k-j} \times p$$

이며 그 후로 남은 $b+j$ 번 시도 동안 b 번 성공했을 확률은

$${}_{b+j-1} C_{b-1} p^{b-1} (1-p)^j \times p$$

이다. 따라서

$${}_{a+b+k-1}C_k p^{a+b} (1-p)^k = \sum_{j=0}^k {}_{a+k-j-1}C_{k-j} p^a (1-p)^{k-j} {}_{b+j-1}C_j p^b (1-p)^j$$

이므로, 양변을 잘 정리해주면 원래의 결과를 얻을 수 있다.

2) 첫째 식은 $n+m-1$ 번의 독립 시행에서 n 번 이상 성공할 확률과 비례하며, 둘째 식은 n 번 성공하는 시점이 n 번 시도했을 때부터 $n+m-1$ 번 시도했을 때의 사이일 확률과 비례한다. 둘은 같음을 쉽게 확인할 수 있다. 세 번째 식은 두 번째 식에서 $j = k-n$ 으로 두면 쉽게 보일 수 있다.

문제 3. 33. X 와 Y 가 독립인 음이항분포를 따르는 확률변수라고 하자. 이들의 성공 확률은 p 로 동일하며, 원하는 성공 횟수가 각각 a, b 이다. 조건부 확률 $P(Y = b+j | X+Y = a+b+k)$ 을 구하여라. 앞선 문제의 결과를 적극 활용하여라.

앞선 문제에서 보인 결과를 이용하자.

$$\begin{aligned} P(Y = b+j | X+Y = a+b+k) &= \frac{P(Y = b+j, X = a+k-j)}{P(X+Y = a+b+k)} \\ &= \frac{P(X = a+k-j)P(Y = b+j)}{P(X+Y = a+b+k)} \\ &= \frac{{}_{a+k-j-1}C_{k-j} {}_{b+j-1}C_j}{{}_{a+b+k-1}C_k} \end{aligned}$$

이다.

문제 3. 34. 아래를 보여라.

$$\begin{aligned} 1) {}_{n+m}C_r &= \sum_{k=0}^r {}_nC_k {}_mC_{r-k} \\ 2) {}_{2n}C_n &= \sum_{k=0}^n ({}_nC_k)^2 \end{aligned}$$

1) $A = \{a_1, \dots, a_n\}, B = \{b_1, \dots, b_m\}$ 이라고 하자 ($A \cap B = \emptyset$). $|M| = r$ 인 $A \cup B$ 의 임의의 부분집합 M 에 대해 항상 0 이상 r 이하인 어떤 k 가 존재하여 $|A \cap M| = k, |B \cap M| = r-k$ 이다. 반대로, 어떤 0 이상 r 이하인 k 가 존재하여 $|A \cap M| = k, |B \cap M| = r-k$ 이면, $|(A \cap M) \cup (B \cap M)| = r$ 이고 $(A \cap M) \cup (B \cap M) \subset A \cup B$ 이다. 다시 말해, 크기가 r 인 $A \cup B$ 의 부분집합들과 $|A \cap M| = k, |B \cap M| = r-k$ 인 집합들 사이에는 일대일 대응이 존재한다. 따라서 ${}_{n+m}C_r = \sum_{k=0}^r {}_nC_k {}_mC_{r-k}$ 이 성립한다.

$$2) {}_{2n}C_n = {}_{n+n}C_n = \sum_{k=0}^n {}_nC_k {}_{n-k}C_{n-k} = \sum_{k=0}^n ({}_nC_k)^2$$

문제 3. 35. 본페로니 부등식을 보여라.

$$P(E_1 \cap E_2 \cap \dots \cap E_n) \geq \sum_{i=1}^n P(E_i) - (n-1)$$

Note : 사실 이 문제는 여기에 오기전에 적절한 문제는 아니지만, 아이디어가 어느 정도 비슷하여 넣어 보았다.

수학적 귀납법을 사용하자. $n = 2$ 일 때 $P(E_1 \cap E_2) = P(E_1) + P(E_2) - P(E_1 \cup E_2) \geq P(E_1) + P(E_2) - 1$. $n = k$ 일 때 본페로니 부등식이 성립한다고 가정하자. 그러면, $P(E_1 \cap \dots \cap E_k \cap E_{k+1}) \geq P(E_1 \cap \dots \cap E_k) + P(E_{k+1}) - 1 \geq \sum_{i=1}^k P(E_i) - (k-1) + P(E_{k+1}) - 1 = \sum_{i=1}^{k+1} P(E_i) - k$. 증명이 완료되었다.

문제 3. 36. 골초로 알려진 가수 A 는 항상 두 개의 라이터를 들고 다닌다. 하나는 왼쪽 주머니에, 한쪽에는 오른쪽 주머니에 넣어 다닌다. 그가 연초에 불을 붙이려 할 때, 그는 두 주머니 중 하나를 골라 라이터를 사용한다. 각 라이터는 N 번 사용 가능하다고 하자. 그 가수가 자신의 한 라이터가 다 된 것을 발견한 순간, 나머지 라이터가 X 번 불을 붙일 수 있을 것이라 하자. $P(X = k) = p_k$ 라 하자. 단, 다 된 것을 발견하는 것은 라이터를 $N+1$ 번 쓰려는 순간이라고 생각하자. 즉 $X = 0$ 일 가능성성이 있다.

- 1) $\sum_{k=0}^N p_k = 1$ 임을 보여라.
 2) X 의 기댓값을 구하여라.

1)

$$\begin{aligned} p_N &= 2 \times \frac{1}{2^{N+1}} = \frac{1}{2^N} \\ p_{N-1} &= 2 \times \binom{N+1}{1} \frac{1}{2^{N+2}} = \binom{N+1}{1} \times \frac{1}{2^{N+1}} \\ p_{N-2} &= 2 \times \binom{N+2}{2} \frac{1}{2^{N+3}} = \binom{N+2}{2} \times \frac{1}{2^{N+2}} \\ &\dots \\ p_0 &= 2 \times \binom{2N}{N} \frac{1}{2^{2N+1}} = \binom{2N}{N} \frac{1}{2^{2N}} \end{aligned}$$

임을 확인할 수 있다.

$$\begin{aligned} \sum_{k=0}^N p_k &= \sum_{k=0}^N \binom{2N-k}{N-k} \frac{1}{2^{2N-k}} \\ &= \sum_{k=0}^N \binom{N+k}{k} \frac{1}{2^{N+k}} \\ &= \frac{1}{2^N} \sum_{k=0}^N \binom{N+k}{N} \frac{1}{2^k} \\ &= 2 \sum_{k=1}^N (\text{성공할 확률이 실패할 확률과 같은 사건을 } N+k+1 \text{번의 시행을 통해 } N+1 \text{번의 성공을 해낼 확률}) \\ &= 2(N+1 \sim 2N+1 \text{번의 시행을 통해 } N+1 \text{번 성공할 확률}) \\ &= 2(2N+1 \text{번의 시행 동안 } N+1 \text{번 이상 성공할 확률}) \\ &= (2N+1 \text{번의 시행 동안 } N+1 \text{번 이상 성공할 확률}) + (2N+1 \text{번의 시행 동안 } N \text{번 이하 성공할 확률}) = 1 \end{aligned}$$

2)

$$\begin{aligned}
 E(X) &= \sum_{k=0}^N kp_k \\
 &= \sum_{k=0}^N (N-k) \binom{N+k}{k} \frac{1}{2^{N+k}} \\
 &= N - \sum_{k=1}^N k \binom{N+k}{k} \frac{1}{2^{N+k}} \\
 &= N - \sum_{k=1}^N (N+k) \binom{N+k-1}{k-1} \frac{1}{2^{N+k}} \\
 &= N - \sum_{k=0}^{N-1} (N+k+1) \binom{N+k}{k} \frac{1}{2^{N+k+1}} \\
 &= N - (N+1) \sum_{k=0}^{N-1} \binom{N+k}{k} \frac{1}{2^{N+k+1}} - \sum_{k=0}^{N-1} k \binom{N+k}{k} \frac{1}{2^{N+k+1}} \\
 &= N - \left(\frac{N+1}{2} - \binom{(N+1)2N}{N} \frac{1}{2^{2N+1}} \right) - \frac{1}{2} \left(N - E(X) - N \binom{2N}{N} \frac{1}{2^{2N}} \right) \\
 &= \frac{1}{2} E(X) + (2N+1) \binom{2N}{N} \frac{1}{2^{2N+1}} - \frac{1}{2}
 \end{aligned}$$

이다. 이를 기댓값에 대해 정리하면

$$E(X) = (2N+1) \binom{2N}{N} \frac{1}{2^{2N}} - 1$$

문제 3. 37. 성공률이 p 인 베르누이 독립시행을 한없이 행할 때, X_n 을 n 번째 시행에서 성공이면 1, 실패면 0인 확률변수라고 하자.

1) $X = X_1 + X_2 + \dots$ 일 때 $E[X] = \infty$ 임을 보여라.

2) $Y_n = X_n X_{n+1}$ 이라고 정의하자. $Y = Y_1 + Y_2 + \dots$ 라고 할 때, $E[Y] = \infty$ 임을 보여라.

1)

$$\begin{aligned}
 E(X) &= \lim_{s \rightarrow \infty} E(X_1 + X_2 + \dots + X_s) \\
 &= \lim_{s \rightarrow \infty} E(X_1) + \dots + E(X_s) \\
 &= \lim_{s \rightarrow \infty} sp = \infty
 \end{aligned}$$

2) Y_i 의 기댓값을 구하여 보자. $E[Y_i] = E[X_i X_{i+1}] = E[X_i]E[X_{i+1}] = p^2$ 이다. 이때, 중간의 식은 X_i 와 X_{i+1} 이 독립이기에 성립한다.

$$\begin{aligned}
 E(Y) &= \lim_{s \rightarrow \infty} E(Y_1 + Y_2 + \dots + Y_s) \\
 &= \lim_{s \rightarrow \infty} E(Y_1) + E(Y_2) + \dots + E(Y_s) \\
 &= \lim_{s \rightarrow \infty} sp^2 = \infty
 \end{aligned}$$

문제 3. 38. 위의 문제에서, $2n-1$ 번째 시행에서 처음으로 성공의 누적횟수가 실패의 누적횟수보다 커지는 사건을 C_n 이라고 하자. Z 는 $P(Z=x) = P(C_x)$ 인 이산확률변수라고 하자. Z 의 확률밀도함수를 구하여라.

$P(Z=x) = P(C_x)$ 라고 하였다. 그러므로, $x = 1, 2, \dots$ 임을 알 수 있다. 먼저, 1일 때를 생각하여 보자. 그러면 1번째 시행에서 처음으로 성공의 횟수가 실패의 누적횟수보다 커져야 하므로, 처음 시행에서 원하는

것이 나올 확률인 p 다. 따라서, $P(Z = 1) = p$.

둘째로, $x = 2$ 일 때를 생각해 보자. 그러면 3번째 시행에서 처음으로 성공 횟수가 실패 횟수보다 커져야 하므로, 성공이 2회, 실패가 1회이며 실패 - 성공 - 성공의 순서로 나왔어야 한다. $P(Z = 2) = (1 - p)p^2$.

셋째로, $x = 3$ 일 때를 생각해 보자. 5번째 시행에서 처음으로 성공 횟수가 실패 횟수보다 커져야 한다. 따라서, 5번째 시행은 성공이며, 4번째 시행에서는 실패 - 실패 - 성공 - 성공 혹은 실패 - 성공 - 실패 - 성공 순으로 나왔어야 한다. 즉, $P(Z = 3) = 2(1 - p)^2 p^3$ 이다.

이때, $x = n + 1$ 이라면 $2n + 1$ 번째 시행에서는 성공을 이루어야 하고, $2n$ 번째 시행까지는 성공과 실패의 수가 같되 실패가 항상 성공보다 많아야 한다. 그 가짓수는 카탈란수로, ${}_{2n}C_n - {}_{2n}C_{n-1}$ 임을 알고 있다. 따라서, $x \in \mathbb{N} - \{1\}$ 에 대해

$$P(Z = x) = ({}_{2x-2}C_{x-1} - {}_{2x-2}C_{x-2})(1 - p)^{x-1} p^x$$

이며, $P(Z = 1) = p$ 이다.

문제 4. 1. 연속확률변수 X 의 확률밀도함수가

$$f(x) = ae^{-x}, \quad 0 \leq x$$

로 정의될 때, a 의 값을 구하여라.

$$\int_0^\infty ae^{-x} dx = [-ae^{-x}]_{x=0}^{x=\infty} = a$$

이다. 따라서, $a = 1$ 이어야 한다.

문제 4. 2. 연속확률변수 X 의 누적분포함수가

$$F(x) = \frac{x}{1+x}, \quad 0 \leq x$$

로 주어질 때, $f(x)$ 를 구하여라.

$f(x)$ 는 $F(x)$ 를 미분함으로써 얻어낼 수 있다. 따라서,

$$f(x) = \frac{1}{(1+x)^2} \quad 0 \leq x$$

문제 4. 3. 연속확률변수 X 에 대하여, $E[X]$ 가 존재한다고 가정하자.

$$E[X] = \int_0^\infty P(X > x)dx - \int_0^\infty P(X < -x)dx$$

임을 보여라. 단, $E(X)$ 가 존재하면 $\lim_{t \rightarrow \infty} tF(-t) = 0$ 임은 증명 없이 사용해도 괜찮다.

$$\begin{aligned} E(X) &= \int_{-\infty}^\infty xf(x)dx \\ &= \lim_{t \rightarrow \infty} \lim_{s \rightarrow \infty} [xF(x)]_0^s - \int_0^s F(x)dx + [xF(x)]_{-t}^0 + \int_{-t}^0 F(x)dx \\ &= \lim_{t \rightarrow \infty} \lim_{s \rightarrow \infty} sF(s) - \int_0^s P(X > x)dx - s + tF(-t) + \int_{-t}^0 P(X < x)dx \\ &= \lim_{t \rightarrow \infty} \lim_{s \rightarrow \infty} s(F(s) - 1) + tF(-t) + \int_0^s P(X > x)dx + \int_0^t P(X < -x)dx \end{aligned}$$

그런데, $E(X)$ 가 존재한다면 $\lim_{t \rightarrow \infty} tF(-t) = 0$ 임을 알고 있다. 이를 잘 응용하면 $\lim_{s \rightarrow \infty} s(F(s) - 1) = 0$ 역시도 동치임을 확인할 수 있다. 따라서, 극한을 이상적분 형태로 다시 바꿔주면

$$E[X] = \int_0^\infty P(X > x)dx - \int_0^\infty P(X < -x)dx$$

문제 4. 4. X 가 연속확률변수라고 하며, $a \leq X \leq b$ 이고 $E[X] = \mu$ 라고 하자.

(1) $a \leq \mu \leq b$ 임을 보여라.

(2) $V(X) \leq \frac{1}{4}(b-a)^2$ 임을 보여라.

(1)

$$E(X) = \int_a^b xf(x)dx$$

이다. 이때, $f(x)$ 는 X 의 확률밀도함수이다. 그런데, $a \leq x \leq b$ 므로

$$a \int_a^b f(x)dx \leq E(X) \leq b \int_a^b f(x)dx$$

이면, $f(x)$ 는 확률밀도함수이므로 적분값이 1이다. 따라서,

$$a \leq \mu \leq b$$

임을 알 수 있다.

(2)

$$\int_a^b (x-a)(b-x)f(x)dx$$

는 피적분함수가 항상 양수이므로 양수이다. 이를 잘 풀어내면,

$$\int_a^b (x-a)(b-x)f(x)dx = - \int_a^b x^2 f(x)dx + (a+b) \int_a^b xf(x)dx - ab \int_a^b f(x)dx = -E(X^2) + (a+b)\mu - ab$$

이다. 그런데

$$\begin{aligned} V(X) &= E(X^2) - E(X)^2 \\ &= - \int_a^b (x-a)(b-x)f(x)dx - \mu^2 + (a+b)\mu - ab \\ &\leq -\mu^2 + (a+b)\mu - ab \\ &= -(\mu - \frac{a+b}{2})^2 + \frac{1}{4}(b-a)^2 \end{aligned}$$

임을 확인할 수 있다.

문제 4. 5. 연속확률변수 X 에 대해서도, 마코프 부등식과 채비셰프 부등식이 성립함을 보여라.

마코프 부등식이 성립함을 먼저 보이자. $f(x)$ 는 확률밀도함수, $F(x)$ 는 누적분포함수이다.

$$\begin{aligned} P(X > a) &= \int_a^\infty f(x)dx \\ &\leq \frac{\int_a^\infty xf(x)dx}{a} \\ &\leq \frac{E(X)}{a} \end{aligned}$$

그 다음으로, 채비셰프 부등식이 성립함은 마코프 부등식이 성립함에 따라 자연스럽게 유도될 수 있다. 이산확률분포에 대해서 했던 것에서 증명 상의 차이점이 존재하지 않는다.

문제 4. 6. 소방서가 수직선 위에 있다. 불의 위치가 $f(x)$ 라는 확률밀도함수를 따라 분포해 있으며, $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$ 를 만족한다고 하자. 그렇다면, 불로부터 소방서 사이의 거리의 기댓값을 최소화할 수 있는 소방서의 위치는 어디인가?

