개념

# 통계

마타수학

## 통계

✓ 통계 단원에서는 기본 개념과 용어의 의미를 확실하게 정리해야한다.

확률변수는 어떤 시행에서 나올 수 있는 경우에 실숫값을 대응 시키는 것을 말합니다.

예를 들어 동전 던지는 시행에서 정할 수 있는 확률 변수는 동전이 앞면(나올 수 있는 경우)이 나오는 개수(실수값)이죠.

통계 문제에서 제일 중요한건 확률변수를 확실하게 이해하고 정하는 것입니다.

이산확률변수는 그 확률변수가 유한개 또는 셀 수 있는 값일 때 그 확률변수를 이산확률변수라 합니다.

✓ **확률질량함수**

이산확률변수 $X$가 갖는 모든 값을 $x_1, x_2, x_3, \ldots, x_n$이라 하면 $X$가 이들 값을 가질 확률 $p_1, p_2, p_3 \cdots, p_n$사이의 대응되는 함수 $P(X=x)=p_i (i=1,2,3\ldots,n)$을 이산 확률 변수 $X$의 질량 함수라 한다.

✓ **확률질량 함수의 성질**(문제풀이에 중요하니반드시 암기)

1. $0 \le p_i \le 1$
2. $p_1+p_2+p_3+\ldots+p_n=\sum_{i=1}^{n} p_i=\sum_{i=1}^{n} P(X=x_i)=1$
3. $P(x_i \le X \le x_j)=p_i+p_{i+1}+p_{i+2}+\ldots+p_j$(단, $i, j=1,2,..,n$이고 $i \le j$)

✓ 이항분포

이항분포는 문제를 볼 때 같은 행동을 반복하여 시행할 때 그것을 이항분포라고 생각하면 됩니다.

시행횟수를 n 사건이 일어나 확률을 p라 한다.

✓ 통계단원의 문제 풀이법

통계단원은 개념을 이해하면 더 쉽게 풀수있지만 개념이해가 안된다 하더라도 문제해석만 잘하면 쉽게 풀수 있는 단원입니다.

1. 확률변수가 무엇인지 확인한다.
2. 구하고자 하는 것을 확인한다.
3. 문제지를 읽으면서 각 조건들을 정리한다.
4. 관련 개념,성질 등을 이용해 조건들을 활용한다.
5. 답나옴 끝

**9.** 확률변수 $X$가 이항분포 $B(9, p)$를 따르고 $\{E(X)\}^2 = V(X)$일 때, $p$의 값은? (단, $0 < p < 1$) [3점]

① $\frac{1}{13}$ ② $\frac{1}{12}$ ③ $\frac{1}{11}$ ④ $\frac{1}{10}$ ⑤ $\frac{1}{9}$

간단하죠? 예제를 통해 연습을 해볼까요?

이항분포 문제입니다.
1step) 확률변수는 X입니다.
2step) p확률값을 물어보고있죠?
3step) $1. B(9,p)$
$2. E(X)^2 = V(X)$

문제에서 주어진 조건입니다.
4step) 이항분포 개념을 통해 E(X)와 V(X)를 구한다.
E(X)=9p V(x)=9p(1-p)입니다. 조건식에 각각 대입해주면

$81p^2 = 9p - 9p^2$이고 정리해주면
$90p^2 - 9p = 0$이 되고 인수분해를 해주면
$9p(10p-1) = 0$
$\therefore p = 0, \frac{1}{10}$이 되고 조건상 $0 < p < 1$이므로
$p = \frac{1}{10}$이 됩니다.

✓ 연속확률변수와 확률밀도함수

연속확률변수도 이산확률변수과 확률변수의 특징이 다를 뿐 평균과 분산,표준편차를 구하는 방법은 동일합니다.

✓ 이산확률변수과 연속확률변수 개념 비교

이산확률변수는 셀 수 있는 값들이고 연속확률변수는 확률변수가 연속적으로 변하는 값들이 됩니다.

확률밀도함수에서 x에 해당하는 것이 이산확률변수의 $x_n$이 되고 x에 대응되는 함수값은 이산확률변수에서의 확률 $p_i$에 대응 됩니다.

이후 공식은 똑같습니다.

✓ 정규분포

정규분포는 확률밀도함수의 f(x)가 m(평균) δ(표준편차)(δ>0)에 대하여 특정 함수(개념서에 나와있어요.)로 주어질 때 확률변수 X가 정규분포를 따른다고 하고 그 그래프를 정규분포곡선이라고 합니다.

