통계학 개론

제4장 확률분포와 표본분포

4.1 확률분포

확률변수 X의 확률분포: 확률변수 X의 값에 따라 확률이 어떻게 분포하는지를 합이 1이 되도록 나타낸 것

이산형 확률분포: 이항분포, 초기하분포, 포아송분포

연속형 확률분포: 정규분포

1) 이항분포(Binomial distribution)

베르누이 시행(Bernoulli trial): 가능한 결과가 두 가지(성공, 실패)이고, 이 실험이 반복되는 것

베르누이 확률변수: 성공확률을 p라고 할 때 '성공'이면 1, '실패'면 0으로 대응시키는 확률변수

베르누이분포: 베르누이 확률변수 X의 확률분포

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

❖ 이항분포

성공확률이 p인 베르누이 실험을 n번 독립적으로 반복 시행할 때 '성공횟수(X)'가 x일 확률

$P(X=x)={}_nC_r p^x(1-p)^{n-x},\ x=0,1,2,\ \cdots,n$

이항분포의 평균은 $E(X)=np$ 이고, 분산은 $Var(X)=np(1-p)$

이항분포의 모수: n, p

2) 초기하분포(Hypergeometric distribution)

❖ 초기하분포

N은 모집단의 크기, D는 모집단에서 특성값 1의 개수, n은 표본크기, x는 표본에서 특성값 1의 개수일 때

$$P(X=x)=\frac{{}_DC_x \times {}_{N-D}C_{n-x}}{{}_NC_n},\ x=0,1,2,\ \cdots\ ,n$$

이다. 단, n≤N, x≤D이다.

평균: $E(X)=np,\quad$ 단, $p=\dfrac{D}{N}$

분산: $Var(X) = np(1-p) \times \frac{N-n}{N-1}$

3) 포아송분포(Poisson distribution)

일어날 확률이 아주 작은 경우에 적용가능한 확률분포임

포아송분포가 적용되기 위한 조건

① 독립성: 한 단위시간이나 공간에서 출현하는 성공횟수와 중복되지 않는 다른 단위시간이나 공간에서 출현하는 성공횟수는 서로 독립적이다.

② 비집락성: 극히 작은 시간이나 공간에서 둘 또는 그 이상의 성공이 같이 일어날 확률은 매우 작으며 0으로 간주된다.

③ 비례성: 단위시간이나 공간에서 성공의 평균출현횟수는 일정하며, 이는 시간이나 공간에 따라 변하지 않는다.

❖ 포아송분포

확률변수 X가 위의 세 가지 조건을 만족할 때, 성공의 평균출현횟수를 m이라고 하면 X의 확률분포는 다음의 포아송분포를 따른다.

$$P(X=x) = \frac{e^{-m} m^x}{x!}, \ x = 0, 1, 2, \cdots$$

평균 = m

분산 = m

4) 정규분포

❖ 정규분포함수(normal distribution function)

= 가우스분포함수(Gaussian distribution function)

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

정규분포의 특징

① 종모양의 연속함수이다

② 평균 μ에 대해 서로 대칭이다. 따라서 평균의 왼쪽과 오른쪽의 확률은 각각 0.5이다

③ μ나 σ의 값에 따라 정규분포는 무한히 많이 있을 수 있다.

④ x축의 구간 [$\mu-\sigma$, $\mu+\sigma$]의 확률은 0.68, 구간 [$\mu-2\sigma$, $\mu+2\sigma$]의 확률은 0.95, 구간 [$\mu-3\sigma$, $\mu+3\sigma$]의 확률은 0.997이 된다. 즉, 정규확률변수의 평균 주위에

대부분의 값을 가지며, 평균에서 좌우로 표준편차의 3배 이상 떨어진 값은 거의 없다.

확률변수 X가 $N(\mu, \sigma^2)$인 정규확률변수일 때 구간 $[a, b]$의 확률 $P(a \le X \le b)$는 다음과 같다.

$$P(a \le X \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left[- \frac{(x-\mu)^2}{2\sigma^2} \right] dx$$

다행히 X가 $N(\mu, \sigma^2)$인 정규확률변수일 때 $Z = (X-\mu)/\sigma$ 변환을 취하면 Z는 평균이 0이고, 표준편차가 1인 정규분포 $N(0, 1)$을 따르게 된다.

☞ $N(0, 1)$인 분포의 모든 확률을 구할 수 있다면, 임의의 정규분포도 확률을 구할 수 있다

❖ 표준정규분포(standard normal distribution)

평균이 0이고, 표준편차가 1인 정규분포: N(0, 1)

❖ 표준화 변환(standardization)

X가 평균이 μ이고, 분산이 σ^2인 정규분포 $N(\mu, \sigma^2)$일 때, 변환

$$Z = \frac{X-\mu}{\sigma}$$

는 평균이 0이고, 표준편차가 1인 표준정규분포 $N(0, 1)$을 따른다.