소방서의 거리가 a 라고 하자. 그러면 소방서의 위치는

$$\int_{-\infty}^{\infty} |x-a|f(x)dx$$

의 값을 최소화시키는 a 이다. $|x-a| \leq |x| + |a|$ 므로,

$$\int_{-\infty}^{\infty} |x-a|f(x)dx \leq \int_{-\infty}^{\infty} |x|f(x)dx + \int_{-\infty}^{\infty} |a|f(x)dx = \int_{-\infty}^{\infty} |x|f(x)dx + |a| < \infty$$

임 역시 확인해줄 수 있다. 그런 a 의 자리를 찾아보자. a 로 식을 미분하게 된다면,

$$\begin{aligned} \frac{d}{da} \int_{-\infty}^{\infty} |x-a| f(x) dx &= \frac{d}{da} \int_{-\infty}^a (a-x) f(x) dx + \frac{d}{da} \int_a^{\infty} (x-a) f(x) dx \\ &= \int_{-\infty}^a f(x) dx - af(a) + af(a) - af(a) + af(a) - \int_a^{\infty} f(x) dx \\ &= 2F(a) - 1 \end{aligned}$$

이다. 따라서, $F(a) = 0.5$ 일 때 이것이 최소가 됨을 알 수 있다. 즉, 소방서는 $F(x) = 0.5$ 가 되는 점에 설치해야 한다.

문제 4. 7. 연속확률변수에 대해서도 적률생성함수를 만들 수 있다. $\phi(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ 라고 정의할 때,

$$\phi'(0) = E[X]$$

$$\phi''(0) = E[X^2]$$

임을 보여라.

$$\phi'(t) = \int_{-\infty}^{\infty} xe^{tx} f(x) dx$$

이므로 $\phi'(0) = \int_{-\infty}^{\infty} xf(x) dx = E(X)$ 이고,

$$\phi''(t) = \int_{-\infty}^{\infty} x^2 e^{tx} f(x) dx$$

이므로 $\phi''(0) = \int_{-\infty}^{\infty} x^2 f(x) dx = E(X^2)$ 이다.

문제 4. 8. 연속확률변수 X 가 확률밀도함수

$$f(x) = \begin{cases} cx^3, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

를 가진다고 하자. c 의 값을 구하여라.

$$\int_0^1 cx^3 dx = \frac{1}{4}c = 1$$

이어야 하므로, $c = 4$ 이다.

문제 4. 9. X_1, X_2, \dots, X_n 은 독립인 확률변수이며, 그들의 확률밀도함수는 모두

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

라고 하자. $M = \max(X_1, X_2, \dots, X_n)$ 이라고 정의할 때, M 의 확률밀도함수를 구하여라.

$M < 0$ 인 경우나, $M > 1$ 인 사건은 일어나지 않는다. 따라서 M 의 확률밀도함수 $g(x)$ 는 $0 \leq x \leq 1$ 에서만 정의된다.

$$P(M \leq x) = P(\max(X_1, X_2, \dots, X_n) \leq x) = P(X_1 < x)P(X_2 < x) \cdots P(X_n < x) = x^n$$

임을 확인할 수 있으니, $g(x) = nx^{n-1}$ 이다. 따라서,

$$g(x) = \begin{cases} 0 & x < 0 \text{ or } x > 1 \\ nx^{n-1} & 0 \leq x \leq 1 \end{cases}$$

이다.

문제 4. 10. 확률변수 X 가 확률밀도함수

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

를 따른다고 하자. 확률변수 $Y = X^3$ 의 확률밀도함수를 구하여라.

Y 의 확률밀도함수를 $g(x)$, 누적분포함수를 $G(x)$ 라고 하자. X 는 0과 1 사이에 존재하므로, Y 도 0과 1 사이에 존재한다. 따라서 $x < 0$ 이거나 $x > 1$ 일 경우 $g(x) = 0$ 이다. $0 \leq x \leq 1$ 에서 $g(x)$ 를 구하여 보자.

$$\begin{aligned} G(x) &= P(Y \leq x) \\ &= P(X^3 \leq x) \\ &= P(X \leq \sqrt[3]{x}) \\ &= F(\sqrt[3]{x}) = \frac{1}{2}x^{\frac{2}{3}} \end{aligned}$$

이다. 따라서, 양변을 x 로 미분하면

$$g(x) = \begin{cases} 0 & x < 0 \text{ or } x > 1 \\ \frac{1}{3}x^{-\frac{1}{3}} & 0 \leq x \leq 1 \end{cases}$$

이다.

문제 4. 11. X 의 확률밀도함수가

$$f(x) = \begin{cases} a + bx^2, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

로 주어진다. $E[X] = 0.6$ 일 때, a, b 를 구하여라.

$$\begin{aligned} \int_0^1 f(x)dx &= a + \frac{1}{3}b = 1 \\ \int_0^1 xf(x)dx &= \frac{1}{2}a + \frac{1}{4}b = \frac{3}{5} \end{aligned}$$

이므로, $a = 3/5, b = 6/5$ 이다.

문제 4. 12. 한 음반이 발표되어 차트아웃 될 때까지의 시간은 아래와 같은 확률밀도함수 $f(x)$ 를 가지는 확률변수 X 로 표현된다.

$$f(x) = a^2 xe^{-ax}, \quad x \geq 0$$

$E[X]$ 를 a 로 표현하고, a 가 2일 때의 값을 구하라.

먼저, a 의 값을 구하여 보자.

$$\int_0^\infty a^2 x e^{-ax} dx = [-axe^{-ax}]_0^\infty + \int_0^\infty ae^{-ax} dx = [-e^{-ax}]_0^\infty = 1$$

따라서, a 가 양수임만 확인할 수 있을 뿐 그 값을 정확하게 알 수는 없다.

$$\begin{aligned} E(X) &= \int_0^\infty a^2 x^2 e^{-ax} dx \\ &= [-ax^2 e^{-ax}]_0^\infty + \int_0^\infty 2axe^{-ax} dx \\ &= \frac{2}{a} \int_0^\infty a^2 x e^{-ax} dx \\ &= \frac{2}{a} \end{aligned}$$

이다. $a = 2$ 면 그 값은 1이다.

문제 4. 13. X 가 연속확률변수이고 누적분포함수 F 를 가질 때, 그 중간값은

$$F(m) = 0.5$$

인 m 으로 정의된다.

$f(x) = e^{-x}, x \geq 0$ 의 중간값을 구하여라.

$$F(m) = \int_0^m e^{-x} dx = [-e^{-x}]_0^m = 1 - e^{-m}$$

이 0.5여야 한다. 즉, $e^m = 2$ 여야 하기에 중간값은 $m = \ln 2$ 다.

문제 4. 14. 확률변수 X, Y 가 누적분포함수 F_X 와 F_Y 를 가지고 $a, b > 0$ 일 때,

$$F_X(x) = F_Y\left(\frac{x-a}{b}\right)$$

이다.

- 1) $E[X]$ 를 $E[Y]$ 로서 표시하라.
- 2) $V(X)$ 를 $V(Y)$ 로 표현하라.

1)

$$P(X < x) = F_X(x) = F_Y\left(\frac{x-a}{b}\right) = P(Y < \frac{x-a}{b})$$

이다. 따라서, $Y = \frac{X-a}{b}$ 라면

$$P(X < x) = P\left(\frac{X-a}{b} < \frac{x-a}{b}\right) = P(Y < \frac{x-a}{b})$$

임을 확인할 수 있게 된다. 따라서, $X = a + bY$ 로 표현된다.

$$E(X) = a + bE(Y)$$

2)

$$V(X) = b^2 V(Y)$$

임을 알 수 있다.

문제 4. 15. X_1, X_2, \dots 는 독립이며 동일한 연속확률분포를 가지는 확률변수의 수열이다. $N \geq 2$ 를

$$X_1 \geq X_2 \geq \dots \geq X_{N-1} < X_N$$

인 N 이라 정의하자. 즉, 감소를 멈추는 첫 점이라고 생각하자.

- 1) $P(N \geq n)$ 의 값을 구하여라.
- 2) $E[N] = e$ 임을 보여라.

1) 동일한 확률변수를 가지므로, 어느 확률변수가 다른 확률변수보다 작을 확률과 클 확률은 0.5로 같다. 즉, 문제를 풀 때 연속확률분포가 무엇인지 알 필요가 없다. $P(N \geq n)$ 은 $X_1, X_2, X_3, \dots, X_{N-1}$ 의 계속 감소해야 하므로, $N - 1$ 개의 숫자를 늘어놓았을 때 그것이 내림차순으로 배열될 확률이다. 즉,

$$P(N \geq n) = \frac{1}{(n-1)!}$$

임을 알 수 있다. 물론, n 은 2 이상의 자연수일 것이다.

2)

$$\begin{aligned} E[N] &= \sum_{k=2}^{\infty} k P(N = k) \\ &= \sum_{k=2}^{\infty} k (P(N \geq k) - P(N \geq k+1)) \\ &= \sum_{k=2}^{\infty} \frac{1}{(k-2)!} \\ &= \sum_{s=0}^{\infty} \frac{1}{s!} = e \end{aligned}$$

문제 4. 16. 문제를 제거한다.

문제 4. 17. 확률변수 R 의

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}$$

이라는 확률밀도함수를 가질 때 Rayleigh 분포라고 부른다. $E[R]$ 을 σ 를 이용해 표현하여라.

단, $\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$ 임을 이용하여도 좋다.

먼저, 확률밀도함수로부터 이 분포는 $r \geq 0$ 일 때만 정의됨을 알 수 있다.

$$\begin{aligned} E(X) &= \int_0^\infty r f_R(r) dr \\ &= \int_0^\infty \frac{r^2}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} dr \\ &= [-re^{-\frac{r^2}{2\sigma^2}}]_0^\infty + \int_0^\infty e^{-\frac{r^2}{2\sigma^2}} dr \\ &= \frac{1}{2} \int_{-\infty}^\infty e^{-\frac{r^2}{2\sigma^2}} dr \\ &= \frac{1}{2} \sqrt{2\pi} \sigma \\ &= \sqrt{\frac{\pi}{2}} \sigma \end{aligned}$$

이다.

문제 5. 1. 연속확률변수 X 가 확률밀도함수를

$$f(x) = \frac{1}{b-a} (a \leq x \leq b)$$

로 가질 때, 누적분포함수 $F(x)$ 를 구하여라.

$F(x)$ 는 x 가 a 보다 작을 때는 0이고, b 보다 클 때는 1임이 자명하다. 따라서, x 가 a 와 b 사이에 있을 때의 상황을 보자.

$$F(x) = \int_a^x f(t)dt = \int_a^s \frac{1}{b-a} dt = \frac{s-a}{b-a}$$

즉,

$$F(x) = \begin{cases} 0 & x < a \\ \frac{s-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

이다.

문제 5. 2. 규찬이와 규만이는 함께 콘서트장에 가려 하고 있다. 둘이 나오는 시각을 X, Y 라고 하면 X, Y 는 6과 7 사이의 균등분포이다. 즉, 6시부터 7시 사이에 둘이 집 밖으로 나오며, 그 분포는 균등분포 $U(6, 7)$ 을 따른다. 둘은 나온 이후 10분 동안 친구가 나오기를 기다리며, 10분 이후에도 나오지 않으면 그냥 먼저 출발해버린다. 규찬이와 규만이가 함께 콘서트장에 갈 확률은?

이 문제를 다르게 해석하면, 규찬이와 규만이가 나오는 시간 X 와 Y 에 대하여 $X - Y$ 의 절댓값이 10분 이하일 경우에는 둘이 같이 가고, 그렇지 않다면 따로 간다. X 와 Y 는 6과 7 사이에서 균등분포를 따르므로, X 를 x 좌표로 하고 Y 를 y 좌표로 하는 영역을 그리면 한 변의 길이가 1인 정사각형이며, 그 내부에서 점들이 골라질 확률은 모두 같다. 따라서 $|X - Y| \leq \frac{1}{6}$ 인 영역을 찾아주면 되는데, 그 방塊 모양의 넓이는 $\frac{11}{36}$ 이다. 따라서, 같이 갈 확률도 $\frac{11}{36}$ 이다.

문제 5. 3. 위에서 빠진 적분 부분을 계산해 위의 결과가 맞음을 보여라.

$$\begin{aligned} E[X] &= \int_0^\infty x \lambda e^{-\lambda x} dx \\ &= [-xe^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x} dx \\ &= [-\frac{1}{\lambda}e^{-\lambda x}]_0^\infty = \frac{1}{\lambda} \end{aligned}$$

$$\begin{aligned} V(X) &= \int_0^\infty (x - \frac{1}{\lambda})^2 \lambda e^{-\lambda x} dx \\ &= [-(x - \frac{1}{\lambda})^2 e^{-\lambda x}]_0^\infty + \int_0^\infty 2(x - \frac{1}{\lambda}) e^{-\lambda x} dx \\ &= \frac{1}{\lambda^2} + [-\frac{2}{\lambda}(x - \frac{1}{\lambda} e^{-\lambda x})]_0^\infty + \int_0^\infty \frac{2}{\lambda} e^{-\lambda x} dx \\ &= \frac{1}{\lambda^2} - \frac{2}{\lambda^2} + [-\frac{2}{\lambda^2} e^{-\lambda x}]_0^\infty \\ &= \frac{1}{\lambda^2} - \frac{2}{\lambda^2} + \frac{2}{\lambda^2} \\ &= \frac{1}{\lambda^2} \end{aligned}$$

문제 5. 4. 걸그룹이 1위를 하기 위해 걸리는 시간은 평균이 100인 지수분포를 따른다고 한다. 어느 걸그룹이 300일 동안 1위를 하지 못했을 때, 400일 안에 1위를 할 확률은 얼마인가?

기하분포의 무기억성에 의하여, 구하는 확률은 데뷔 100일 안에 1위를 할 확률과 같다. 평균이 100인 지수분포는 $\lambda = \frac{1}{100}$ 에 대하여 누적분포함수 $F(x)$ 가

$$F(x) = 1 - e^{-\frac{1}{100}x}$$

를 따르므로,

$$F(100) = 1 - \frac{1}{e}$$

이다. 즉, 구하는 확률은 $1 - e^{-1}$ 이다.

문제 5. 5. 확률변수 X 가 $X \sim N(m, \sigma^2)$ 일 때, 그 확률밀도함수 $f(x)$ 가 m 을 기준으로 대칭임을 보여라. 이를 통하여, 평균이 0인 정규분포의 누적분포함수 $F(x)$ 에 대하여

$$F(x) = 1 - F(-x)$$

임을 보여라.

$f(2m - x) = f(x)$ 임을 보이면 $f(x)$ 가 m 을 기준으로 대칭임을 알 수 있다.

$$f(2m - x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(2m-x-m)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} = f(x)$$

이므로, 증명이 완료된다. 평균이 0이면 $m = 0$ 이므로, $f(x)$ 는 0을 기준으로 대칭인 함수다. 그러면,

$$F(x) = \int_{-\infty}^x f(t)dt = \int_{-\infty}^x f(-t)dt = \int_{-x}^{\infty} f(t)dt = 1 - F(-x)$$

임을 알 수 있다.

문제 5. 6. 시간이 된다면,

$$\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}$$

임을 보여라.

범위가 아니므로 몰라도 된다. 이것은 아래의 식과 동치명제다.

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = 2\pi$$

$$\begin{aligned} \text{int}_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dxdy \\ &= \int_0^{\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr \\ &= 2\pi \int_0^{\infty} r e^{-\frac{r^2}{2}} dr \\ &= 2\pi [-e^{-\frac{r^2}{2}}]_0^{\infty} = 2\pi \end{aligned}$$

따라서,

$$\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}$$

이 성립한다.

문제 5. 7. 서울과학고 기초통계학 시험점수는 평균이 50, 표준편차가 10인 정규분포를 따른다고 한다. 시험에서 30점 이하인 학생의 비율은 대략 얼마인가?

평균 - 2표준편차보다 작은 학생의 비율은

$$\frac{1 - 0.954}{2} = 0.023$$

임을 알 수 있다.

문제 5. 8. $N(10, 2)$ 를 따른 정규분포 X 를 표준화하려면 Z 를 무엇으로 두어야 하는가? X 에 대해 표시하라.

만약 정규분포 X 가 평균 m , 표준편차 σ 인 분포를 따른다면 표준정규분포로 만들기 위해서는 평균을 0, 표준편차를 1로 만들어야 하기에

$$Z = \frac{X - 10}{\sqrt{2}}$$

이면 된다.

문제 5. 9. 한 개의 주사위를 720회 던져서 1의 눈이 나오는 횟수를 X 라 할 때, $P(90 \leq X \leq 140)$ 의 값을 구하여라.

표본의 크기가 충분히 크므로, $B(720, \frac{1}{6})$ 을 따른 확률변수 X 는 $N(120, 100)$ 으로 근사될 수 있다. 따라서,

$$P(90 \leq X \leq 140) = P\left(\frac{90 - 120}{10} \leq \frac{X - 120}{10} \leq \frac{140 - 120}{10}\right) \approx P(-3 \leq Z \leq 2) = 0.976$$

임을 알 수 있다. (연속성 수정 생략)

문제 5. 10. 어느 가수의 곡이 성공할 확률은 0.3이라 한다. 이 가수가 25개의 곡을 냈을 때, 10곡 이상 히트했을 확률을 정규분포를 이용해 근사하시오.

