7주

제4장 확률변수(3) – 이항분포의 정규근사

이항분포의 정규근사

- 이항분포
  - 어느 특정한 속성을 가지는 개체의 비율을 p, 이러한 집단으로부터 임의로 추출된 크기가 n인 확률표본에서 그 특정한 속성을 가지는 원소의 수를 X 라고 하자.
  - 확률변수 X는 평균이 np 이고 표준편차가 $\sqrt{np(1-p)}$인 이항확률분포를 따름, 이항확률분포

$$f(x) = P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0,\ 1,\ \ldots, n$$

  - $X \sim B(n, p)$일 때, np와 $\sqrt{np(1-p)}$ 가 둘 다 크면(대체로 5보다 클 때) X는 평균이 np이고 표준편차가 $\sqrt{np(1-p)}$ 정규분포에 매우 잘 근사함

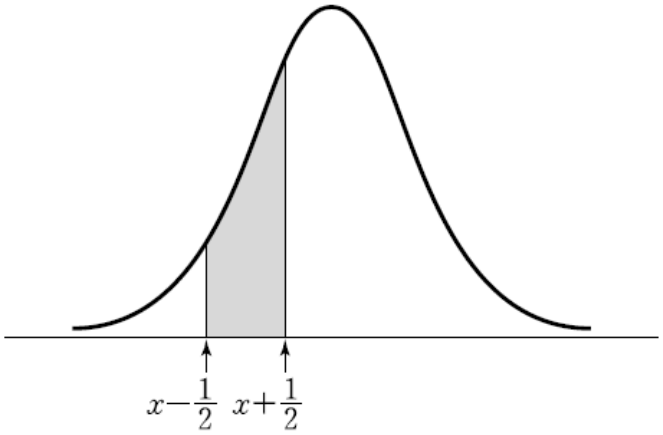
$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

    이 근사적으로 표준정규분포를 따른다

- 연속성의 수정(continuity correction)
  - 이항확률변수 X에 대한 확률을 표준정규분포를 이용하여 계산
  - 근사 확률을 계산할 때 이산형의 확률 값을 연속형의 표준정규분포로 근사시키는 것이기 때문에 연속성의 수정을 하여야함
  - 이산확률변수 X에 대한 확률 $P[X = x]$를 다음과 같이 수정

$$P[X = x] \simeq P[x - 1/2 \leqq X \leqq x + 1/2]$$

  - 연속확률변수의 경우는 한 값을 취할 확률이 항상 0이기 때문에 이산확률변수의 확률을 얻기 위해서는 1/2을 더해주고 빼서 구간의 형태로 수정한 후 연속확률변수에 대한 확률을 근사 시킴



| 그림 4. 13 | 연속성의 수정

- 예) 우리나라 중・고교생 가운데 근시인 학생은 20 % 라고 한다. 이제 우리나라 중・고교생 중에서 100 명을 표본으로 추출하여 시력을 검사하였다. 그 중에서 근시인 학생이 15 명에서 18 명 사이로 나타날 확률을 구하여 보자. (출처: 통계학의 이해)
  - X~B(100, 0.2)
  - 물론 근시인지 아닌지 두 가지 결과로 나타나는 현상이기 때문에 근시인 학생 수는 이항분포를 한다. 그러므로 이항분포로부터 확률을 구할 수는 있다.
  - n = 100, p = 0.20 이므로 X 가 15 에서 18 사이에 갖게 될 확률

$$P(15 \le X \le 18) = P(X = 15) + P(X = 16) + P(X = 17) + P(X = 18)$$
$$= \frac{100!}{15!85!}(0.20)^{15} \cdot (0.80)^{85} + \cdots + \frac{100!}{18!82!}(0.2)^{18}(0.8)^{82}$$
$$= ????$$

- 풀이
  - 이항확률변수 X 는 n = 100 이고 p = 0.20, X ~ B (100, 0.2)
    - 평균=np=20
    - 분산=np(1-p)=16
    - 표준편차=4
  - 표본 크기가 충분히 크기 때문에 정규분포로 접근시켜 확률을 구함
    무수히 많은 표본을 뽑는다면 근시인 학생 수는 확률변수 X 로서 평균이 20 이고 분산이 16 인 정규분포에 접근한다.
    따라서 $P(15 \le X \le 18)$ 은 연속성 수정하여 $P(14.5 < X < 18.5)$ 로 구함

$$P(15 \le X \le 18) = P\ ((14.5-\mu)/\sigma < Z < (18.5-\mu)/\sigma$$

P(15≤ X ≤ 18)