이 때 평균과 표준편차를 정규분포 기호로 $N(m,\sigma^2)$으로 나타냅니다.

✓ 정규분포 곡선의 성질

1. x=m에 대하여 좌우대칭인 종 모양의 곡선이다
2. x=m 일 때 최댓값을 가지고 x축을 점근선으로 한다.
3. 곡선과 x축 사이의 넓이는 1이다.
4. 평균이 일정할 때 표준편차의 값이 커질수록 곡선은 가운데 부분이 낮아지고 양쪽으로 퍼진다.(**이해하기: 표준편차는 각 확률변수들의 차이라고 이해하면됩니다. 차이가 커지면 커질수록 평균에서 멀어질 것이니깐 당연히 가운데 부분이 낮아지고 양쪽으로 퍼지겠죠.)

5.표준편차가 일정할 때 , 평균이 변하면 곡선의 모양은 동일하고 곡선이 x축 방향으로 평행이동을 한다.

<정규분포곡선의 성질은 읽어서 이해하고 넘어가면 충분합니다.>

✓ 표준정규분포

표준정규 분포는 평균이 0 ,표준편차가 1인 정규분포 N(0,1)을 표준정규분포라 합니다. 표준정규분포의 확률변수는 Z라 한다.

중요!

✓ 정규분포의 표준정규분포화

확률변수 X가 N(m,$\sigma^2$)을 따를 때, 확률변수 Z=$\frac{X-m}{\sigma}$ 이고 표준정규분포를 따른다.

✓ 이항분포와 정규분포

확률변수 X가 이항분포 B(n,p)를 따를 때, n이 충분히 크면 X는 근사적으로 정규분포 N(np,np(1-p))를 따른다.

(문제에서 n은 충분히 큰 경우로 주어지니깐 개념적으로만 이해하면됩니다.)

연속확률분포,정규분포등은 통계단원풀이법을 앞에서 언급한바와 같이 풀이를 해주면 됩니다.

문제를 통해 연습해보죠.

**19.** 확률변수 $X$가 평균이 $\frac{3}{2}$, 표준편차가 2인 정규분포를 따를 때, 실수 전체의 집합에서 정의된 함수 $H(t)$는

$$H(t)=\mathrm{P}(t \le X \le t+1)$$

이다. $H(0)+H(2)$의 값을 오른쪽 표준정규분포표를 이용하여 구한 것은? [4점]

| $z$ | $\mathrm{P}(0 \le Z \le z)$ |
|---|---|
| 0.25 | 0.0987 |
| 0.50 | 0.1915 |
| 0.75 | 0.2734 |
| 1.00 | 0.3413 |

① 0.3494 ② 0.4649 ③ 0.4852
④ 0.5468 ⑤ 0.6147

1. 확률변수는 X

2. H(0)+H(2)를 하는거죠. 조건을 이용해 다시표현해 보면 H(0)=$P(0 \le X \le 1), H(2)=P(2 \le X \le 3)$을 구하는겁니다.

3. $N(\frac{3}{2}, 2^2)$인 정규분포를 따른다는 조건이 있죠.

4.정규분포를 표준정규분포화해서 위 H(0)과 H(2)를 구하는 겁니다. X->Z(표준정규분포의 확률변수)로 표현해야하는거죠.

$H(0), H(2)$을 표준정규분포 $Z$로 표현해보면

$$H(0)=P(\frac{0-\frac{3}{2}}{2} \le \frac{X-\frac{3}{2}(=m)}{2(=\sigma)}(=Z) \le \frac{1-\frac{3}{2}}{2})$$
$$=P(-0.75 \le Z \le -0.25)$$

$H(2)=P(0.25 \le Z \le 0.75)$가됩니다.

표준정규분포곡선은 $m=0$에 대하여 대칭이므로

$H(0)=0.2734-0.0987=0.1747$
$H(2)=0.2734-0.0987=0.1747$이므로
$H(0)+H(2)=0.3494$가됩니다.

27. A 과수원에서 생산하는 귤의 무게는 평균이 86, 표준편차가 15인 정규분포를 따르고, B 과수원에서 생산하는 귤의 무게는 평균이 88, 표준편차가 10인 정규분포를 따른다고 한다. A 과수원에서 임의로 선택한 귤의 무게가 98 이하일 확률과 B 과수원에서 임의로 선택한 귤의 무게가 $a$ 이하일 확률이 같을 때, $a$의 값을 구하시오.
(단, 귤의 무게의 단위는 g이다.) [4점]

1.확률변수를 확인하세요. 확률변수는 귤의 무게입니다.