히트한 곡의 개수를 확률변수 X 에 대응시키면, $X \sim B(25, 0.3)$ 이다. 이를 정규분포로 근사하면, X 는 $N(7.5, 5.25)$ 을 근사적으로 따른다.

$$P(X \geq 10) \approx P(Z \geq 1.09) = 0.138$$

임을 확인할 수 있다.

문제 5. 11. U 는 $[0, 1]$ 에서 정의된 균등확률분포를 가지는 확률변수이다. $a + (b - a)U$ 의 확률분포를 구하라.

$a = b$ 이면 $a + (b - a)U = a$ 이므로 항상 같은 값을 가지는 확률변수이다. $a < b$ 라고 가정하였을 때 $a + (b - a)U$ 의 누적분포함수 $F(x)$ 는 다음과 같다.

$$F(x) = P(a + (b - a)U \leq x) = P\left(U \leq \frac{x - a}{b - a}\right) = \begin{cases} 0 & (x < a) \\ \frac{x-a}{b-a} & (a \leq x \leq b) \\ 1 & (x > b) \end{cases}$$

이로부터 $a + (b - a)U$ 의 확률밀도함수 $f(x)$ 를 아래와 같이 얻을 수 있다.

$$f(x) = \frac{d}{dx}F(x) = \frac{1}{b-a} \quad (a \leq x \leq b)$$

따라서 U 는 $[a, b]$ 에서 정의된 균등 확률분포를 가지는 확률변수이다. $a > b$ 인 경우에도 같은 결과를 얻을 수 있다.

문제 5. 12. 특정 전구는 일반적으로 수명이 평균이 2000이고 표준편차가 85인 정규분포를 따른다고 알려져 있다. 오직 5퍼센트만의 전구만이 L 보다 작은 수명을 가지게 하는 L 의 값을 구하시오. 단, $P(Z \geq z_\alpha) = \alpha$ 라고 정의한다. 또한, $z_{0.05} = 1.645$ 이다.

전구의 수명을 확률변수 X 로 놓았을 때, $X \sim N(2000, 85^2)$ 이다.

$$P(X < L) = P\left(Z < \frac{L - 2000}{85}\right) = P\left(Z > -\frac{L - 2000}{85}\right) = 0.05$$

따라서 $-\frac{L - 2000}{85} = z_{0.05} = 1.645$ 이고 $L = 1860.175$ 이다.

문제 5. 13. 기계를 고치는 데 필요한 시간은 $\lambda = 1$ 인 지수분포를 따른다고 한다.

- 1) 고치는 데 2시간 이상 걸릴 확률은?
- 2) 고치는 데 이미 2시간이 지났을 경우, 그 시점부터 3시간 안에 기계가 고쳐질 확률은?

기계를 고치는 데에 필요한 시간을 X 로 놓았을 때, $X \sim \exp(1)$ 이다.

1. 지수확률변수의 누적분포함수는 $F(t) = 1 - e^{-\lambda t}$ 으로 $P(X \geq 2) = e^{-1 \cdot 2} = e^{-2}$ 이다.
2. 구하고자 하는 확률은 $P(X < 3 | X > 2)$ 이다. 지수확률변수의 무기역성에 의하여 아래와 같이 구할 수 있다.

$$P(X < 3 | X > 2) = 1 - P(X > 3 | X > 2) = 1 - P(X > 1) = 1 - e^{-1 \cdot 1} = 1 - e^{-1}$$

문제 5. 14. 확률분포 X 는 만약 $\log X$ 가 정규분포를 이룬다면 로그정규분포라고 부른다. X 가 로그정규분포이고 $E[\log X] = \mu$, $V(\log X) = \sigma^2$ 이라고 할 때, X 의 누적분포함수와 확률밀도함수를 구하여라.

시작하기에 앞서, $\log X$ 가 잘 정의되어야 하므로 $X > 0$ 임을 유념하자. X 의 누적분포함수는 $F(x)$ 는 아래와 같이 구할 수 있다. 이 함수를 더이상 간단히 정리할 필요는 없다.

$$F(x) = P(X \leq x) = P(\log X \leq \log x) = \int_{\log x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \quad (x > 0)$$

확률밀도함수 $f(x)$ 는 $F(x)$ 를 미분하여 구할 수 있다. 미적분학의 기본정리와 연쇄 법칙을 사용하면,

$$f(x) = \frac{d}{dx} F(x) = \frac{d}{dx} \int_{\log x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \quad (x > 0)$$

문제 5. 15. 이차원 과녁에 총을 쏘 때, 수평 방향의 오차는 평균 0에 분산 4이며, 수직 방향의 오차도 동일한 평균과 동일한 분산을 가진다. 그리고 오차들은 모두 정규분포를 따른다고 한다. D 를 과녁으로부터 실제 총이 쏘인 부분 사이의 거리라고 둘 때, $E[D^2]$ 를 구하여라.

수평 방향 오차를 X , 수직 방향 오차를 Y 로 놓자. $X \sim N(0, 2^2)$, $Y \sim N(0, 2^2)$ 이고, $D^2 = X^2 + Y^2$ 이다. D^2 의 기댓값은 아래와 같이 구할 수 있다.

$$\begin{aligned} E[D^2] &= E[X^2 + Y^2] = E[X^2] + E[Y^2] \\ &= V(X) + V(Y) = 4 + 4 = 8 \end{aligned}$$

문제 5. 16. 신명호의 3점슛 성공률은 50퍼센트이며, 한 게임에서 그가 40번의 삼점슛을 시도하였다고 하자. 이때, 각 삼점슛의 성공 여부는 다른 삼점슛의 성공 여부와는 독립이다. 이때, 정규분포로의 근사를 통해 그가 3점슛으로 60점을 얻어낼 확률을 구하시오. 단, $P(0 \leq Z \leq \frac{1}{2\sqrt{10}}) = 0.0636$ 으로 둔다.

$i = 1, 2, \dots, 40$ 에 대하여 확률변수 X_i 를 아래와 같이 정의하자.

$$X_i = \begin{cases} 1 & i\text{번째 3점슛이 성공함} \\ 0 & i\text{번째 3점슛이 실패함} \end{cases}$$

그러면 X_1, X_2, \dots, X_{40} 은 모두 독립적으로 동일한 베르누이 분포를 따른다. $n = 40$, $\mu = E[X_i] = 0.5$, $\sigma = \sqrt{V(X_i)} = \sqrt{0.5 \cdot 0.5} = 0.5$ 로 정의하자. 중심극한정리에 의하여 $X_1 + X_2 + \dots + X_{40}$ 의 분포는 $S \sim N(n\mu, n\sigma^2)$ 으로 근사할 수 있다.

$$\begin{aligned} P(X_1 + X_2 + \dots + X_{40} = 20) &\approx P(19.5 \leq S \leq 21.5) \\ &= P\left(\frac{19.5 - 20}{\frac{\sqrt{40}}{2}} \leq Z \leq \frac{20.5 - 20}{\frac{\sqrt{40}}{2}}\right) = 2P\left(0 \leq Z \leq \frac{1}{2\sqrt{10}}\right) = 0.1272 \end{aligned}$$

문제 5. 17. 주어진 정규분포 $N(m, \sigma^2)$ 의 확률밀도함수 $f(x)$ 에 대하여, $\phi_{m, \sigma^2}(x) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$ 라고 정의하자.

일반적으로, $\phi(x)$ 가 정해지면 그 분포 역시도 정해진다. 또한, X, Y 가 독립일 경우 $X + Y$ 의 $\phi(x)$ 는 X 의 $\phi(x)$ 와 Y 의 $\phi(x)$ 를 곱한 것임이 알려져 있다.

확률변수 X_1, X_2, \dots, X_n 이 정규분포 $N_1(m_1, \sigma_1^2)$ 부터 $N_n(m_n, \sigma_n^2)$ 을 따른다. 이때, $X_1 + X_2 + \dots + X_n$ 이 어떤 분포를 따르는지 이야기하고, 평균과 분산을 계산하라.

먼저 $X \sim N(m, \sigma^2)$ 의 ϕ 를 직접 계산해보자.

$$\phi_{m, \sigma}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} e^{tx} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m+ta)^2}{2\sigma^2}} e^{mt+\frac{1}{2}\sigma^2 t^2} dx = e^{mt+\frac{1}{2}\sigma^2 t^2}$$

$X_1 + X_2 + \dots + X_n$ 의 ϕ 는 각각의 ϕ 를 곱한 것과 같으므로,

$$\phi(t) = e^{m_1 t + \frac{1}{2}\sigma_1^2 t^2} e^{m_2 t + \frac{1}{2}\sigma_2^2 t^2} \dots e^{m_n t + \frac{1}{2}\sigma_n^2 t^2} = e^{(m_1 + m_2 + \dots + m_n)t + \frac{1}{2}(\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2)t^2}$$

이는 앞서 구한 정규분포의 ϕ 와 같은 형태를 가진다. 따라서 $X_1 + X_2 + \dots + X_n$ 은 정규분포를 따르며, 평균과 분산은 각각 $m_1 + m_2 + \dots + m_n$ 과 $\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2$ 이다.
(즉, 정규분포를 따르는 확률변수들의 집합은 덧셈에 대해 닫혀 있고, 합하여 얻은 정규분포의 평균/분산은 각각의 평균/분산의 합과 같다.)

문제 5. 18. 나PD는 신서유기 프로그램에서 n 명의 멤버들에게 적어도 k 명이 지압마사지를 버티면 그동안 식사를 제공하는 k out of n 시스템을 도입하려 한다. 특히 $k = n$ 이면 직렬시스템, $k = 1$ 이면 병렬시스템이라 부른다. 각 멤버들이 지압마사지를 버티는 시간 T_i , $i = 1, 2, \dots, n$ 이 서로 독립이고 모두 $EXP(\lambda)$ 를 따를 때, 식사를 제공받는 시간 T_s 에 대하여 다음 물음에 답하라.

- (a) 직렬시스템의 T_s 의 기댓값을 구하라.
- (b) 병렬시스템의 T_s 의 기댓값을 구하라.

T_s 의 누적분포함수를 $F(x)$ 로 놓자.

1. 직렬시스템의 경우 $T_s = \min\{T_1, T_2, \dots, T_n\}$ 이다. 그러면

$$\begin{aligned} F(x) &= 1 - P(T_s \geq x) \\ &= 1 - P(T_1 \geq x, T_2 \geq x, \dots, T_n \geq x) \\ &= 1 - P(T_1 \geq x) P(T_2 \geq x) \dots P(T_n \geq x) = 1 - e^{-\lambda x} e^{-\lambda x} \dots e^{-\lambda x} = 1 - e^{-n\lambda x} \end{aligned}$$

따라서 $T_s \sim \exp(n\lambda)$ 이고, 기댓값은 $E[T_s] = \frac{1}{n\lambda}$ 이다.

2. 병렬시스템의 경우 $T_s = \max \{T_1, T_2, \dots, T_n\}$ 이다. 그러면

$$\begin{aligned} F(x) &= P(T_s \leq x) \\ &= P(T_1 \leq x, T_2 \leq x, \dots, T_n \leq x) \\ &= P(T_1 \leq x) P(T_2 \leq x) \cdots P(T_n \leq x) = (1 - e^{-\lambda x})^n \end{aligned}$$

이고 확률밀도함수 $f(x)$ 는 다음과 같다.

$$f(x) = \frac{d}{dx} F(x) = \lambda n e^{-\lambda x} (1 - e^{-\lambda x})^{n-1} \quad (x \geq 0)$$

기댓값은 아래와 같이 이항전개하여 구할 수 있다.

$$\begin{aligned} E[T_s] &= \int_0^\infty \lambda n e^{-\lambda x} (1 - e^{-\lambda x})^{n-1} x dx \\ &= \lambda n \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \int_0^\infty e^{-(k+1)\lambda x} x dx \\ &= \lambda n \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \frac{1}{((k+1)\lambda)^2} = \frac{1}{\lambda} \sum_{k=0}^{n-1} \binom{n}{k+1} \frac{(-1)^k}{k+1} \end{aligned}$$

이때, $I_n = \sum_{k=0}^{n-1} \binom{n}{k+1} \frac{(-1)^k}{k+1} = \sum_{k=1}^n \binom{n}{k} \frac{(-1)^{k-1}}{k}$ 으로 두면 $I_1 = 1$ 이다.

$$\begin{aligned} I_{n+1} - I_n &= \frac{(-1)^n}{n+1} + \sum_{k=1}^n \left\{ \binom{n+1}{k} - \binom{n}{k} \right\} \frac{(-1)^{k-1}}{k} \\ &= \frac{(-1)^n}{n+1} + \sum_{k=1}^n \binom{n+1}{k} \frac{k}{n+1} \frac{(-1)^{k-1}}{k} \\ &= \frac{1}{n+1} \end{aligned}$$

이므로 $I_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ 이다. 따라서 $E[T_s] = \frac{1}{\lambda} + \frac{1}{2\lambda} + \cdots + \frac{1}{n\lambda}$ 이다.

문제 5.19. 서로 독립인 확률변수 U_1, U_2, \dots, U_n 이 모두 $U(0, 1)$ 을 따를 때, $X = \max(U_1, U_2, \dots, U_n), Y = \min(U_1, U_2, \dots, U_n)$ 이라고 정의하자. X 와 Y 의 확률밀도함수를 구하여라.

X 와 Y 의 누적분포함수를 각각 F_X 와 F_Y , 확률밀도함수를 각각 f_X 와 f_Y 로 놓자.

$$F_X(x) = P(X \leq x) = P(U_1 \leq x) P(U_2 \leq x) \cdots P(U_n \leq x) = \begin{cases} 0 & x < 0 \\ x^n & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

$$\text{이므로 } f_X(x) = \frac{d}{dx} F_X(x) = \begin{cases} nx^{n-1} & 0 \leq x \leq 1 \\ 0 & x < 0, x > 1 \end{cases} \text{이다.}$$

$$F_Y(x) = 1 - P(Y \geq x) = 1 - P(U_1 \geq x) P(U_2 \geq x) \cdots P(U_n \geq x) = \begin{cases} 0 & x < 0 \\ 1 - (1-x)^n & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

$$\text{이므로 } f_Y(x) = \frac{d}{dx} F_Y(x) = \begin{cases} n(1-x)^{n-1} & 0 \leq x \leq 1 \\ 0 & x < 0, x > 1 \end{cases} \text{이다.}$$

문제 5. 20. U_1, U_2, U_3 는 독립인 확률변수이며 모두 $U(0, 1)$ 을 따른다.

- 1) $U_1 + U_2$ 의 확률밀도함수를 구하라.
- 2) $U_1 + U_2 + U_3$ 의 확률밀도함수를 구하여라.

1. 누적분포함수를 F_1 , 확률밀도함수를 f_1 으로 두자. $[0, 1] \times [0, 1]$ 영역에서 넓이를 계산하면 아래와 같이 F_1 을 구할 수 있다.

$$F_1(x) = P(U_1 + U_2 \leq x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}x^2 & 0 \leq x \leq 1 \\ 1 - \frac{1}{2}(1-x)^2 & 1 \leq x \leq 2 \\ 1 & x > 2 \end{cases}$$

$$\text{따라서 } f_1 \stackrel{\text{def}}{=} f_1(x) = \frac{d}{dx} F_1(x) = \begin{cases} 0 & x < 0, x > 2 \\ x & 0 \leq x \leq 1 \\ 1-x & 1 \leq x \leq 2 \end{cases} \text{이다.}$$

2. 누적분포함수를 F_2 , 확률밀도함수를 f_2 으로 두자. $[0, 1] \times [0, 1] \times [0, 1]$ 공간에서 적분을 계산하면 아래와 같이 F_2 을 구할 수 있다.

$$F_2(x) = P(U_1 + U_2 + U_3 \leq x) = \begin{cases} 0 & x < 0 \\ \frac{1}{3}x^3 & 0 \leq x \leq 1 \\ \frac{1}{3}x^3 - 3 \cdot \frac{1}{3}(x-1)^3 & 1 \leq x \leq 2 \\ 1 - \frac{1}{3}(3-x)^3 & 2 \leq x \leq 3 \\ 1 & x > 3 \end{cases}$$

$$\text{따라서 } f_2 \stackrel{\text{def}}{=} f_2(x) = \frac{d}{dx} F_2(x) = \begin{cases} 0 & x < 0, x > 3 \\ x^2 & 0 \leq x \leq 1 \\ -2x^2 + 6x - 3 & 1 \leq x \leq 2 \\ x^2 - 6x + 9 & 2 \leq x \leq 3 \end{cases} \text{이다.}$$

문제 5. 21. 임의의 점 P 가 수직선 위에 있으며, P 의 위치는 $\lambda = 1$ 인 지수분포를 따른다. 반면 점 U 는 구간 $[0, l]$ 에서 정의되는 균등분포를 따라 분포한다. P 와 U 로 수직선을 잘랐을 때, P 가 포함된 구간의 길이를 X 라 하자. $E[X]$ 의 값은? 단, P 와 U 의 위치는 독립이다.