= P ((14.5-μ)/σ < Z < (18.5-μ)/σ

$$= P\left(\frac{14.5-20}{4} < \frac{X-\mu}{\sigma} < \frac{18.5-20}{4} \right)$$

= P(-1.38 < Z < -0.38) ; 좌우대칭분포이므로

= P(Z < -0.38) – P(Z < -1.38) ; 표준정규분포로부터 값을 구하면

= (1-0.6480)-(1-0.9162)

= 0.3520-0.0838

= 0.2682

∴ 따라서 중・고교생 가운데서 100 명을 뽑았을 때
근시인 학생이 15 명에서 18 명 사이로 나타날 확률은 26.82 %

- [예제 4.21] 5년 전에 행해진 대규모 조사에서 성인의 10%가 왼손을 주로 사용하는 왼손잡이라고 한다. 만일 이 비율이 현재에도 적용된다면 성인 1000명의 확률표본에서 왼손잡이인 사람이

  (a) 120명 이하일 확률은 얼마인가?

  (b) 90명 이상 100명 이하일 확률은 얼마인가?

확률변수 X를 성인 1000명의 확률표본에서 왼손을 주로 사용하는 사람의 수라고 하자. 그러면 X는 $n=1000$이고 $p=0.1$인 이항분포를 따른다.

$$np=100,\quad \sqrt{np(1-p)}=\sqrt{90}=9.49$$

이므로 확률변수 X는 근사적으로 $N(100, 9.49^2)$을 따른다. 따라서

(a)
$$P[X \leqq 120] \simeq P\left[Z \leqq \frac{120-100}{9.49}\right]$$
$$= P[Z \leqq 2.11]$$
$$= 0.9826$$

(b)
$$P[90 \leqq X \leqq 100] \simeq P\left[\frac{90-100}{9.49} \leqq Z \leqq \frac{100-100}{9.49}\right]$$
$$= P[-1.05 \leqq Z \leqq 0]$$
$$= 0.5-0.1469=0.3531$$

▨

- 예) 어느 영어시험에서의 점수 X는 평균이 497점이고 표준편차가 120인 정규모집단에서의 관측 값으로 간주할 수 있다고 한다. 이 때 다음을 구하라.

  (a) 600점 이상의 사람들만 합격한다고 했을 때 합격자의 비율을 구하라.

  (b) 상위 10%에 해당되는 사람들만 합격한다고 했을 때, 합격자 중에서 최저 점수는 얼마인가?

$$X \sim N(497, 120^2)$$

$$(a) P(X \geq 600) = ?$$

$$P(X \geq 600) = P(Z. \geq \frac{600-497}{120}) = P(Z \geq 0.86)$$

$$= 1 - P(Z \leq 0.86) = 1 - 0.8051 = 0.1949$$

$$\therefore 19.49\%$$

$$(b) Z_{0.1} = 1.28, X = ?$$

$$\frac{X-497}{120} = 1.28$$

$$\therefore X = 650.6\text{점}$$

- 예) A 지역에 거주하는 성인의 30%가 습관적으로 술을 마신다고 한다. 이 지역에서 성인 1000명을 임의로 추출했을 때, 그 중에 습관적으로 술을 마시는 사람이

  (a) 280명 이하일 확률은 얼마인가?

  (b) 300명 이상 330명 이하일 확률은 얼마인가?

$$(a)\, X \sim B(1000, 0.3) \approx N(300, 210^2)$$

- $E(X) = np = 1000 \times 0.3 = 300$
- $Var(X) = npq = 1000 \times 0.3 \times 0.7 = 210$

$$P(X \le 280) = ?$$

$$P(X \le 280) = P(Z < \frac{280.5 - 300}{\sqrt{210}}) = P(Z < -1.35)$$

$$= 1 - P(Z < 1.35) = 1 - 0.9115 = 0.0885$$

$$(b)\, P(300 \le X \le 330) = P(\frac{299.5 - 300}{\sqrt{210}} < Z < \frac{330.5 - 300}{\sqrt{210}})$$

$$= P(\frac{-0.5}{\sqrt{210}} < Z < \frac{30.5}{\sqrt{210}}) = P(-0.035 < Z < 2.10)$$

$$= P(Z < 2.10) - (1 - P(Z < 0.035))$$

$$= 0.9821 - .484 = .4981$$

- 예) 어느 회사에서 직원들의 60%가 아파트에 거주하고 있음을 알았다. 회사 직원 20명을 대상으로 아파트 거주여부를 조사하였다.