2. 귤의 무게a를 구하는 겁니다

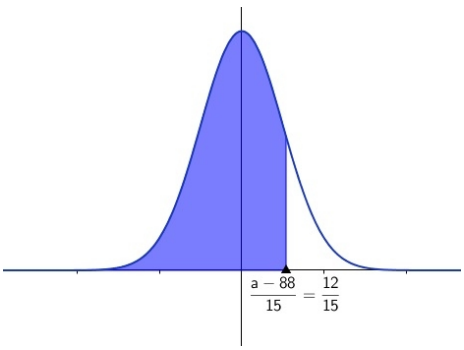
3. A과수원은 A-N(86,$15^2$) B과수원은 B-N(88,$10^2$) 따른다는 정규분포에 대한 조건과
$P(X(A) \le 98) = P(X(B) \le a)$조건이 주어집니다.이때
$X(A)$는 $A$과수원의 확률변수 $X$를 의미하고
$X(B)$는 $B$과수원의 확률변수 $X$를 의미합니다.

4. A,B 과수원의 정규분포를 표준정규분포화하여 확률을 구하여 비교해줍니다.

$A$과수원에서의 확률변수를
표준정규분포의 확률변수로 바꾸어 확률을 구하면
$P(Z \le \frac{98-86}{15} = \frac{12}{15})$이 됩니다.

$B$과수원에서의 확률변수를
표준정규분포의 확률변수로 바꾸어 확률을 구하면
$P(Z \le \frac{a-88}{10})$이 됩니다.
두 확률이 같으려면 $\frac{12}{15} = \frac{a-88}{10}$ 이어야 하므로
$10 \times \frac{12}{15} = a - 88$
$8 = a - 88$
$a = 88 + 8 = 96$
$\therefore a = 96$

그래프를 그리면 아래와 같겠죠? 그래프를 반드시 그리면서 풀이하세요.




## 통계적 추정

통계적 추정 단원 개념에 앞서 통계적 추정을 왜 배우는지에 대해서 설명하고 넘어가겠습니다.

건전지 회사의 품질관리팀장이 직원보고 건전지 사용시간을 조사하여 통계를 내오라고 했습니다.

직원이 건전지 사용시간을 조사를 끝나고 나서
"우리 회사의 건전지 사용시간은 평균24시간이네."라는 통계를 냈지만 회사에서 생산한 모든 건전지를 다 사용해버려서 판매할 건전지가 없다.

이와같이 통계는 어떤 물건이나 사회에서 필요하지만 전부 다 조사하게 되면 경제적 비용이 크게 듭니다.

그래서 조사하고자 하는 것의 전체가 아닌 일부를 조사하여 전체의 특성을 예측하고자 하는 것이 바로 통계적 추정입니다.

모집단(조사하고자하는 집단의 전체)에서 일부분을 추출해서 그것을 조사하여 모집단의 특성을 예측한다.-통계적 추정

이때 모집단에서 뽑아낸 일부의 자료로 된 집합을 표본이라한다.

추출된 표본집단의 원소의개수를 표본의크기(n)라한다.

✓ 표본평균의 분포

모집단에 대한 확률변수 X에 대하여
표본의 평균을 "표본평균", 표준편차를 "표본표준편차"라 한다. 표본의 평균은 기호로 $\overline{X}$라한다.

1,2,3,4의 숫자가 각각 쓰여진 4개의 공이 들어있는 주머니에서 한 개의 공을 꺼내는 시행을 한다.
이 때 ,공의 숫자를 확률변수 X라 할 때, 확률변수 X의 분포는

| X | 1 | 2 | 3 | 4 | 합계 |
|---|---|---|---|---|---|
| P | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | 1 |

이다.

확률변수 X의 평균과 표준편차를 구해보면

평균=$\frac{5}{2}$ 표준편차=$\frac{\sqrt{5}}{2}$이다.

이때 크기가 2인 표본을 추출한다고 했을 때 표본은
(1,1)(1,2),(1,3),(1,4) (2,1)(2,2)(2,3)(2,4)
(3,1)(3,2)(3,3)(3,4) (4,1)(4,2)(4,3)(4,4)의
16가지 경우가 생기고 이들 표본의 평균($\overline{X}$)은

(1,1)=1 (1,2)=1.5 (1,3)=2 (1,4)=2.5
(2,1)=1.5 (2,2)=2 (2,3)=2.5 (2,4)=3
(3,1)=2 (3,2)=2.5 (3,3)=3 (3,4)=3.5
(4,1)=2.5 (4,2)=3 (4,3)=3.5 (4,4)=4이다.