확률변수 Y 를 $Y = P - U$ 로 정의하자. Y 의 누적분포함수 F_Y 는 아래와 같이 구할 수 있다.

$$F_Y(x) = \Pr(P - U \leq x) = \int_0^l \int_0^{x+u} f_U(u) f_P(v) dv du = \begin{cases} 0 & x < -l \\ 1 + \frac{x-1}{l} + \frac{e^{-x-l}}{l} & -l \leq x \leq 0 \\ 1 - \frac{e^{-x}(1-e^{-l})}{l} & x > 0 \end{cases}$$

확률변수 X 는 $X = |Y|$ 이므로, $x \geq 0$ 에 대해 누적분포함수 F_X 는 다음과 같이 구할 수 있다.

$$F_X(x) = F_Y(x) - F_Y(-x) = \begin{cases} \frac{x+1-e^{-x}+e^{-x-l}-e^{x-l}}{l} & 0 \leq x \leq l \\ 1 - \frac{e^{-x}(1-e^{-l})}{l} & x > l \end{cases}$$

$$\text{확률밀도함수 } f_X \stackrel{\text{def}}{=} f_X(x) = \frac{d}{dx} F_X(x) = \begin{cases} \frac{1+e^{-x}-e^{-x-l}-e^{x-l}}{l} & 0 \leq x \leq l \\ \frac{e^{-x}}{l} & x > l \end{cases} \text{이고, 기댓값은 아래와 같이 구할}$$

수 있다.

$$\begin{aligned}
 E[X] &= \int_0^l x \frac{1 + e^{-x} - e^{-x-l} - e^{x-l}}{l} dx + \int_l^\infty x \frac{e^{-x}}{l} dx \\
 &= \frac{1}{l} \left[\frac{x^2}{2} + (1 - e^{-l})(-xe^{-x} - e^{-x}) - e^{-l}(xe^x - e^x) \right]_0^l + \frac{1}{l} \left[-xe^{-x} - e^{-x} \right]_l^\infty \\
 &= \frac{l}{2} + (1 - e^{-l})(-le^{-l} - e^{-l} + 1) - e^{-l}(le^l - e^l + 1) + \frac{1}{l}(le^{-l} + e^{-l}) \\
 &= le^{-2l} + \left(\frac{1}{l} - l - 2 \right) e^{-l} + \left(2 - \frac{l}{2} \right)
 \end{aligned}$$

문제 5. 22. 이 문제는 삭제하도록 한다.

문제 5. 23. $\{X_k\}$ 를 독립된 균등확률분포를 따르는 확률변수의 수열이라고 두자. a_k 가 양의 실수라면, $X_k \sim U(0, a_k)$ 를 따른다. $S_n = \sum_{k=1}^n X_k$ 라고 두며, $f_n(x)$ 와 $F_n(x)$ 를 그 확률밀도함수와 누적분포함수라고 부르자.

1) $F_1(x) = \frac{\max(x, 0) - \max(x - a_1, 0)}{a_1}$ 임을 보여라.

2) $f_{n+1}(x) = \frac{F_n(x) - F_n(x - a_{n+1})}{a_{n+1}}$ 임을 보여라.

3) $F_2(x)$ 를 구하여라.

4) $x^+ = \max(x, 0)$ 으로 정의하자. 만약 모든 a_k 의 값이 a 로 같다면,

$$F_n(x) = \frac{1}{n!a^n} \sum_{r=0}^n (-1)_n^r C_r [(x - ra)^+]^n$$

임을 보여라.

시작하기에 앞서 $b_n = a_1 + a_2 + \dots + a_n$ 으로 두자.

1. $S_1 = X_1$ 의 확률밀도함수는

$$f_1(x) = \begin{cases} 0 & x < 0, x > a_1 \\ \frac{1}{a_1} & 0 \leq x \leq a_1 \end{cases}$$

이므로

$$F_1(x) = \int_{-\infty}^x f_1(t) dt = \begin{cases} 0 & x < 0 \\ \frac{x}{a_1} & 0 \leq x \leq a_1 \\ 1 & x > a_1 \end{cases} = \frac{\max\{x, 0\} - \max\{x - a_1, 0\}}{a_1}$$

이다.

2.

$$\begin{aligned}
 F_{n+1}(x) &= \int_{x-a_{n+1}}^x \frac{x-s}{a_{n+1}} f_n(s) ds + F_n(x - a_{n+1}) \\
 &= \int_{x-a_{n+1}}^x \frac{x}{a_{n+1}} f_n(s) ds - \int_{x-a_{n+1}}^x \frac{s}{a_{n+1}} f_n(s) ds + F_n(x - a_{n+1})
 \end{aligned}$$

이므로 양변을 미분하면

$$\begin{aligned}
 a_{n+1} f_{n+1}(x) &= \int_{x-a_{n+1}}^x f_n(s) ds + x(f_n(x) - f_n(x - a_{n+1})) - xf_n(x) + (x - a_{n+1})f_n(x - a_{n+1}) + a_{n+1}f_n(x - a_{n+1}) \\
 &= F_n(x) - F_n(x - a_{n+1})
 \end{aligned}$$

이를 정리하면

$$f_{n+1}(x) = \frac{F_n(x) - F_n(x - a_{n+1})}{a_{n+1}}$$

3. $f_2(x) = \frac{F_1(x) - F_1(x - a_2)}{a_2} = \frac{\max\{x, 0\} - \max\{x - a_1, 0\} - \max\{x - a_2, 0\} + \max\{x - a_1 - a_2, 0\}}{a_1 a_2}$ 이다. $m = \min\{a_1, a_2\}$, $M = \max\{a_1, a_2\}$ 로 놓고 전개하면

$$f_2(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{a_1 a_2} & 0 \leq x \leq m \\ \frac{1}{M} & m \leq x \leq M \\ \frac{a_1 + a_2 - x}{a_1 a_2} & M \leq x \leq a_1 + a_2 \\ 0 & x > a_1 + a_2 \end{cases}$$

이를 적분하면 F_2 를 얻는다.

$$F_2(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{2a_1 a_2} & 0 \leq x \leq m \\ \frac{m^2}{2a_1 a_2} + \frac{x-m}{M} & m \leq x \leq M \\ 1 - \frac{\frac{1}{2}(a_1 + a_2)^2 - (a_1 + a_2)x + \frac{1}{2}x^2}{a_1 a_2} & M \leq x \leq a_1 + a_2 \\ 1 & x > a_1 + a_2 \end{cases}$$

4. 수학적 귀납법을 사용하자. $n = 1$ 에 대해서 등식이 성립한다는 것을 쉽게 확인할 수 있고, 어떤 $n \geq 1$ 에 대해서 등식이 성립한다고 가정하자.

$$\begin{aligned} f_{n+1}(x) &= \frac{1}{n! a^{n+1}} \left(\sum_{r=0}^n (-1)^r \binom{n}{r} [(x - ra)^+]^n - \sum_{r=0}^n (-1)^r \binom{n}{r} [(x - a - ra)^+]^n \right) \\ &= \frac{1}{n! a^{n+1}} \left((x^+)^n - (-1)^n [(x - (n+1)a)^+]^n + \sum_{r=1}^n \left\{ (-1)^r \binom{n}{r} - (-1)^{r-1} \binom{n}{r-1} \right\} [(x - ra)^+]^n \right) \\ &= \frac{1}{n! a^{n+1}} \sum_{r=0}^{n+1} (-1)^r \binom{n+1}{r} [(x - ra)^+]^n \end{aligned}$$

이를 적분하면 $F_{n+1}(x) = \frac{1}{(n+1)! a^{n+1}} \sum_{r=0}^{n+1} (-1)^r \binom{n+1}{r} [(x - ra)^+]^{n+1}$ 이다. 수학적 귀납법에 의해 등식은 모든 n 에 대해 성립한다.

문제 5. 24. X_i 를 독립된 확률변수라고 두자. 이들이 각각 모수가 i 인 지수분포를 따를 때, $Z = \min(X_1, \dots, X_{10})$ 으로 정의하자.

- 1) $P(Z > 2)$ 의 값을 구하라.
- 2) $V(Z)$ 의 값을 구하여라.

1.

$$P(Z > 2) = P(X_1 > 2)P(X_2 > 2) \cdots P(X_{10} > 2) = e^{-2}e^{-4} \cdots e^{-20} = e^{-110}$$

2. $P(Z > x) = P(X_1 > x)P(X_2 > x) \cdots P(X_{10} > x) = e^{-x}e^{-2x} \cdots e^{-10x} = e^{-55x}$ 이므로 $Z \sim \exp(55)$ 이다. 따라서 $V(Z) = \frac{1}{55^2} = \frac{1}{3025}$ 이다.

문제 5. 25. $F(x) = \frac{e^x}{e^x + e^{-x}}$ 의 누적분포함수임을 보이고, 확률밀도함수를 찾아라.

$F(x) \geq 0 \forall x$ 임은 쉽게 보일 수 있다. 또한, $\lim_{x \rightarrow -\infty} F(x) = 0$ 이고 $\lim_{x \rightarrow \infty} F(x) = 1$ 이며 전 구간에서 단조증가하고 연속이므로 F 는 누적분포함수이다. 확률밀도함수 $f(x)$ 는 미분하여 구할 수 있다.

$$f(x) = \frac{d}{dx} F(x) = \frac{e^x (e^x + e^{-x}) - e^x (e^x - e^{-x})}{(e^x + e^{-x})^2} = \frac{2}{(e^x + e^{-x})^2}$$

문제 6. 1. 확률변수 X 가 $POI(3)$ 을 따를 때, 크기가 16인 표본평균 \bar{X} 의 분산은 얼마인가?

원래 확률변수 X 의 평균은 3, 분산은 3이다. 표본평균의 분산은 이를 16으로 나눈 $\frac{3}{16}$ 이다.

문제 6. 2. 전국 고등학교 학생들의 모의고사 점수는 평균이 60점이고, 표준편차가 15점이라고 한다. 고등 학생 중 100명을 임의로 추출해 표본평균을 구했을 때, 그 값이 63점 이상일 확률을 구하여라.

표본평균 \bar{X} 의 분포는 $N(60, \frac{15^2}{100})$ 을 따른다. 즉, 평균이 60이고 표준편차가 1.5인 정규분포를 따른다. 따라서

$$\begin{aligned} P(\bar{X} \geq 63) &= P\left(\frac{\bar{X} - 60}{1.5} \geq \frac{63 - 60}{1.5}\right) \\ &= P(Z \geq 2) = 0.023 \end{aligned}$$

문제 6. 3. $\chi_{k,\alpha}^2 > \chi_{k,1-\alpha}^2$ 가 성립하는 α 의 범위를 구하여라.

$\chi_{k,1-\alpha}^2$ 는 자유도가 k 인 카이제곱분포에서의 100α 백분위수에 해당하는 값이며, $\chi_{k,\alpha}^2$ 는 자유도가 k 인 카이제곱분포에서의 $100(1 - \alpha)$ 백분위수에 해당하는 값이다. 따라서, $\chi_{k,\alpha}^2$ 가 더 크기 위해서는 $1 - \alpha > \alpha$ 여야 한다. 즉, $\frac{1}{2} < \alpha \leq 1$ 이다.

문제 6. 4. 문제 삭제

문제 6. 5. 위의 등식에서는 빠진 부분이 조금 있다. 사이사이를 채워넣어라.

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m + m - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n [(X_i - m)^2 + 2(X_i - m)(m - \bar{X}) + (m - \bar{X})^2] \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m)^2 + \frac{2(m - \bar{X})}{n-1} \sum_{i=1}^n (X_i - m) + \frac{n}{n-1} (m - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m)^2 + \frac{2(m - \bar{X})}{n-1} (n\bar{X} - nm) + \frac{n}{n-1} (m - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m)^2 - \frac{2n}{n-1} (m - \bar{X})^2 + \frac{n}{n-1} (m - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - m)^2 - \frac{n}{n-1} (m - \bar{X})^2 \end{aligned}$$

이므로,

$$E\left(\sum_{i=1}^n (X_i - m)^2\right) = \sum_{i=1}^n E((X_i - m)^2) = n\sigma^2$$

이며

$$E((\bar{X} - m)^2) = V(\bar{X}) = \frac{\sigma^2}{n}$$

임을 고려한다면

$$E[S^2] = \frac{1}{n-1} n\sigma^2 - \frac{n}{n-1} \frac{\sigma^2}{n} = \sigma^2$$

이 될 것이다.

문제 6. 6. $\frac{X_i - m}{\sigma}$ 가 어떤 분포를 따르는지 말하라.

$X_i \sim N(m, \sigma^2)$ 이므로, 주어진 식은 X_i 를 표준화한 것과 같다. 따라서 이는 표준정규분포 Z 를 따른다.

문제 6. 7. $\sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2$ 가 어떤 분포를 따르는지 말하라.

서로 독립인 표준정규분포의 제곱 n 개를 더한 것이므로, 이는 $\chi^2(n)$ 을 따른다.

문제 6. 8. $\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$ 가 아래같이 표현됨을 보여라.

$$\sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 - \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2$$

$$\begin{aligned} \sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 &= \sum_{i=1}^n \left(\frac{X_i - \bar{X} + \bar{X} - m}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2 + 2 \sum_{i=1}^n \left(\frac{(X_i - \bar{X})(m - \bar{X})}{\sigma^2} \right) \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2 + 2(m - \bar{X}) \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma^2} \right) \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2 + \frac{2(m - \bar{X})}{\sigma^2} (n\bar{X} - n\bar{X}) \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2 \end{aligned}$$

이다. 따라서 양변을 잘 정리해주면

$$\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 - \sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2$$

문제 6. 9. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 임을 보여라.

위의 문제에서

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

이며

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

이고,

$$\sum_{i=1}^n \left(\frac{\bar{X} - m}{\sigma} \right)^2 = \left(\frac{\bar{X} - m}{\sigma/\sqrt{n}} \right)^2 = \chi^2(1)$$

이다. 서로 독립임을 가정하면, 자유도가 n 인 카이제곱분포에서 1인 카이제곱분포를 뺀 것이므로 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 임을 확인할 수 있다.

문제 6. 10. 서울과학고 기초통계학 시험의 점수는 평균이 500이고 표준편차가 100인 정규분포를 따른다. 만약 5명의 학생이 임의로 골라진다면,

- 1) 모든 학생의 점수가 600 이하일 확률은?
- 2) 그들 중 정확히 3명만 640점 이상일 확률은?
- 3) 5명의 평균점수가 600점 이하일 확률은?
- 4) 5명의 평균점수가 560점 이하일 확률은?

1) 한 학생의 점수가 600 이하일 확률을 먼저 구하여 보자. 학생의 점수를 X 라고 하면, $P(X \leq 600)$ 의 값을 구하면 된다.

$$P(X \leq 600) = P\left(\frac{X - 500}{100} \leq \frac{600 - 500}{100}\right) = P(Z \leq 1) = 0.841$$

이다. 따라서 모든 학생의 점수가 이것 이하일 확률은 $(0.841)^5$ 이다.

- 2) 한 학생의 점수가 640점 이상일 확률을 먼저 구하여 보자.

$$P(X \geq 640) = P\left(\frac{X - 500}{100} \geq \frac{640 - 500}{100}\right) = P(Z \geq 1.4) = 0.081$$

이다. 해당하는 확률분포는 이항분포이므로

$$\binom{5}{3}(0.081)^3(0.919)^2$$

- 3) 5명에 대한 표본평균의 분포는 $N(500, \frac{100^2}{5})$ 이다. 따라서 평균 점수를 \bar{X} 라고 하면,

$$P(\bar{X} \leq 600) = P\left(\frac{\bar{X} - 500}{20\sqrt{5}} \leq \frac{600 - 500}{20\sqrt{5}}\right) = P(Z \leq \sqrt{5}) = 0.987$$

4)

$$P(\bar{X} \leq 560) = P\left(\frac{\bar{X} - 500}{20\sqrt{5}} \leq \frac{560 - 500}{20\sqrt{5}}\right) = P(Z \leq \frac{3\sqrt{5}}{5}) = 0.91$$

문제 6. 11. 방탄소년단의 연간 스트리밍 수는 평균이 40.14억이고 표준편차가 8.7억인 정규분포를 따른다고 하자.

- 1) 올해 신곡의 스트리밍 수가 42억을 넘을 확률은?
- 2) 차후 2년간 스트리밍 수가 84억을 넘을 확률은?
- 3) 차후 3년간 스트리밍 수가 126억을 넘을 확률은?