❖ 표준정규분포표

여러가지 실수값 z에 대해서 왼쪽 끝부분에서 z까지의 면적인 확률 P(Z<z)를 구하여 작성해둔 표

90%: -1.645 < Z < 1.645

95%: -1.96 < Z < 1.96

99%: -2.575 < Z < 2.575

4.2 표본분포

통계적 추론(statistical inference): 모집단에서 일부를 추출한 표본을 잉ㅇ하여 모집단에 관한 추측이나 결론을 이끌어 내는 과정

모수(parameter): 모집단의 특성값

통계량(ststistics): 표본에서 구한 특성값

표본분포(sampling distribution): 통계량의 분포

1) 표본평균의 분포

모평균 μ와 모분산 σ^2를 갖는 모집단에서 추출한 랜덤표본을 X_1. X_2, …, X_n 이라고 하면 이들의 표본평균, 기댓값, 분산은 다음과 같다.

표본평균: $\overline{X}=\frac{1}{n}(X_1+X_2+\cdots+X_n)$

기댓값: $E(\overline{X})=\frac{1}{n}(E(X_1)+E(X_2)+\cdots+E(X_n))=\frac{1}{n}\cdot n\mu=\mu$

분산: $Var(\overline{X})=\frac{1}{n^2}(Var(X_1)+Var(X_2)+\cdots+Var(X_n))=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}$

❖ 표본평균의 분포

정규모집단 $N(\mu,\ \sigma^2)$ 으로부터 추출한 랜덤표본을 X_1, X_2, …, X_n 이라고 할 때, 표본평균 $\overline{X}$의 표본분포는 정규분포 $N(\mu,\frac{\sigma^2}{n})$이다.

❖ 중심극한정리(central limit theorem)

모집단이 무한모집단이고, 표본크기(n)가 충분히 크면 모집단이 어떠한 분포라도 표본평균의 분포는 근사적으로 정규분포이다.

평균이 μ이고, 분산이 σ^2인 임의의 무한모집단에서 표본크기(n)가 충분히 크면, 표본평균 $\overline{X}$의 분포는 근사적으로 평균이 μ이고, 분산이 σ^2인 정규분포를 따른다.

$\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$

❖ 이항분포의 정규근사

이항분포 B(n, p)를 따르는 확률변수 X는 n이 충분히 클 때 근사적으로 평균이 np, 분산이 np(1-p)인 정규분포 N(np, np(1-p))를 따른다. 즉, n이 충분히 크면 다음이 성립한다.

$\frac{X-np}{\sqrt{np(1-p}}\sim N(0,1)$

중심극한정리에 따르면 표본크기가 충분히 큰 경우는 모집단이 정규분포를 따르지 않는다고 해도 표본평균 $\overline{X}$의 분포는 근사적으로 정규분포를 따른다. 이 경우 모분산 σ^2을 모른다면 이를 표본에서 구한 표본분산 S^2으로 대치함으로써 표본평균 $\overline{X}$의 분포를 파악할 수 있다.

❖ t분포

정규분포 $N(\mu, \sigma^2)$을 따르는 모집단으로부터 얻어진 확률표본을 $X_1, X_2, \cdots, X_n$ 이라고 할 때,

$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ 라 하면, T는 자유도 (n-1)인 t분포, 즉 t(n-1)을 따른다.

여기서 표본표준편차 $S = \sqrt{\frac{\sum_{t=1}^{n}(X_i - \overline{X})^2}{n-1}}$ 이다.

정규분포가 평균 μ와 분산 σ^2에 의해 완전히 결정되는 분포인 것처럼 t분포는 자유도에 의해 결정되는 분포이다. 따라서 t분포에는 반드시 자유도가 명시되어야 한다.

t분포는 평균이 0이고, 0을 중심으로 대칭인 분포를 하고, 표준정규분포와 비슷한 형태의 분포를 하지만, 표준정규분포보다는 좌우로 더 멀리 퍼져 봉우리는 낮고 꼬리가 두꺼운 분포개형을 갖는다. 또한 자유도가 작을수록 더 멀리 퍼진 형태를 나타내고, 자유도가 커질수록 표준정규분포에 점점 가까워진다.

2) 표본분산의 분포

모집단의 모분산과 표본에서 얻어지는 표본분산 사이의 관계를 알 수 있다면 역시 미지의 모분산을 추정하는데 많은 도움이 된다.

일반적으로 표본분산의 분포는 모집단이 정규분포이고 모분산이 σ^2일 때, 표본분산의 분포는 χ^2분포(χ^2 distribution)을 따른다.

❖ 표본분산의 분포

모집단이 모분산 σ^2인 정규분포를 따를 때 크기가 n인 표본을 랜덤표본을 추출하면 $(n-1)S^2/\sigma^2$ 은 자유도가 (n-1)인 χ^2분포를 따른다. 즉,

$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$$

이다. 여기서 $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ 이다.