이 표본들에 대한 확률변수 $\overline{X}$의 확률분포는

| $\overline{X}$ | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
|---|---|---|---|---|---|---|---|
| P ( $\overline{X}$) | $\frac{1}{16}$ | $\frac{2}{16}$ | $\frac{3}{16}$ | $\frac{4}{16}$ | $\frac{3}{16}$ | $\frac{2}{16}$ | $\frac{1}{16}$ |

이다.

이 때, 표본평균 $\overline{X}$의 평균과 표준편차를 구해보면

E($\overline{X}$)=$\frac{5}{2}$ $\sigma(\overline{X})$=$\frac{\sqrt{10}}{4}$이 된다.

✓ 표본평균 $\overline{X}$의 확률분포에 대한 성질

모평균을 m, 모표준편차가 $\sigma$인 모집단에서 크기가 n인 표본을 추출 할 때, 표본평균을 $\overline{X}$라 하면
확률변서 $\overline{X}$의 분포에 대하여

1. $E(\overline{X})=m$ , $V(\overline{X})=\frac{\sigma^2}{n}$ , $\sigma(\overline{X})=\frac{\sigma}{\sqrt{n}}$
2.모집단의 분포가 정규분포일 때는 $\overline{X}$의 분포는 n의 크기에 관계없이 정규분포 N(m,$\frac{\sigma^2}{n}$)을 따른다.

이 단원은 표본평균의 정규분포를 구한다음 이후 풀이법은 앞의 통계단원 풀이대로 하면 됩니다.

12. 어느 약품 회사가 생산하는 약품 1병의 용량은 평균이 $m$, 표준편차가 10인 정규분포를 따른다고 한다. 이 회사가 생산한 약품 중에서 임의로 추출한 25병의 용량의 표본평균이 2000 이상일 확률이 0.9772일 때, $m$의 값을 오른쪽 표준정규분포표를 이용하여 구한 것은? (단, 용량의 단위는 mL이다.) [3점]

| $z$ | $P(0 \le Z \le z)$ |
|---|---|
| 1.5 | 0.4332 |
| 2.0 | 0.4772 |
| 2.5 | 0.4938 |
| 3.0 | 0.4987 |

① 2003 ② 2004 ③ 2005 ④ 2006 ⑤ 2007

이 문제에서 확률변수 X는 약품1병의 용량입니다.

확률변수 X의 평균 m을 구하는 문제이구요.

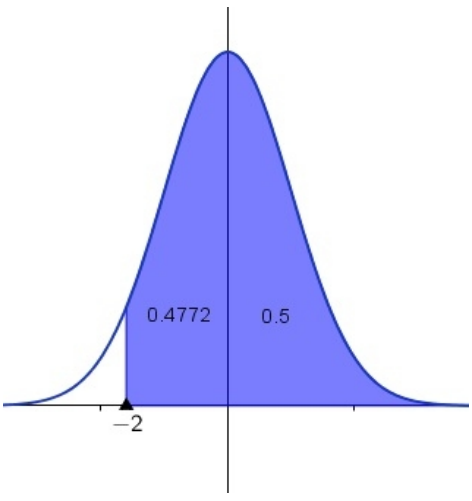
크기가 25로 모집단에서 표본을 추출합니다.

모집단은 N(m,$10^2$)인 정규분포를 따르므로
표본집단은 N(m,$2^2$)인 정규분포를 따른다.

$P(\overline{X} \ge 2000) = 0.9772$라는 조건이 주어져있습니다.

확률변수 $\overline{X}$를 정규분포의 확률변수 Z로 바꾸어 표현하면
$P(Z \ge \frac{2000-m}{2})$=0.9772입니다. 주어진 표준정규분포표를 이용하여 m값을 구해보면

표준정규분포곡선은 $m=0$에 대하여 좌우대칭이고
곡선과 $x$축 사이의 넓이는 1이므로 0을 기준으로해서 좌우의 확률은 0.5이 됩니다. 0.9772의 확률을 가지려면 0.5의 확률과 0.4772의 확률의 합으로 나타내져야합니다.
그래프는 아래와 같습니다.