1)

$$P(X > 42) = P\left(\frac{X - 40.14}{8.7} > \frac{42 - 40.14}{8.7}\right) = P(Z > 0.23) = 0.409$$

2)

$$P(\bar{X} > 42) = P\left(\frac{\bar{X} - 40.14}{8.7/\sqrt{2}} > \frac{42 - 40.14}{8.7/\sqrt{2}}\right) = P(Z > 0.325) = 0.373$$

3)

$$P(\bar{X} > 42) = P\left(\frac{\bar{X} - 40.14}{8.7/\sqrt{3}} > \frac{42 - 40.14}{8.7/\sqrt{3}}\right) = P(Z > 0.398) = 0.345$$

문제 6. 12. 양궁 국대 기보배는 엑스텐(과녁의 정중앙을 맞추는 것)을 노리고 있다. 이때, 수직 오차는 평균이 0이고 분산이 4인 정규분포를 따른며, 수평 오차는 수직 오차와는 독립이며 평균이 0이고 분산이 4인 정규분포를 따른다. D 는 화살이 맞은 점과 정중앙 사이의 거리라고 할 때, $E[D^4]$ 의 값은?

수평 오차를 확률변수 X , 수직 오차를 확률변수 Y 에 대응시키면 $D^2 = X^2 + Y^2$ 이며, X 와 Y 는 $N(0, 4)$ 를 따르므로 $\frac{X}{2}$ 와 $\frac{Y}{2}$ 는 X 와 Y 를 표준화한 것이며 Z 를 따른다. 따라서 $\frac{X^2 + Y^2}{4} \sim \chi^2(2)$ 이다.

$$\frac{D^2}{4} \sim \chi^2(2)$$

카이제곱분포로부터

$$E\left(\frac{D^2}{4}\right) = 2$$

$$V\left(\frac{D^2}{4}\right) = 4$$

임을 알고 있다. 따라서,

$$E(D^4) = 16E\left(\frac{D^4}{16}\right) = V\left(\frac{D^2}{4}\right) + E\left(\frac{D^2}{4}\right)^2 = 4 + 4 = 8$$

임을 확인할 수 있다.

문제 6. 13. X 와 Y 는 서로 독립인 카이제곱분포를 따르는 확률변수이며, 각각의 자유도는 3과 6이다. $X+Y$ 의 값이 10보다 클 확률을 구하여라. $\chi_{0.35}^2(9) = 10$ 이다.

각각이 서로 독립인 카이제곱분포를 따르므로, $X+Y \sim \chi^2(9)$ 임을 확인할 수 있다. 따라서,

$$P(X+Y > 10) = P(X+Y > \chi_{0.35}^2(9)) = 0.35$$

이다.

문제 6. 14. 10개의 주사위가 굴려져서, 나온 값의 총합이 30과 40 사이일 확률을 근사하여라.

하나의 주사위에 대해서는 주사위에서 나오는 값을 확률변수 X 에 대응시킬 때 $E(X) = 3.5, V(X) = 35/12$ 이다. 이를 10번 던지게 된다면, 나온 값의 총합은 $N(35, \frac{350}{12})$ 에 근사시킬 수 있다. 즉,

$$P(30 < \sum_{i=1}^{10} X_i < 40) = P\left(\frac{30 - 35}{5\sqrt{\frac{7}{6}}} < \frac{\sum_{i=1}^{10} X_i - 35}{5\sqrt{\frac{7}{6}}} < \frac{40 - 35}{5\sqrt{\frac{7}{6}}}\right) \approx P(-0.926 < Z < 0.926) = 0.646$$

문제 6. 15. 16개의 서로 독립인 균등분포 $U(0, 1)$ 을 따르는 확률변수들의 합이 10보다 클 확률을 근사하시오.

균등분포 안에서 $E(X) = 0.5, V(X) = \frac{1}{12}$ 이다. 따라서, 확률변수들의 총합은 $N(8, \frac{16}{12})$ 에 근사시킬 수 있다.

$$P\left(\sum_{i=1}^{16} U_i > 10\right) = P\left(\frac{\sum_{i=1}^{16} U_i - 8}{\sqrt{\frac{4}{3}}} > \frac{10 - 8}{\sqrt{\frac{4}{3}}}\right) \approx P(Z > 1.73) = 0.042$$

문제 6. 16. 50개의 숫자들의 가장 가까운 정수로부터의 차이가 -0.5와 0.5 사이에서 균등분포를 이룰 때, 50개를 그냥 합했을 때와 반올림해서 합했을 때의 차이가 3 이상일 확률을 근사하라.

$U(-0.5, 0.5)$ 를 따르는 분포 50개를 더했을 때 그 절댓값이 3 이상이 될 확률을 묻는 문제이다. 해당 분포의 경우 기댓값은 0, 분산은 $\frac{1}{12}$ 이다. 즉, 그 합은 $N(0, \frac{50}{12})$ 를 따른다.

$$P\left(\left|\sum_{i=1}^{50} U_i\right| > 3\right) \approx P(|Z| > \frac{3 - 0}{\sqrt{\frac{50}{12}}}) = 2P(Z > 1.47) = 0.142$$

문제 6. 17. 케이티엔지는 그들의 담배에 포함된 니코틴의 양이 평균 $2.2mg$ 에 표준편차가 $0.3mg$ 인 정규 분포를 따르는 확률변수라고 주장하였다. 그러나, 100개의 담배를 모은 다음 얻은 표본평균은 $3.1mg$ 였다. 만약 담배공사의 주장이 옳을 때, 표본평균이 $3.1mg$ 이상으로 나올 확률은 얼마인지 근사하라.

담배공사의 주장이 옳을 때 표본평균 \bar{X} 의 분포는

$$\bar{X} \sim N(2.2, \frac{0.3^2}{100})$$

여야 한다. 따라서,

$$P(\bar{X} > 3.1) = P\left(\frac{\bar{X} - 2.2}{\frac{0.3}{\sqrt{10}}} > \frac{3.1 - 2.2}{\frac{0.3}{\sqrt{10}}}\right) = P(Z > 30) \approx 0$$

이다.

문제 6. 18. 기초통계학 선생님께서는 오랜 경험을 통해 시험 점수가 평균 77점에, 표준편차 15점임을 알아냈다. 이 교사는 현재 두 반을 가르치고 있는데, 한 반은 25명이고 한 반은 64명이다.

- 1) 25명 분반에서 평균 시험 점수가 72와 82 사이일 확률을 근사하라.
- 2) 64명 분반에 대해서는 어떤가?
- 3) 25명 분반의 평균 점수가 64명 분반의 평균점수보다 높을 확률은?
- 4) 두 분반의 평균 점수가 순서 상관 없이 76점과 83점이었다. 두 분반 중 83점일 가능성성이 더 높은 분반은 어디인가?

1) 25명 분반의 평균 시험 점수를 확률변수 \bar{X} 라고 하자. 그러면 \bar{X} 은 근사적으로 $N(77, \frac{15^2}{25})$ 을 따른다.

$$P(72 < \bar{X} < 82) = P\left(\frac{72 - 77}{\frac{15}{\sqrt{25}}} < \frac{\bar{X} - 77}{\frac{15}{\sqrt{25}}} < \frac{82 - 77}{\frac{15}{\sqrt{25}}}\right) \approx P(-1.67 < Z < 1.67) = 0.095$$

2) 64명 분반의 평균 시험 점수를 확률변수 \bar{Y} 라고 하자. 그러면 \bar{Y} 는 근사적으로 $N(77, \frac{15^2}{64})$ 을 따른다.

$$P(72 < \bar{Y} < 82) = P\left(\frac{72 - 77}{\frac{15}{\sqrt{64}}} < \frac{\bar{Y} - 77}{\frac{15}{\sqrt{64}}} < \frac{82 - 77}{\frac{15}{\sqrt{64}}}\right) \approx P(-2.67 < Z < 2.67) = 0.008$$

3) 두 분반의 평균 시험 점수는 분산만 다를 뿐 평균은 같으며, 평균을 기준으로 대칭인 확률분포를 이룬다. 따라서, 해당하는 확률은 0.5이다.

4) 평균인 77점으로부터 더욱 면 83점은 표본 수가 작을 때 나타나기 더욱 쉽다. 따라서, 25명 분반일 확률이 더욱 높다.

문제 6. 19. 어떤 나라의 풋살리그는 두 리그로 나누어져서 운영된다. 한 팀은 총 60게임을 치루며, 32 게임은 같은 그룹에 있는 팀과, 28게임은 다른 그룹에 있는 팀과 한다고 가정해보자. 설곽FC는 같은 그룹에 속한 팀에게는 0.5의 확률로, 다른 그룹에 속한 팀에게는 0.7의 확률로 승리한다. X 를 해당 시즌의 총 승리 수라고 하자.

- 1) X 는 이항분포를 따르는가?
 - 2) X_A 와 X_B 는 각각 같은 리그, 다른 리그와 경기한 경우의 승리 수라고 정의하자. X_A 와 X_B 의 분포는 어떠한가?
 - 3) X_A , X_B 와 X 의 관계는?
 - 4) 설곽FC가 40게임 이상 승리할 확률을 근사하라.
- 1) 시행마다 승리 확률이 달라지므로, X 는 이항분포를 따르지 않는다.
 - 2) X_A 와 X_B 는 각각 $B(32, 0.5)$, $B(28, 0.7)$ 을 따르는 확률변수이다. 즉, 이항분포를 따른다.
 - 3) $X = X_A + X_B$
 - 4) X_A , X_B 는 이항분포를 정규분포로 근사할 수 있을 고려하면 각각 $N(16, 8)$, $N(19.6, 5.88)$ 을 근사적으로 따른다. 서로 독립이다. 따라서 그 합인 X 는 근사적으로 $N(35.6, 13.88)$ 을 따르게 된다.

$$P(X \geq 40) \approx P(Z \geq \frac{40 - 35.6}{\sqrt{13.88}}) = P(Z \geq 1.18) = 0.119$$

문제 6. 20. 대형 시계를 구성하는 부품 중 톱니바퀴A는 구동에 매우 중요하여 고장나면 바로 바꿔 주어야 한다. 이 톱니바퀴의 평균 수명은 100시간이며 표준편차가 30시간이다. 2000시간 동안 0.95 이상의 확률로 구동시킬 수 있으려면, 최소 몇 개의 톱니바퀴A가 준비되어 있어야 하는가?

톱니바퀴가 n 개 준비되어 있다고 한다. 그러면, n 개의 톱니바퀴 구동 시간 합이 2000이 넘을 확률이 0.95보다 커야 한다. i 번째 톱니바퀴의 구동 시간 X_i 는 $E(X_i) = 100$ 이며 $V(X_i) = 900$ 이다. 그러면 중심극한정리에 의하여 \bar{X} 는 근사적으로 $N(100, \frac{900}{n})$ 을 따른다. 이것이 톱니바퀴가 평균적으로 버티는 시간인 $\frac{2000}{n}$ 보다 클 확률이 0.95보다는 커야 한다는 것이다. 즉,

$$P(\bar{X} > \frac{2000}{n}) = 0.95 = P(Z > -1.645)$$

라는 것이다. \bar{X} 를 표준화하면

$$P\left(\frac{\bar{X} - 100}{30/\sqrt{n}} > \frac{2000 - 100n}{30}\frac{1}{\sqrt{n}}\right)$$

이다. 따라서,

$$\frac{200 - 10n}{3\sqrt{n}} < -1.645$$

인 n 을 찾아주면 된다. 그런 최소의 n 은 23이다. 즉, 최소 23개의 톱니바퀴가 준비되어 있어야 한다.

문제 6. 21. X_1, X_2, X_3 은 평균이 12이고 분산이 9인 정규분포를 따르는 모집단에서 뽑힌 표본이다.

$$1) P(X_1 + X_2 + X_3 > 35) \text{ 는?}$$

$$2) P(\bar{X} > 14) \text{ 는?}$$

$$3) P(\min(X_1, X_2, X_3) < 9) \text{ 는?}$$

$$4) P(\max(X_1, X_2, X_3) < 15) \text{ 는?}$$

1) $P(\bar{X} > \frac{35}{3})$ 을 구해주면 된다. $\bar{X} \sim N(12, 3)$ 이므로,

$$P(X_1 + X_2 + X_3 > 35) = P\left(\frac{\bar{X} - 12}{\sqrt{3}} > \frac{\frac{35}{3} - 12}{\sqrt{3}}\right) = P(Z > -\frac{1}{3\sqrt{3}}) = 0.575$$

2)

$$P\left(\frac{\bar{X} - 12}{\sqrt{3}} > \frac{14 - 12}{\sqrt{3}}\right) = P(Z > \frac{2}{\sqrt{3}}) = 0.124$$

3)

$$P(\min(X_1, X_2, X_3) < 9) = 1 - P(X_1 \geq 9)P(X_2 \geq 9)P(X_3 \geq 9) = 1 - P(Z \geq -1) = 1 - (0.841)^3$$

4)

$$P(\max(X_1, X_2, X_3) < 15) = P(X_1 < 15)P(X_2 < 15)P(X_3 < 15) = P(Z < 1)^3 = (0.841)^3$$

문제 6. 22. X_1, X_2, \dots, X_{25} 는 평균이 3이고 분산이 100인 정규분포에서 뽑은 표본이다. $P(0 < \bar{X} < 4, 56.2 < S^2 < 164)$ 는? 단, 표준정규분포 Z 에서 $P(Z > z_\alpha) = \alpha$, 자유도가 k 인 카이제곱분포를 따르는 확률변수 X 에 대해 $P(X > \chi_\alpha^2(k)) = \alpha$ 라고 정의한다.

단, 표본평균과 표본분산은 독립이라고 가정한다.

$$P(0 < \bar{X} < 4, 56.2 < S^2 < 164) = P(0 < \bar{X} < 4)P(56.2 < S^2 < 164)$$

이다. 먼저 표본평균에 대해 구하여 보자. 원래의 모집단 평균이 3이고 분산이 100이므로,

$$\bar{X} \sim N(3, 4)$$

를 따름을 알 수 있다. 따라서,

$$P(0 < \bar{X} < 4) = P\left(\frac{0 - 3}{2} < \frac{\bar{X} - 3}{2} < \frac{4 - 3}{2}\right) = P(-1.5 < Z < 0.5) = 0.624$$

다음으로는 표본분산에 대하여 보자. 분산이 100이므로

$$\frac{24S^2}{100} \sim \chi^2(24)$$

를 따른다. 따라서,

$$P(56.2 < S^2 < 164) = P\left(\frac{24S^2}{100} < \frac{24S^2}{164} < \frac{24S^2}{56.2}\right) = P(13.488 < \frac{24S^2}{100} < 39.36)$$

따라서 $\chi^2_{\alpha}(24) = 39.36$ 인 α 와 $\chi^2_{\beta}(164) = 13.488$ 인 β 를 찾은 이후 $\beta - \alpha$ 를 해주면 된다. $\alpha = 0.025$, $\beta = 0.957$ 이므로 원하는 확률은 0.932이다. 따라서, 구하는 확률은 $0.624 \times 0.932 = 0.582$ 다.

문제 7. 1. $Z(1 - \alpha)$ 를 $Z(\alpha)$ 로써 표시하라.

표준정규분포는 0을 기준으로 대칭형의 분포를 가진다. 따라서, $Z(1 - \alpha) = -Z(\alpha)$ 이다.

문제 7. 2. $Z(\frac{\alpha}{2})$ 와 $Z(\alpha)$ 중 어느 것이 더 큰지 확인하라.

자신보다 큰 값일 확률이 더 작다는 것은 더 크다는 것을 의미한다. 따라서,

$$Z\left(\frac{\alpha}{2}\right) \geq Z(\alpha)$$

문제 7. 3. $t_{n,1-\alpha}$ 을 $t_{n,\alpha}$ 을 이용하여 표시하라.

t 분포는 0을 기준으로 대칭형의 분포를 가진다. 따라서, $t_{n,1-\alpha} = -t_{n,\alpha}$ 이다.

문제 7. 4. 가수 장혜진의 앨범에서 임의추출된 9개의 곡 길이가 평균 240초이고, 분산이 100이라고 한다. 곡 길이의 표준편차를 모른다고 할 때,

- 1) 곡 길이 평균의 점추정값은?
- 2) 곡 길이 평균의 95% 신뢰구간을 구하여라.

- 1) 곡 길이의 평균이라는 모평균은 표본평균이 점추정값이다. 따라서, 240초가 점추정값이다.
- 2) 95% 신뢰구간은

$$\left[240 - t_{8,0.025} \frac{10}{\sqrt{9}}, 240 + t_{8,0.025} \frac{10}{\sqrt{9}} \right] = [232.31, 247.69]$$

문제 7. 5. 동욱이는 자신이 하고 있는 게임에서 '기초통계학의 검'이라는 아이템을 뽑고 싶어한다. 동욱이가 100번 뽑았을 때, 기초통계학의 검은 30번 등장하였다. 기초통계학의 검 등장 확률의 95% 신뢰구간을 구하여라.

표본비율이 0.3이므로, 모비율 p 에 대하여 신뢰구간을 구하면

$$\left[0.3 - 1.96 \sqrt{\frac{0.3 \times 0.7}{100}}, 0.3 + 1.96 \sqrt{\frac{0.3 \times 0.7}{100}} \right] = [0.210, 0.390]$$

문제 7. 6. 걸그룹 여자친구의 곡별 BPM은 정규분포를 따른다고 한다. 25개의 활동곡을 골라 BPM을 계산했더니, 분산값이 0.32였다. 모표준편차 σ 의 95% 신뢰구간을 구하여라.