(a) 아파트에 거주하는 직원의 수(X)가 10명 이상 14명 미만일 확률을 구하라.

$$X \sim B(20, 0.6)$$

$$P(10 \leq X < 14) = P(X \leq 13) - P(X \leq 9)$$

$$= 0.750 - 0.128$$

$$= 0.622$$

(b) 만약 150명의 직원을 대상으로 아파트 거주여부를 조사하였다면, 아파트에 거주하는 직원의 수(Y)가 82명 초과 101명 이하일 확률을 구하라.

$$Y \sim B(150, 0.6) \approx N(90, 36)$$

$$P[82 < Y \leq 101] \approx P[82.5 < Y < 101.5]$$

$$= P\left[\frac{82.5-90}{6} < Z < \frac{101.5-90}{6}\right]$$

$$= P[-1.25 < Z < 1.92]$$

$$= 0.9726 - (1 - 0.8944) = 0.867$$

SPSS를 이용한 실습

- SPSS에 정의된 다음의 함수들을 이용하면 이항확률분포, 정규분포 등의 확률분포의 확률 값을 계산할 수 있다. 관련된 함수들을 정리하면 다음과 같다.

PDF.BINOM(x, n, p) : 모수가 n, p인 이항확률변수가 x값이 되는 확률 값을 반환한다.

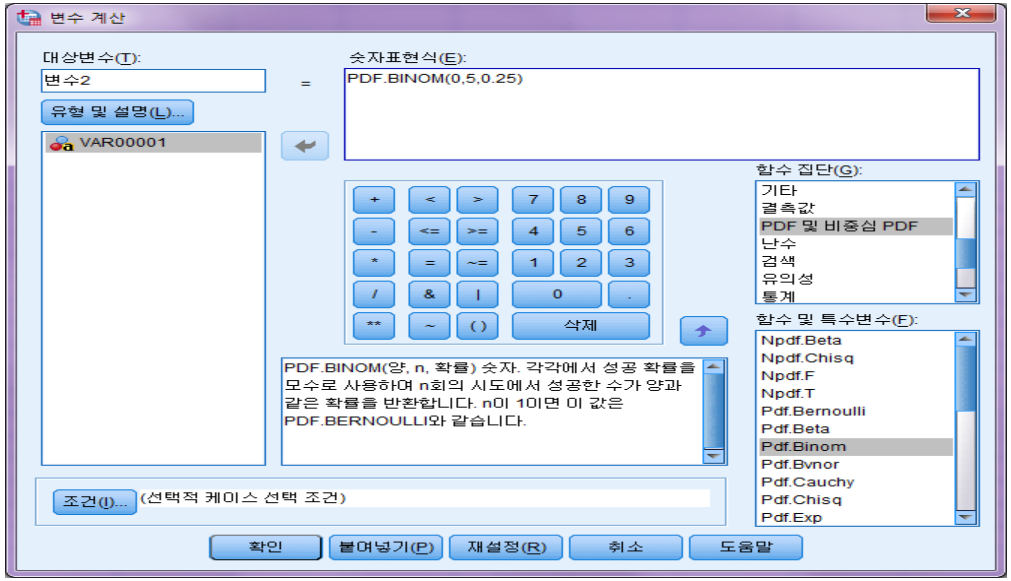
CDF.BINOM(x, n, p) : 모수가 n, p인 이항확률변수가 x값보다 작거나 같을 누적 확률을 반환한다.

CDF.NORM(z) : 표준정규확률변수가 z값보다 작을 확률을 반환한다.

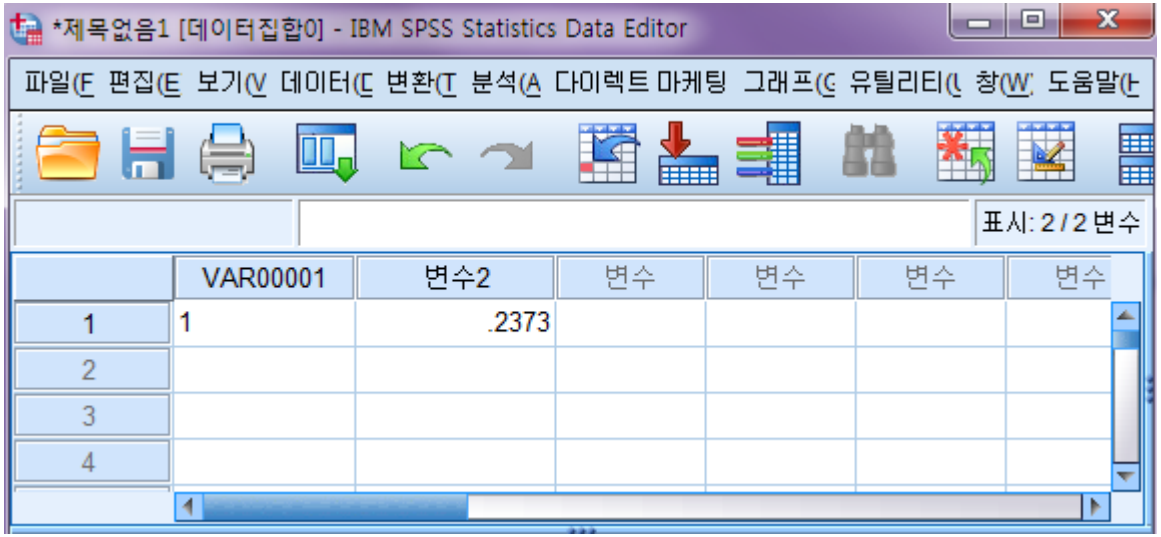
CDF.NORMAL(x, μ, σ) : 평균이 μ이고 표준편차가 σ인 정규확률변수가 x보다 작을 확률값을 반환한다. 따라서 CDF.NORMAL(x, 0, 1)와 CDF.NORM(x)는 동일하다.