$\frac{2000-m}{2}$의 값은 $-2$이고, 정리하여 $m$을 구해보면

$2000-m=-4$

$2000+4=m$

$\therefore m=2004$

✓ 모평균의 추정

자 이제 이 개념만 잘 이해하면 통계는 끝!! 잘이해해 봅시다.

모집단에서 평균,표준편차 등을 알지 못할 때 역으로 표본을 이용하여 모집단의 평균과 표준편차의 값을 추측하는 방법이 모평균의 추정입니다.

✓ 신뢰도의 의미

표본을 이용하여 구한 값들이 적중할 확률을 그 추정의 '신뢰도'라고 합니다.

예를들어, 수능수학시험을 친 수험생 전체에서 임의로 50명을 뽑아 조사하였더니 평균이 80점 표준편차 15점이었다.라는 표본조사의 결과를 가지고 수험생 전체의 수학성적을 말할 때

㉠평균점수가 76점에서 84점 사이이다.
㉡평균점수가 75점에서 85점 사이이다.

라고했을 경우. 1번과 2번의 추정이 옳을 확률을 각각 95%,99%라고 한다면

㉠은 95%의 신뢰도를 갖는다고 하고
㉡은 99%의 신뢰도를 갖는다고 합니다.

또한, ㉠의 점수구간 [76,84]과 ㉡의 점수구간 [75,85]을 '신뢰구간'이라고 합니다.

✓ 모평균의 추정

표본평균은 개념설명했을 때와 같이 서로 다른값이 나옵니다. 그러므로 신뢰구간 또한 변합니다.
하지만 이 구간들의 99%정도는 모평균 m을 포함것으로 추정된다라고 하는 것이 신뢰도 99%의 신뢰구간의 뜻입니다.

신뢰구간을 통해 모평균 m이 속하는 범위를 추정하는 것입니다.

✓ 표본의 크기를 n,표본평균을 $\overline{X}$,표본표준편차를 $\sigma$라 할 때

1)신뢰도 95%일 때 모평균 m의신뢰구간

$$[\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \le m \le \overline{X}+1.96\frac{\sigma}{\sqrt{n}}]$$

2)신뢰도가 99%일 때 모평균 m의신뢰구간

$$[\overline{X}-2.58\frac{\sigma}{\sqrt{n}} \le m \le \overline{X}+2.58\frac{\sigma}{\sqrt{n}}]$$

17. 어느 밭에서 수확한 딸기의 무게는 정규분포를 따른다고 한다. 이 딸기 중에서 임의추출한 $n$개의 무게를 조사하였더니 평균이 20 g, 표준편차가 5 g이었다. 이 결과를 이용하여 이 밭에서 수확한 딸기 무게의 평균을 신뢰도 95 %로 추정한 신뢰구간이 $[19.02, a]$이다. $n+a$의 값은? (단, 표준정규분포를 따르는 확률변수 $Z$에 대하여 $P(0 \le Z \le 1.96) = 0.4750$이다.) [4점]

① 84.98 ② 85.96 ③ 101.02
④ 120.98 ⑤ 121.96

확률변수 X는 딸기의 무게입니다.

신뢰구간과 표본의 크기 n을 물어보고있습니다.

표본평균$\overline{X}$=20 표본표준편$\sigma$=5 라는 조건이 주어졌습니다. (여기서 표본평균이 왜 20인지 이해 못할 수 있어서 위에서 개념설명한걸 다시 반복하자면 표본평균은 표본들의 평균입니다. 문제에서n의 크기로 표본을 뽑아 무게를 조사하고 그 평균이 20 이라했으므로 표본평균이 20인거에요.)

그리고 신뢰구간이 [19.02,a]라고 했습니다.

주어진 조건으로 문제를 풀이해보면

$[20-1.96\frac{5}{\sqrt{n}}, 20+1.96\frac{5}{\sqrt{n}}]$입니다.
$[19.02, a]$와 동일해야하므로

$20-1.96\frac{5}{\sqrt{n}} = 19.02$
$1.96\frac{5}{\sqrt{n}} = 0.98$
$\frac{5}{\sqrt{n}} = \frac{1}{2}$
$\sqrt{n} = 10 \quad \therefore n = 100$
$a$는 $19.02$와 $0.98$의 두 배만큼 차이가 나므로
$a = 19.02 + 1.96 = 20.98$

$a = 20.98$ , $n = 100$
$\therefore a + n = 120.98$