모분산 σ^2 에 대한 95% 신뢰구간을 구하면 아래와 같다.

$$\left[\frac{24 \cdot 0.32}{\chi_{24,0.025}^2}, \frac{24 \cdot 0.32}{\chi_{24,0.975}^2} \right] = [0.195, 0.619]$$

모표준편차 σ 에 대해서는 여기에 제곱근을 씌워주면 되므로, [0.442, 0.787]가 될 것이다.

문제 7. 7. T 가 자유도가 8인 t 분포를 따른다고 할 때, 표를 이용하여

- 1) $P(T \geq 1)$
 - 2) $P(T \leq 2)$
 - 3) $P(-1 < T < 1)$
의 값을 구하여라.
- 1) 0.173
 - 2) 0.96
 - 3) 0.653

문제 7. 8. 2013년 8월, 뉴욕타임즈는 오바마 대통령의 국정수행에 대한 긍정률이 50퍼센트라고 보도하였으며, 95퍼센트 신뢰도에서 오차는 $\pm 4\%$ 이하라고 주장하였다. 표본 크기의 최솟값을 구하여라.

모비율에 대한 추정에서 95퍼센트 신뢰구간을 만들 때, 오차는 $\pm 1.96 \sqrt{\frac{0.5 \times 0.5}{n}}$ 이다. 이 값의 절댓값이 0.04보다는 작아야 하므로,

$$1.96 \sqrt{\frac{0.5 \times 0.5}{n}} \leq 0.04$$

로부터

$$\frac{1}{4} \left(\frac{1.96}{0.04} \right)^2 \leq n$$

이다. 따라서, n 의 최솟값은 601이다.

문제 7. 9. 앞선 문제와 같이, 모비율의 구간추정에서 신뢰도가 $1 - \alpha$ 일 때 오차가 $\pm b\%$ 이하가 되기 위해 필요한 표본의 수를 결정하길 원한다. 표본비율에 대한 확신이 없이 계획을 수립하려 할 때, 표본 크기 n 은 최소 얼마여야 하는가?

$$Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{0.5 \times 0.5}{n}} < \frac{b}{100}$$

이어야 하므로, 이를 정리하면

$$\frac{1}{4} \left(\frac{Z(\alpha/2)}{b/100} \right)^2 \leq n$$

이다. 따라서,

$$2500 \left(\frac{Z(\alpha/2)}{b} \right)^2$$

이상의 표본 크기가 있어야 한다.

문제 7. 10.

$$P(Z \leq z_\alpha) = \alpha$$

임을 이용하여, 표본평균이 \bar{X} 이고 모분산이 σ^2 이라 알려져 있을 때 95%의 신뢰도로 모평균을 포함하되, 구간의 오른쪽 끝은 ∞ 인 구간을 만들고 싶다. 원하는 구간은 무엇이 되는가?

$$\begin{aligned} 0.95 &= P(Z < Z(0.05)) \\ &= P\left(\frac{\bar{X} - m}{\sigma/\sqrt{n}} < Z(0.05)\right) \\ &= P\left(\bar{X} - m < \frac{\sigma}{\sqrt{n}} Z(0.05)\right) \\ &= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} Z(0.05) < m < \infty\right) \end{aligned}$$

이므로 원하는 구간은

$$\left[\bar{X} - 1.645 \frac{\sigma}{\sqrt{n}}, \infty \right)$$

문제 7. 11. 위의 문제와 유사한 방법으로, 구간의 왼쪽 끝은 $-\infty$ 인 신뢰구간을 만들어라.

$$\begin{aligned}
0.95 &= P(Z > -Z(0.05)) \\
&= P\left(\frac{\bar{X} - m}{\sigma/\sqrt{n}} > -Z(0.05)\right) \\
&= P\left(\bar{X} - m > -\frac{\sigma}{\sqrt{n}}Z(0.05)\right) \\
&= P\left(\bar{X} + \frac{\sigma}{\sqrt{n}}Z(0.05) > m > -\infty\right)
\end{aligned}$$

이므로

$$\left(-\infty, \bar{X} + 1.645 \frac{\sigma}{n}\right]$$

문제 7. 12. 문제 7.10, 문제 7.11과 유사한 방법으로 모분산을 모를 때의 한쪽 끝의 절댓값이 ∞ 인 모평균의 신뢰구간을 만들어라.

모분산을 모를 경우에는

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

임을 알고 있으니 이를 이용하자.

$$\begin{aligned}
1 - \alpha &= P(T < t_{\alpha}(n-1)) \\
&= P\left(\bar{X} - m < \frac{S}{\sqrt{n}}t_{n-1,\alpha}\right) \\
&= P\left(\bar{X} - \frac{S}{\sqrt{n}}t_{n-1,\alpha} < m < \infty\right)
\end{aligned}$$

이므로, 오른쪽 끝이 무한한 구간은

$$\left[\bar{X} - \frac{S}{\sqrt{n}}t_{n-1,\alpha}, \infty\right)$$

동일한 이유로,

$$\begin{aligned}
1 - \alpha &= P(T > -t_{n-1,\alpha}) \\
&= P\left(\bar{X} - m > -\frac{S}{\sqrt{n}}t_{n-1,\alpha}\right) \\
&= P\left(\bar{X} + \frac{S}{\sqrt{n}}t_{n-1,\alpha} > m > -\infty\right)
\end{aligned}$$

이므로, 왼쪽 끝이 무한한 구간은

$$\left(-\infty, \bar{X} + \frac{S}{\sqrt{n}}t_{n-1,\alpha}\right]$$

문제 7. 13. 문제 7.10, 문제 7.11과 유사한 방법으로, 모비율의 신뢰구간을 만들어라.

표본평균의 표준편차 역할을 하는 σ/\sqrt{n} 대신에 모비율의 추정에서 사용되는 표준편차인

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

을 넣어주면 될 것이다.

즉, 각각은

$$\left[\bar{X} - Z(\alpha)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \infty\right)$$

과

$$\left(-\infty, \bar{X} + Z(\alpha) \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

문제 7. 14. 문제 7.10, 문제 7.11과 유사한 방법으로, 모분산의 신뢰구간을 만들어라.

확률변수 $U = \frac{(n-1)S^2}{\sigma^2}$ 는 $\chi^2(n-1)$ 을 따른다.

$$\begin{aligned} 1 - \alpha &= P(U > \chi_{n-1, 1-\alpha}^2) \\ &= P\left(\frac{(n-1)S^2}{\sigma^2} > \chi_{n-1, 1-\alpha}^2\right) \\ &= P\left(\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha}^2} > \sigma^2 > 0\right) \end{aligned}$$

이므로, 아래에 대한 제한이 없는 구간은

$$\left[0, \frac{(n-1)S^2}{\chi_{n-1, 1-\alpha}^2} \right]$$

이다. 반면

$$\begin{aligned} 1 - \alpha &= P(U < \chi_{n-1, \alpha}^2) \\ &= P\left(\frac{(n-1)S^2}{\sigma^2} < \chi_{n-1, \alpha}^2\right) \\ &= P\left(\frac{(n-1)S^2}{\chi_{n-1, \alpha}^2} < \sigma^2 < \infty\right) \end{aligned}$$

이므로, 위에 대한 제한이 없는 구간은

$$\left[\frac{(n-1)S^2}{\chi_{n-1, \alpha}^2}, \infty \right)$$

이다.

문제 7. 15. 모표준편차의 신뢰구간은 어떻게 구할지 이야기하고, 위의 문제와 같이 한 쪽으로는 가능한 데 까지 뻗어 있는 신뢰구간을 만들어라. 기준에 배운 신뢰구간을 양측신뢰구간, 한쪽이 뻗어 있는 신뢰구간을 단측신뢰구간이라 부르기로 하자.

모표준편차에 대해서는 모분산의 구간에 제곱근을 씌우면 된다. 즉, 각각

$$\left[0, \sqrt{\frac{(n-1)S^2}{\chi_{n-1, 1-\alpha}^2}} \right]$$

$$\left[\sqrt{\frac{(n-1)S^2}{\chi_{n-1, \alpha}^2}}, \infty \right)$$

이다.

문제 7. 16. 저울이 보여주는 값은 사실 원래 물체의 무게와 비교하여, 오차가 포함된 값이다. 오차는 정규 분포를 따르며, 평균은 0이고 표준편차가 $0.1mg$ 이라고 한다. 동일한 다섯 개의 물체를 저울에 올려본 결과, $3.142mg$, $3.163mg$, $3.155mg$, $3.150mg$, $3.141mg$ 이 나왔다. 실제 물체의 무게에 대한 95퍼센트 신뢰구간을 구하라.

저울에 쳐히는 무게는 원래 물체의 무게 m 에 대하여 $N(m, 0.01)$ 을 따른다. 따라서 $n = 5$, $\bar{X} = 3.150$, $\sigma = 0.1$ 을 대입하면,

$$[3.062, 3.238]$$

이 실제 물체의 무게 m 에 대한 95퍼센트 신뢰구간이다.

문제 7. 17. 랜덤으로 뽑힌 학생 81명에 대해 기초통계학 점수의 표본평균이 74.6이고 표본표준편차가 11.3이었을 때, 전체 모평균에 대한 90퍼센트 신뢰구간을 구하여라.

신뢰구간은

$$\left[74.6 - \frac{11.3}{\sqrt{81}} t_{80,0.05}, 74.6 + \frac{11.3}{\sqrt{81}} t_{80,0.05} \right] = [72.1, 77.1]$$

문제 7. 18. 알지 못하는 평균 μ 와 분산 1을 가지는 정규분포를 따르는 모집단에서 표본 X_1, X_2, \dots, X_{n+1} 을 뽑아냈다. 처음 n 개의 표본평균을 \bar{X}_n 이라고 정의하자. 서로 독립인 정규분포는 더해도 독립이며, 평균과 분산은 둘의 합과 같다.

1) $X_{n+1} - X_n$ 의 분포는?

2) $\bar{X}_n = 4$ 일 때, X_{n+1} 값의 95퍼센트 신뢰구간을 구하여라.

1) 빼지는 두 분포는 서로 독립이며, 동일한 분포 $N(\mu, 1)$ 을 따르고 있다. 따라서, 서로 빼면 그 분포는 $N(0, 2)$ 을 따른다.

2) \bar{X}_n 은 표본평균의 분포이므로, $N(\mu, \frac{\sigma^2}{n})$ 을 따른다. 그러면, $\bar{X}_n - X_{n+1}$ 은 서로 독립인 정규분포를 뺀 것이므로 평균은 그 평균의 차인 0이며 분산은 둘의 분산의 합인 $(1 + 1/n)\sigma^2$ 인 정규분포이다. 따라서, 표준화를 해주면

$$\frac{\bar{X}_n - X_{n+1}}{\sigma \sqrt{1 + \frac{1}{n}}}$$

가 표준정규분포를 따른다. 그러면

$$\begin{aligned} 0.95 &= P(-1.96 < Z < 1.96) \\ &= P(-1.96 < \frac{\bar{X}_n - X_{n+1}}{\sigma \sqrt{1 + \frac{1}{n}}} < 1.96) \\ &= P(-1.96\sigma\sqrt{1 + \frac{1}{n}} < \bar{X}_n - X_{n+1} < 1.96\sigma\sqrt{1 + \frac{1}{n}}) \\ &= P(\bar{X}_n - 1.96\sigma\sqrt{1 + \frac{1}{n}} < X_{n+1} < \bar{X}_n + 1.96\sigma\sqrt{1 + \frac{1}{n}}) \end{aligned}$$

이므로, 신뢰구간은

$$\left[\bar{X}_n - 1.96\sigma\sqrt{1 + \frac{1}{n}}, \bar{X}_n + 1.96\sigma\sqrt{1 + \frac{1}{n}} \right]$$

이다. 따라서 n 을 제외한 아는 값들을 모두 대입하면

$$\left[4 - 1.96\sqrt{1 + \frac{1}{n}}, 4 + 1.96\sqrt{1 + \frac{1}{n}} \right]$$

문제 7. 19. U_1, U_2, \dots 는 서로 독립이고 $U(0, 1)$ 을 따르는 확률변수들이라고 하자. 그 다음, N 이라는 확률변수를

$$N = \min(n : U_1 + \dots + U_n > 1)$$

이라고 정의하자. 계산을 통해 $E[N]$ 의 값을 계산하여라.

추가하여, 계산이 맞는지 확인하기 위해 N 에서 표본을 100개 뽑아 95퍼센트 신뢰구간을 구하고자 한다. N 의 모표준편차는 알 수 없지만, N 은 근사적으로 정규분포를 따른다는 사실을 이용할 수는 있다. 어떤 방식을 사용해 신뢰구간을 구해야 하는지 이야기하여라. 결과적으로 표본평균은 \bar{N} , 표본분산은 S^2 으로 둔다.

먼저 계산을 통해 이 값을 구해 보자.

$$N_x = \min(n : U_1 + \cdots + U_n > x)$$

라고 하자. 또한, $m_x = E[N_x]$ 라고 두자. 그러면 구하는 값은 m_1 이 된다.

만약 $U_1 = u$ 라는 값을 얻었다고 하자. 그러면 만약 $u \geq x$ 일 경우 $N_x = 1$ 이며, $u \leq x$ 일 경우 $x - u$ 가 새로운 x 가 되며 필요한 U 의 개수는 하나 추가되는 것이다. 이를 식으로 표현하면

$$m_x = P(U_1 > x) + \int_0^x 1 \cdot (m_{x-u} + 1) du = (1-x) + x + \int_0^x m_{x-u} du = 1 + \int_0^x m_y dy$$

이다. 양변을 미분하게 되면, $m'_x = m_x$ 형태이며 이에 따라 $m_x = ae^x$ 꼴이다. 그런데 $m_0 = 1$ 므로, $a = 1$ 이다. 따라서, $m_1 = 1e^1 = e$ 가 된다. 즉, $E[N] = e$.

N 은 근사적으로 정규분포를 따르고 표본의 크기는 100이다. 모표준편차를 모르므로, 표본표준편차를 이용해야 한다. 따라서 그 신뢰구간은

$$\left[\bar{N} - t_{99,0.025} \frac{S}{10}, \bar{N} + t_{99,0.025} \frac{S}{10} \right]$$

문제 7. 20. 어떤 과학자는 성인의 나트륨 섭취에 관하여 연구를 하고 있다. 남성 9명을 골라 설문한 결과, 그들의 하루 평균 나트륨 섭취량은 1560그램이었으며 표준편차는 33그램이었다. 이를 바탕으로, 성인 남성의 하루 평균 나트륨 섭취량의 95퍼센트 신뢰구간을 완성하여라.

모표준편차를 모르는 상황에서 표본표준편차를 이용하여 표본평균을 이용한 모평균의 구간추정을 시행하고 있다. 따라서 그 신뢰구간은

$$\left[1560 - t_{8,0.025} \frac{33}{\sqrt{9}}, 1560 + t_{8,0.025} \frac{33}{\sqrt{9}} \right] = [1534.6, 1585.4]$$

가 된다.

문제 7. 21. 효범이는 동급생 81명에게 물어 자신의 노래점수를 평가해달라고 했다. 그 결과, 표본평균은 95점이었으며 표본분산은 5로 나타났다.

1) 모분산의 점추정량은?

2) 모분산에 대한 95퍼센트 신뢰구간을 만들어라. 단, $\chi^2_{80,0.025}$ 등을 계산할 필요는 없다.

1) 모분산의 점추정량은 표본분산이다. 즉, 5이다.

2)

$$\left[\frac{80 \times 5}{\chi^2_{80,0.025}}, \frac{80 \times 5}{\chi^2_{80,0.975}} \right] = [3.75, 7]$$

문제 8. 1. 당근에 들어 있는 카로틴 성분이 눈에 좋다는 것을 보이고자 한다. 사람들을 모아 당근 섭취량에 따른 야맹증 발병률을 비교함으로써 이를 보이고자 한다. 귀무가설과 대립가설은 각각 무엇인가?

귀무가설 (H_0) : 당근 섭취량에 따른 야맹증 발병률 차이는 없다. 즉, 당근의 섭취량과 야맹증의 발병 가능성 사이에는 관계가 없다.

대립가설 (H_1) : 당근 섭취량에 따라 야맹증 발병률에는 차이가 있다.

문제 8. 2.

$$H_0 : m \geq m_0$$

일 때 H_1 을 쓰고, 채택역과 기각역을 써라.

$$H_1 : m < m_0$$

m_0 이 예상하는 m_0 보다 충분히 작을 때 우리는 귀무가설을 기각할 수 있다. 즉,

$$P(-Z(\alpha) \leq Z < \infty) = 1 - \alpha$$

인 것으로부터

$$\left[m_0 - \frac{\sigma}{\sqrt{n}} Z(\alpha), \infty \right)$$

가 채택역임을 알 수 있으며 기각역은 실수에서 이를 제외한

$$\left(-\infty, m_0 - \frac{\sigma}{\sqrt{n}} Z(\alpha) \right)$$

이 된다.

문제 8. 3. 작년의 자영업자 월세는 모평균 200만원, 모표준편차 20만원인 정규분포로 나타났다. 올해 자영업자 100명에게 월세를 조사한 결과, 평균이 205만원이었다.