PROBIT($1-\alpha$) : 표준정규확률분포의 누적 확률이 $1-\alpha$가 되는 표준정규확률변수의 값을 반환한다. 즉, 본문에서 다루었던 z_α값을 반환한다. 예를 들어 PROBIT(0.95)는 $z_{0.05}$로서 1.645이다.

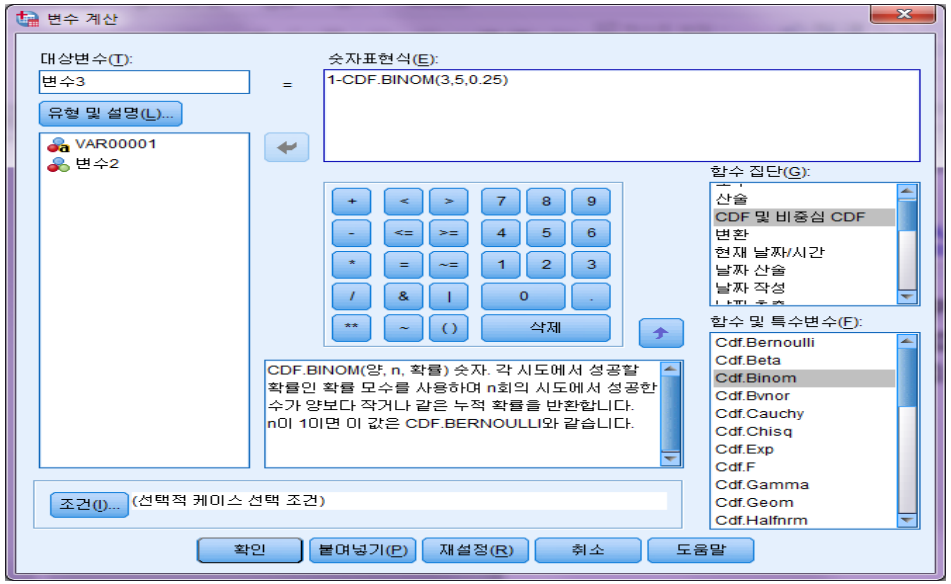
- [예제 4.22] 예제 4.11의 문제를 SPSS를 이용해 풀어보자.
  - 확률변수 X가 n=5이고 p=0.25인 이항확률변수일 때 (a)는 P[X=0] 을 구하는 문제로서 PDF.BINOM(0, 5, 0.25)가 된다. 이 확률 값은 다음의 절차를 통해 얻을 수 있음
  - 변환(T) ▷ 변수 계산(C)
    - 변수계산 창에서 새 변수 "변수2"를 정의하고 숫자 표현식에 해당되는 함수 PDF.BINOM(, ,)를 찾아 옮긴 후 각각의 값을 입력하여 PDF.BINOM(0, 5, 0.25) 을 정한 후 확인을 클릭



- 다음과 같이 같이 데이터 편집기 창에 그 확률 값을 갖는 "변수2"가 생성



- 문제 (b)는 확률 P[X≥4]을 구하는 것으로 이것은 1-CDF.BINOM(3,5,0.25) 로 나타내어진다. 따라서 그림 4.16과 같이 입력하면 확률 값 0.016을 얻을 수 있다.



	VAR00001	변수2	변수3	변수
1	1	.2373	.02	

- [예제 4.23] 예제 4.18의 (a) 문제를 해결해 보자.
  - 확률변수 X가 $N(3, 2^2)$을 따를 때 P[2<X<5] 의 값은 CDF.NORMAL(5,3,2)-CDF.NORMAL(2,3,2)로서 다음 그림 4.17과 같이 입력하면 구할 수 있다.