- 1) 올해 평균 월세가 올랐는지 유의수준 0.01에서 검정하라.
- 2) 올해 평균 월세가 변했는지 유의수준 0.01에서 검정하라.

1) 올해 자영업자 월세의 모평균을 μ 라고 하자. 또한, 작년과 모표준편차는 같다고 가정하자. 그러면, 귀무가설 H_0 은 $\mu \leq 200$, 대립가설 H_1 은 $\mu > 200$ 로 나타난다. 이 경우 가설의 기각역은

$$\bar{X} > 200 + Z(0.01) \frac{20}{\sqrt{100}} = 204.654$$

일 때이다. 따라서 \bar{X} 의 값이 205로 나타났으니 이는 기각역에 포함된다. 따라서, 귀무가설을 기각하고 올해의 평균 월세가 올랐다고 주장할 수 있다.

2) 이 경우에는 $H_0 : \mu = 200$, $H_1 : \mu \neq 200$ 이 원하는 귀무가설과 대립가설이다. 가설의 기각역은

$$\bar{X} > 200 + Z(0.005) \frac{20}{\sqrt{100}} \quad \text{or} \quad \bar{X} < 200 - Z(0.005) \frac{20}{\sqrt{100}}$$

이다. 그러나 205는 채택역인 [194.92, 205.08]에 포함되므로 귀무가설을 기각할 만한 충분한 근거가 없다. 따라서, 올해 평균 월세가 변했다고 주장할 수 없다.

- 1) $H_0 : \sigma^2 \leq \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$ 일 때, $S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha}^2}{n-1}$ 이면 H_0 기각
- 2) $H_0 : \sigma^2 \geq \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$ 일 때, $S^2 < \frac{\sigma_0^2 \chi_{n-1,1-\alpha}^2}{n-1}$ 이면 H_0 기각
- 3) $H_0 : \sigma^2 = \sigma_0^2$, $H_1 : \sigma^2 \neq \sigma_0^2$ 일 때, $S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha/2}^2}{n-1}$ 이거나 $S^2 < \frac{\sigma_0^2 \chi_{n-1,1-\alpha/2}^2}{n-1}$ 면 H_0 기각

문제 8. 4. 위의 박스를 증명해 보아라.

1)

$$P\left(0 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,\alpha}^2\right) = 1 - \alpha$$

이면, 귀무가설 하에서는 $\sigma = \sigma_0$ 이다. 따라서,

$$\begin{aligned} 1 - \alpha &= P\left(0 \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi_{n-1,\alpha}^2\right) \\ &= P\left(0 \leq S^2 \leq \frac{\sigma_0^2 \chi_{n-1,\alpha}^2}{(n-1)}\right) \end{aligned}$$

이므로 S^2 가 0부터 $\sigma_0^2 \chi_{n-1,\alpha}^2 / (n-1)$ 내부에 들어올 확률이 $1 - \alpha$ 인 경우,

$$S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha}^2}{(n-1)}$$

일 확률이 α 보다 작다는 것이다. 즉, 유의수준 α 에서 위의 부등식이 성립하면 귀무가설이 맞다는 가정 하에 해당 표본분산이 관찰될 확률이 매우 낮다는 것이다. 따라서, 기각역은

$$S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha}^2}{(n-1)}$$

이다.

2)

$$P(\chi_{n-1,1-\alpha}^2 \leq \frac{(n-1)S^2}{\sigma_0^2} < \infty) = 1 - \alpha$$

이므로, 위와 같은 식으로 정리하게 된다면

$$P\left(\frac{\sigma_0^2 \chi_{n-1,1-\alpha}^2}{n-1} \leq S^2 < \infty\right) = 1 - \alpha$$

이다. 따라서, 유의수준 α 에서

$$S^2 < \frac{\sigma_0^2 \chi_{n-1,1-\alpha}^2}{n-1}$$

일 경우 기각할 수 있다.

3)

$$P(\chi_{n-1,1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}^2) = 1 - \alpha$$

이므로,

$$P\left(\frac{\sigma_0^2 \chi_{n-1,1-\alpha/2}^2}{n-1} \leq S^2 \leq \frac{\sigma_0^2 \chi_{n-1,\alpha/2}^2}{n-1}\right) = 1 - \alpha$$

으로부터

$$S^2 < \frac{\sigma_0^2 \chi_{n-1,1-\alpha/2}^2}{n-1}$$

이거나

$$S^2 > \frac{\sigma_0^2 \chi_{n-1,\alpha/2}^2}{n-1}$$

일 경우 기각해야 함을 안다.

\bar{X}_1 과 \bar{X}_2 에 대하여, 이들이 각각 $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$ 를 따르는 서로 독립인 정규분포에서 추출된 크기가 n_1, n_2 인 표본의 표본평균이라 하면

$$\bar{X}_1 - \bar{X}_2 \sim N(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

이다.

문제 8. 5. 위의 박스를 증명하여라. 서로 독립인 정규분포의 합과 차는 역시 정규분포임을 알고 있다고 가정하자.

서로 독립인 정규분포의 차는 정규분포임을 알고 있다. 그러면

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = m_1 - m_2$$

이며,

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

이 될 것이다. 이때, 표본평균의 평균과 분산에 대한 정보를 이용하였다. 따라서, $\bar{X}_1 - \bar{X}_2$ 는

$$N(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

을 따른다.

- 1) $H_0 : m_1 - m_2 \leq D_0, H_1 : m_1 - m_2 > D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이면 H_0 기각
- 2) $H_0 : m_1 - m_2 \geq D_0, H_1 : m_1 - m_2 < D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이면 H_0 기각
- 3) $H_0 : m_1 - m_2 = D_0, H_1 : m_1 - m_2 \neq D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha/2) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이거나,
 $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha/2) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 이면 H_0 기각

문제 8. 6. 위를 증명하여라.

1) $m_1 - m_2 = D_0$ 라는 가정 하에 검정을 수행하므로, 귀무가설 하에서 $\bar{X}_1 - \bar{X}_2 \sim N(D_0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ 를 따른다. 그러면

$$\begin{aligned} 1 - \alpha &= P \left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z(\alpha) \right) \\ &= P \left((\bar{X}_1 - \bar{X}_2) - D_0 \leq Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \\ &= P \left(\bar{X}_1 - \bar{X}_2 \leq D_0 + Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 > D_0 + Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

일 확률은 매우 적다. 따라서, 이 범위일 때는 귀무가설을 기각할 수 있다.

2)

$$\begin{aligned} 1 - \alpha &= P \left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq -Z(\alpha) \right) \\ &= P \left((\bar{X}_1 - \bar{X}_2) - D_0 \geq -Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \\ &= P \left(\bar{X}_1 - \bar{X}_2 \geq D_0 - Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 < D_0 - Z(\alpha) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

이 기각역이 될 것이다.

3)

$$\begin{aligned} 1 - \alpha &= P \left(-Z\left(\frac{\alpha}{2}\right) \leq \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z\left(\frac{\alpha}{2}\right) \right) \\ &= P \left(-Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\bar{X}_1 - \bar{X}_2) - D_0 \leq Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \\ &= P \left(D_0 - Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \bar{X}_1 - \bar{X}_2 \leq D_0 + Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 < D_0 - Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

와

$$\bar{X}_1 - \bar{X}_2 > D_0 + Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

이 기각역이다.

- 1) $H_0 : m_1 - m_2 \leq D_0, H_1 : m_1 - m_2 > D_0$ 일 경우 $(\bar{X}_1 - \bar{X}_2) > D_0 + t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 면 H_0 기각
- 2) $H_0 : m_1 - m_2 \geq D_0, H_1 : m_1 - m_2 < D_0$ 일 경우 $(\bar{X}_1 - \bar{X}_2) < D_0 - t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 면 H_0 기각
- 3) $H_0 : m_1 - m_2 = D_0, H_1 : m_1 - m_2 \neq D_0$ 일 경우 $(\bar{X}_1 - \bar{X}_2) > D_0 + t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
이거나 $(\bar{X}_1 - \bar{X}_2) < D_0 - t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 일면 H_0 기각

문제 8. 7. 위 박스를 증명하여라.

1)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

임을 이용해주면 된다.

$$\begin{aligned} 1 - \alpha &= P \left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2,\alpha} \right) \\ &= P \left((\bar{X}_1 - \bar{X}_2) - D_0 \leq t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \\ &= P \left(\bar{X}_1 - \bar{X}_2 \leq D_0 + t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \end{aligned}$$

임을 고려하면 기각역은

$$\bar{X}_1 - \bar{X}_2 > D_0 + t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

이다.

2)

$$\begin{aligned} 1 - \alpha &= P \left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq -t_{n_1+n_2-2,\alpha} \right) \\ &= P \left((\bar{X}_1 - \bar{X}_2) - D_0 \geq -t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \\ &= P \left(\bar{X}_1 - \bar{X}_2 \geq D_0 - t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 < D_0 - t_{n_1+n_2-2,\alpha} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

이 기각역이 될 것이다.

3)

$$\begin{aligned}
1 - \alpha &= P \left(-t_{n_1+n_2-2,\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2,\alpha/2} \right) \\
&= P \left(-t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\bar{X}_1 - \bar{X}_2) - D_0 \leq t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \\
&= P \left(D_0 - t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \bar{X}_1 - \bar{X}_2 \leq D_0 + t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)
\end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 < D_0 - t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

와

$$\bar{X}_1 - \bar{X}_2 > D_0 + t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

가 기각역이다.

- 1) $H_0 : m_1 - m_2 \leq D_0, H_1 : m_1 - m_2 > D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이면 H_0 기각
- 2) $H_0 : m_1 - m_2 \geq D_0, H_1 : m_1 - m_2 < D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이면 H_0 기각
- 3) $H_0 : m_1 - m_2 = D_0, H_1 : m_1 - m_2 \neq D_0$ 일 때 $(\bar{X}_1 - \bar{X}_2) > D_0 + Z(\alpha/2) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이거나,
 $(\bar{X}_1 - \bar{X}_2) < D_0 - Z(\alpha/2) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ 이면 H_0 기각

임을 알 수 있다.

문제 8. 8. 위를 증명하여라.

- 1) $m_1 - m_2 = D_0$ 라는 가정 하에 검정을 수행하므로, 귀무가설 하에서 $\bar{X}_1 - \bar{X}_2$ 는 $N(D_0, \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})$ 를 따른다. 그러면

$$\begin{aligned}
1 - \alpha &= P \left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq Z(\alpha) \right) \\
&= P \left((\bar{X}_1 - \bar{X}_2) - D_0 \leq Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) \\
&= P \left(\bar{X}_1 - \bar{X}_2 \leq D_0 + Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)
\end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 > D_0 + Z(\alpha) \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

일 확률은 매우 적다. 따라서, 이 범위일 때는 귀무가설을 기각할 수 있다.

2)

$$\begin{aligned}
 1 - \alpha &= P\left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \geq -Z(\alpha)\right) \\
 &= P\left((\bar{X}_1 - \bar{X}_2) - D_0 \geq -Z(\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \\
 &= P\left(\bar{X}_1 - \bar{X}_2 \geq D_0 - Z(\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)
 \end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 < D_0 - Z(\alpha)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

이 기각역이 될 것이다.

3)

$$\begin{aligned}
 1 - \alpha &= P\left(-Z\left(\frac{\alpha}{2}\right) \leq \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq Z\left(\frac{\alpha}{2}\right)\right) \\
 &= P\left(-Z\left(\frac{\alpha}{2}\right)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq (\bar{X}_1 - \bar{X}_2) - D_0 \leq Z\left(\frac{\alpha}{2}\right)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \\
 &= P\left(D_0 - Z\left(\frac{\alpha}{2}\right)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \bar{X}_1 - \bar{X}_2 \leq D_0 + Z\left(\frac{\alpha}{2}\right)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)
 \end{aligned}$$

이므로,

$$\bar{X}_1 - \bar{X}_2 < D_0 - Z\left(\frac{\alpha}{2}\right)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

와

$$\bar{X}_1 - \bar{X}_2 > D_0 + Z\left(\frac{\alpha}{2}\right)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

이 기각역이다.

문제 8. 9. $F_{df_1, df_2, 1-\alpha}$ 을 $F_{df_2, df_1, \alpha}$ 로 표현하라.

자유도가 df_1, df_2 인 F분포를 따르는 확률변수 F를 생각하자. 그러면

$$1 - \alpha = P(F > F_{df_1, df_2, 1-\alpha})$$

이다. 그런데, $\frac{1}{F}$ 는 F의 역수이므로 자유도가 df_2, df_1 인 F분포이다. 따라서

$$P(F > F_{df_1, df_2, 1-\alpha}) = P\left(\frac{1}{F_{df_1, df_2, 1-\alpha}} > \frac{1}{F}\right) = 1 - \alpha$$

이므로

$$\alpha = P\left(\frac{1}{F_{df_1, df_2, 1-\alpha}} \leq \frac{1}{F}\right) = P(F_{df_2, df_1, \alpha} \leq \frac{1}{F})$$

이기애

$$F_{df_1, df_2, 1-\alpha} = \frac{1}{F_{df_2, df_1, \alpha}}$$

이다.

$H_0 : \sigma_1^2 = \sigma_2^2, H_1 : \sigma_1^2 \neq \sigma_2^2$ 일 때,

$$\frac{\text{큰 표본분산}}{\text{작은 표본분산}} > F_{n_1-1, n_2-1, \alpha/2}$$

이면 H_0 을 기각한다.

문제 8. 10. 위의 박스를 보여라.

귀무가설 하에서 두 집단의 모분산이 같으므로,

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

이다. 이때, 가정에 의하여 S_1^2 이 S_2^2 보다 큰 것으로 생각한다.

$$1 - \alpha = P(F_{n_1-1, n_2-1, 1-\alpha/2} < \frac{S_1^2}{S_2^2} < F_{n_1-1, n_2-1, \alpha/2})$$

인데, 큰 표본분산을 위로 올렸으므로 채택역은

$$1 \leq \frac{S_1^2}{S_2^2} \leq F_{n_1-1, n_2-1, \alpha/2}$$

이다. 따라서, 기각역은

$$\frac{S_1^2}{S_2^2} > F_{n_1-1, n_2-1, \alpha/2}$$

이 될 것이다.

문제 8. 11. T_n 이 자유도가 n 인 t 분포를 따를 때, T_n^2 은 문자의 자유도가 1, 분모의 자유도가 n 인 F 분포임을 보여라.

T_n 은 자유도가 n 인 t 분포를 따르므로, 표준정규분포를 따르는 확률변수 Z 와 이와 독립인 자유도가 n 인 카이제곱분포를 따르는 확률변수 X 에 대하여

$$T_n = \frac{Z}{\sqrt{X/(n)}}$$

으로 표현된다. 그러면 양변에 제곱을 취하고 Z^2 이 $\chi^2(1)$ 을 따름을 감안하면

$$T_n^2 = \frac{Z^2/1}{X/n} = F(1, n)$$

임을 확인할 수 있다.

문제 8. 12. 두 개의 독립표본이 있으며, 그들은 모두 정규분포를 따르는 모집단으로부터 왔다. 그리고 그

모집단은 모두 같은 표준편차 σ 를 가지고 있다고 한다. 자료가

16, 17, 19, 20, 18

과

3, 4, 8

일 때, σ 의 점추정량을 구하라.

σ^2 이라는 두 집단의 공통된 모분산은 공통분산

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

로서 추정된다. 첫 자료의 표본평균은 18, 표본분산은 2.5이며 둘째 자료의 표본평균은 5, 표본분산은 7이다. 따라서

$$S_p^2 = \frac{10 + 14}{6} = 4$$

이므로, 표준편차의 점추정량은 $S_p = 2$ 이다.

문제 8. 13. 만약 두 표본이 서로 독립적이지 않은 상태에서 둘의 평균을 비교하고 싶으면 어떻게 할까? 예를 들어, 학생들의 영어 말하기 점수와 듣기 점수 비교를 통해 말하기 선생님과 듣기 선생님 중 어느 분이 더 잘 가르치시는지 알고 싶다. 따라서 한 반 26명의 학생들에게 수업 이후 시험을 보게 해 성취도를 평가하였다. 근데 문제는 각 학생의 말하기 점수와 듣기 점수는 독립이 아니라는 점이다. 공부를 잘하는 학생이 둘 다 잘 봤을 확률이 높다. 이런 경우에는 대응비교를 통해 수행해낸다.

1) 말하기 점수의 모평균을 m_1 , 듣기 점수의 모평균을 m_2 라고 하며, 말하기 점수의 표본평균이 \bar{X}_1 , 듣기 점수의 표본평균이 \bar{X}_2 라고 하자. 앞서 배운 것들을 이용하여, $\bar{X}_1 - \bar{X}_2$ 의 평균을 구하여라.

2) 모분산을 모를 때의 단일표본 가설검정을 이용하여, 적절한 가설을 세우고 신뢰수준이 α 일 때의 채택역을 구하여라.

1) $m_1 - m_2$ 가 그 평균이 된다.

2) 말하기 선생님과 듣기 선생님의 가르치는 효율이 같다는 가정이 귀무가설에서 적용된다. 즉, 귀무가설은 $m_1 - m_2 = 0$ 이며, 대립가설은 $m_1 - m_2 \neq 0$ 이다. 두 표본평균의 차는 평균이 $m_1 - m_2$ 이다. 또한 그 차의 모분산은 알려져 있지 않지만, σ^2 이라 두면 $\bar{X}_1 - \bar{X}_2$ 는 그 차의 표본평균이므로 귀무가설 하에서

$$\bar{X}_1 - \bar{X}_2 \sim N(0, \frac{\sigma^2}{n})$$

임을 확인할 수 있다. 즉, 이는 모분산을 모를 때의 단일표본 가설검정을 이용할 수 있으며

$$\frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}} \sim t(n-1)$$

임을 고려할 수 있다. 신뢰수준이 α 일 때의 채택역은

$$-t_{25,\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{26}} \leq t_{25,\alpha/2}$$

이며, 이를 정리하면

$$-t_{25,\alpha/2} \frac{S}{\sqrt{26}} \leq \bar{X}_1 - \bar{X}_2 \leq t_{25,\alpha/2} \frac{S}{\sqrt{26}}$$

이다.

문제 8. 14. 생산 라인의 두 기계로부터 물품이 나오는데, 각 기계에서 생산되는 물품의 무게는 독립표본이라고 하자. 첫째 기계에서는 36개의 물품에 대해 표본평균이 120그램, 표본분산이 4였다. 두번째 기계에서는 64개의 물품에 대해 표본평균이 130그램, 표본분산이 5였다. 첫째 기계에서 생산되는 물품의 무게는 $N(\mu_1, \sigma^2)$ 을 따르며 둘째 기계에서 생산되는 물품의 무게는 $N(\mu_2, \sigma^2)$ 을 따른다고 가정하자. $\mu_1 - \mu_2$ 의 95% 신뢰구간을 구하여라.

$\bar{X}_1 - \bar{X}_2$ 는 평균이 $\mu_1 - \mu_2$ 이고, 분산이 $\sigma^2(1/n_1 + 1/n_2)$ 인 정규분포를 따른다. 그런데 σ 를 모르는 상황이므로, S_p 를 이용하여 이를 t분포로 변형시켜주어야 한다. 따라서,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

임을 사용해줄 수 있다. 그러면 이를 잘 변형시킬 경우

$$\begin{aligned} 0.95 &= P \left(-t_{98,0.025} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{98,0.025} \right) \\ &= P \left(-t_{98,0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \leq t_{98,0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \\ &= P \left((\bar{X}_1 - \bar{X}_2) - t_{98,0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{98,0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \end{aligned}$$

이다. 따라서 95% 신뢰구간은

$$\left[(\bar{X}_1 - \bar{X}_2) - t_{98,0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{X}_1 - \bar{X}_2) + t_{98,0.025} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

이다. 이때 $S_p^2 = 4.64$ 로 계산되고 나머지를 모두 대입해 계산해주면,

$$[-10.89, -9.11]$$

이다.

문제 8. 15. 문제 8.14를 모분산이 각각 4, 5인 것으로 바꾸어서 풀어보아라.

모분산이 각각 4, 5일 경우에는

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{4}{36} + \frac{5}{64}}} \sim Z$$

임을 이용해줄 수 있다. 따라서,

$$\begin{aligned} 0.95 &= P \left(-1.96 \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{4}{36} + \frac{5}{64}}} \leq 1.96 \right) \\ &= P((\bar{X}_1 - \bar{X}_2) - 0.85 \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + 0.85) \end{aligned}$$

이다. 따라서 95% 신뢰구간은

$$[-10.85, -9.15]$$

이다.

문제 8. 16. 시장에서 유통되는 모든 담배는 현재 평균 니코틴 함유량이 최소 $1.6mg$ 이다. 그런데 어떤 담배회사가 담뱃잎을 새롭게 가공하는 법을 개발해내 그 방법으로 담배를 생산할 시 담배의 평균 니코틴 함유량이 $1.6mg$ 보다 작아질 수 있다고 주장하였다. 담배공사는 20개의 해당 담배를 얻어내 분석하였다. 일반적으로, 담배의 니코틴 함유량 모분산은 0.8임 이 알려져 있다. 표본평균이 $1.54mg$ 이었을 때, 유의수준 5% 에서 이 회사의 주장이 맞다고 이야기할 수 있는가?

귀무가설 H_0 : 새로운 가공 방법을 통해 얻은 담배의 평균 니코틴 함량은 $1.6mg$ 보다 높거나 같다.

대립가설 H_1 : 새로운 가공 방법을 통해 얻은 담배의 평균 니코틴 함량은 $1.6mg$ 보다 작다.

이 경우의 기각역은 $\bar{X} < 1.6 - 1.645 \frac{\sqrt{0.8}}{\sqrt{20}} = 1.271$ 이다. 그러나 얻은 표본평균은 여기에 포함되지 않고, 채택역에 포함된다. 따라서 귀무가설을 기각할 수 없기에 이 회사의 주장이 맞다고 이야기할 충분한 근거가 없다.

문제 8. 17. 병원 환자들의 혈중 콜레스테롤 농도를 줄일 수 있는 약을 새롭게 개발하였다. 50명의 환자들에 대해 이 약을 1달 동안 투여한 이후, 혈중 콜레스테롤 농도의 감소를 확인하였다. 표본평균은 $14.8mg/ml$, 모표준편차는 $6.4mg/ml$ 로 나타났다. 신약의 효과가 있다고 이야기할 수 있는가? 유의수준 0.05 에서 검정하라.

귀무가설 H_0 : 약을 투여한 환자들에게서 평균 혈중 콜레스테롤 농도가 같거나 높아진다.

대립가설 H_1 : 약을 투여한 환자들에게서 평균 혈중 콜레스테롤 농도가 감소한다.

귀무가설 하에서 감소량의 평균은 0이어야 한다. 이 경우의 기각역은 $\bar{X} > 0 + 1.645 \frac{6.4}{\sqrt{50}} = 1.489$ 인데, 표본평균은 이것보다 훨씬 큰 값인 14.8 이다. 따라서, 이 경우에는 귀무가설을 기각할 충분한 근거가 있다. 따라서 대립가설을 채택하고 신약의 효과가 있다고 주장할 수 있다.

문제 8. 18. 뼈째로 기계는 뼈째로에 묻혀진 초콜릿의 길이를 조절한다. 만약 초콜릿 길이의 표준편차 σ 가 $0.15cm$ 보다 작다면, 이 기계는 효과적이라고 판정할 수 있다. 20개의 뼈째로를 뽑아 표본분산을 확인한 결과, 0.025제곱센치미터 였다. 우리는 이 기계가 비효과적이라고 이야기할 수 있는가? 유의수준 0.05 에서 검정하라.

귀무가설 H_0 : 뼈째로에 묻혀진 초콜릿의 길이의 표준편차 편 0.15 보다 작거나 같다.

대립가설 H_1 : 뼈째로에 묻혀진 초콜릿의 길이의 표준편차는 0.15 보다 크다.

이 경우 기각역은 $S^2 > \frac{0.0225\chi^2_{19,0.05}}{19} = 0.0357$ 이다. 그러나 표본분산은 0.025 로 기각역에 포함되지 않는다. 따라서 귀무가설을 기각할 만한 충분한 근거가 존재하지 않으므로, 기계가 비효과적이라 이야기할 수 없다.

문제 8. 19. 어떤 효소의 활성을 막는 비가역성 억제제에는 *suicide inhibitor*와 *transition state analog*가 있다. 어떤 억제제를 사용해도 효소의 활성에서 분산 차이가 없다는 것을 밝히기 위해, 첫째 억제제에 대해서는 10번, 둘째 억제제에 대해서는 12번의 실험을 수행하였다. 그 결과, 표본분산은 각각 0.14 와 0.28 로 나타났다. 유의수준 0.05 에서 우리는 둘이 다른 분산을 가진다고 이야기할 수 있는가?

귀무가설 H_0 : 사용하는 비가역성 억제제의 종류에 상관없이 효소 활성도의 분산 차이가 없다.

대립가설 H_1 : 사용하는 비가역성 억제제의 종류에 따라 효소 활성도의 분산 차이가 존재한다.

검정은 $0.28/0.14 = 2$ 의 값이 기각역에 포함되는지 보면 된다.

$$\frac{S_1^2}{S_2^2} > F_{11,9,0.025} = 3.912$$

이 기각역인데, 이는 실제 표본분산의 비인 2를 포함하지 않는다. 따라서, 채택역에 포함되므로 귀무가설을 기각할 만한 충분한 근거가 없기에 우리는 둘이 다른 분산을 가진다고 주장할 수 없다.

문제 8. 20. 실제 모평균은 10인데, 우리는 귀무가설에서 모평균이 15라고 가정하고 크기가 100인 표본을 뽑아 유의수준 0.05 에서 양측검정을 진행하고 있다. 모집단이 정규분포를 따르며 모분산이 20이라는 사실은 우리가 이미 알고 있었을 때, 제 2종의 오류를 저지를 확률은?

제 2종의 오류를 저지른다는 것은 대립가설이 맞음에도 귀무가설을 선택한다는 것이다. 따라서 실제로는 모평균이 10이고 모분산이 20인 정규분포를 따르는 경우에서 크기가 100인 표본을 뽑았는데 모평균이 15일 때의 채택역에 포함되면 된다.

귀무가설 H_0 , 즉 모평균이 15라는 가정 하에서 채택역은

$$\left[15 - 1.96 \frac{\sqrt{20}}{\sqrt{100}}, 15 + 1.96 \frac{\sqrt{20}}{\sqrt{100}} \right] = [14.123, 15.877]$$

이다. 제 2종의 오류를 저지르려면 $N(10, \frac{20}{100})$ 을 따르는 분포에서 여기에 포함되는 값이 등장해야 한다. 따라서, 구하는 확률은

$$P\left(\frac{14.123 - 10}{\sqrt{20}/10} \leq Z \leq \frac{15.877 - 10}{\sqrt{20}/10}\right) = P(9.22 < Z < 13.1) \approx 0$$

이 된다.

문제 8. 21. 어떤 제약사가 실험을 진행하고 있다. 이 제약사는 자신들의 약을 투여했을 때 근육량이 늘어난다고 주장한다. 일반적으로 성인 남성의 근육량은 20kg, 표준편차는 5kg이라 하자. 해당 제약사는 자신들의 제품을 복용하면 근육량이 5kg 정도 증가할 것이라고 예상하고 있다. 만약 그들의 예상이 맞을 때, 유의수준이 0.05인 가설검정 결과 제 2종의 오류를 저지를 확률이 20퍼센트 이하가 되려면 표본의 크기는 최소 얼마여야 하는가?

귀무가설 하에서는 모평균이 20이고, 유의수준이 0.05인 가설검정을 할 경우 채택역이

$$\left[20 - 1.96 \frac{5}{\sqrt{n}}, 20 + 1.96 \frac{5}{\sqrt{n}} \right]$$

이다. 그런데 그들의 예상이 맞다면 사실 실험 참가자들의 근육량의 모평균은 25kg이다. 따라서, 모평균이 25kg이고 표준편차가 5kg인 집단에서 표본을 n 개 뽑아서 표본평균을 구하였을 때 채택역에 포함될 확률이 0.2보다 작아야 한다. 즉,

$$P\left(20 - 1.96 \frac{5}{\sqrt{n}} < \bar{X} < 20 + 1.96 \frac{5}{\sqrt{n}}\right) \leq 0.2$$

이다.

$$\begin{aligned} P\left(20 - 1.96 \frac{5}{\sqrt{n}} < \bar{X} < 20 + 1.96 \frac{5}{\sqrt{n}}\right) &= P\left(\frac{20 - 1.96 \frac{5}{\sqrt{n}} - 25}{5/\sqrt{n}} < Z < \frac{20 + 1.96 \frac{5}{\sqrt{n}} - 25}{5/\sqrt{n}}\right) \\ &= P(-\sqrt{n} - 1.96 < Z < -\sqrt{n} + 1.96) \leq 0.2 \end{aligned}$$

만약 $n = 7$ 일 경우, 해당하는 확률은 약 0.248이며, $n = 8$ 일 경우 약 0.195이다. 따라서, n 이 8 이상일 때부터 해당하는 확률이 20퍼센트 이하가 됨을 알 수 있다. 따라서, 표본의 크기는 최대 8이어야 한다.

문제 8. 22. 특정 트랜지스터가 버틸 수 있는 전류의 세기가 최소 210A는 된다고 믿어진다. 표본을 구해 확인한 결과, 그들의 평균 한계전류는 200A였으며, 표본표준편차가 35였다. 유의수준 0.05에서 귀무가설을 기각할 수 있는지

- 1) 표본의 크기가 25일 때
- 2) 표본의 크기가 64일 때 판정하라.

1)

귀무가설 H_0 : 트랜지스터가 버틸 수 있는 전류의 세기 평균은 210A 이상이다.

대립가설 H_1 : 트랜지스터가 버틸 수 있는 전류의 세기 평균은 210A 미만이다.

이 경우의 기각역은

$$\bar{X} < 210 - t_{24,0.05} \frac{35}{\sqrt{25}} = 198.0$$

이다. 따라서 표본평균 200은 기각역에 포함되지 않는다. 따라서 귀무가설을 기각할 만한 충분한 근거가 없다.

2) 이 경우의 기각역은

$$\bar{X} < 210 - t_{63,0.05} \frac{35}{\sqrt{64}} = 202.7$$

이다. 따라서 표본평균 200은 기각역에 포함되고, 귀무가설을 기각할 수 있다.

문제 8. 23. 어떤 교수는 장어가 한우에 비해 더욱 비싸다고 주장하고 있다. 16개의 장어집에 대해 평균 가격은 72700원이었으며, 표본표준편차는 2400원이었다. 16개 고깃집에서 한우는 표본평균이 71400원, 표본표준편차가 2200원이었다. 교수의 주장이 옳은가? 유의수준 0.05에서 검정하라.

장어 가격과 한우 가격의 모평균이 각각 μ_1, μ_2 라고 하자. 그러면 귀무가설과 대립가설은 각각

$$H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_1 : \mu_1 - \mu_2 > 0$$

이 될 것이다. 이 때의 기각역은

$$\bar{X}_1 - \bar{X}_2 > t_{n_1+n_2-2,0.05} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

이다. 공통분산 S_p^2 의 공식을 이용하여 구하면 $S_p = 2302$ 이며, $n_1 = n_2 = 16$ 이다. 따라서, 기각역은

$$\bar{X}_1 - \bar{X}_2 > 1381$$

이다. 실제 표본평균의 차는 1300원으로, 기각역에 포함되지 못한다. 따라서 귀무가설을 기각할 충분한 근거가 없으며, 교수의 주장이 옳다고 확신할 수 없다.

문제 8. 24. 모비율에 대한 신뢰구간을 구하였듯이, 가설 $H_0 : p = p_0, H_1 : p \neq p_0$ 에 대하여 표본비율 \hat{p} 일 때의 가설검정 방법을 개발하라.

귀무가설이 옳은 경우,

$$\hat{p} \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$$

임을 배웠었다. 따라서 주어진 귀무가설과 대립가설에 대하여

$$\begin{aligned} 1 - \alpha &= P(-Z(\frac{\alpha}{2}) \leq Z \leq Z(\frac{\alpha}{2})) \\ &= P\left(-Z(\frac{\alpha}{2}) \leq \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq Z(\frac{\alpha}{2})\right) \\ &= P\left(p_0 - Z(\frac{\alpha}{2})\sqrt{\frac{p_0(1-p_0)}{n}} \leq \hat{p} \leq p_0 + Z(\frac{\alpha}{2})\sqrt{\frac{p_0(1-p_0)}{n}}\right) \end{aligned}$$

이므로 채택역은

$$\left[p_0 - Z(\frac{\alpha}{2})\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + Z(\frac{\alpha}{2})\sqrt{\frac{p_0(1-p_0)}{n}} \right]$$

이며,

$$\hat{p} > p_0 + Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{p_0(1-p_0)}{n}}$$

이거나

$$\hat{p} < p_0 - Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{p_0(1-p_0)}{n}}$$

일 경우 귀무가설을 기각한다.

문제 8. 25. 모비율에 대한 단측검정을 개발하라.

만약

$$H_0 : p \geq p_0, H_1 : p < p_0$$

일 경우에는 표본비율이 p_0 보다 충분히 낮을 경우 기각함을 알 수 있다. 따라서 기각역은

$$0 \leq \hat{p} < p_0 - Z(\alpha) \sqrt{\frac{p_0(1-p_0)}{n}}$$

이다. 반면, 채택역은

$$\left[p_0 - Z(\alpha) \sqrt{\frac{p_0(1-p_0)}{n}}, 1 \right]$$

만약

$$H_0 : p \leq p_0, H_1 : p > p_0$$

일 경우에는 표본비율이 p_0 보다 충분히 클 경우 기각함을 알 수 있다. 따라서 기각역은

$$p_0 + Z(\alpha) \sqrt{\frac{p_0(1-p_0)}{n}} < \hat{p} \leq 1$$

이다. 반면, 채택역은

$$\left[0, p_0 + Z(\alpha) \sqrt{\frac{p_0(1-p_0)}{n}} \right]$$

이다.