제 6 장 여러 가지 확률분포

정보통계학과 : 김 덕 기

http://cafe.daum.net/cb-stat

toby123@cbnu.ac.kr

이산형 확률분포(베르누이 시행-이항분포)

베르누이 시행(Bernoulli Trial)

베르누이 시행은 다음의 두 조건을 만족시키는 시행으로 정의된다.

(1) 시행의 결과는 일반적으로 성공(S)과 실패(F)로 구분된다.

(2) 각 시행에서 성공의 확률을 $p = P(S)$, 실패의 확률을 $q = P(F)$라 하면, 두 확률의 합 $p + q = 1$이 된다.

베르누이 분포의 확률분포함수

x	0	1	계
$P(X = x)$	$1-p$	p	1

성공확률이 p인 베르누이 시행이 n번 독립적으로 반복 시행되었을 때 성공 횟수를 나타내는 확률변수를 X라 하면, 이때의 확률분포를 시행횟수 n과 성공확률 p를 갖는 **이항분포(binomial distribution)**라 한다.

이산형 확률분포(이항분포)

이항분포(Binomial Distribution)
확률변수 X가 다음과 같은 확률분포함수를 가질 때 시행횟수 n, 성공률 p인 이항분포에 따른다고 하며, 기호로는 $X \sim B(n, p)$를 사용한다. $$f(x)=\begin{cases}\binom{n}{x}p^x q^{n-x} & , x=0,1,2,\cdots,n \\ 0 & , 그 외\end{cases}$$

($Y_1, \ldots, Y_n$)~베르누이 확률변수

평균 : $\mu = E(X) = E(Y_1 + Y_2 + \cdots + Y_n) = np$

분산 : $\sigma^2 = Var(X) = Var(Y_1 + Y_2 + \cdots + Y_n) = npq$

표준편차 : $\sigma = \sqrt{np(1-p)}$

이산형 확률분포(이항분포:평균,분산)

(예 1) 동전을 다섯 번 던졌을 때 앞면이 두 번 나올 확률을 구하여라.

n=5, x=2, p=1/2 : x~B(n,p)=B(5,1/2)

$$f(x)=\binom{n}{x}p^{x}(1-p)^{n-x}=\binom{5}{2}(1/2)^{2}(1-1/2)^{3}=5/16$$

(예 2) S공장에서 생산된 제품의 불량률이 15%라고 한다. 임의로 100개의 제품을 선택했을 때 불량품의 개수에 대한 기대치와 분산은?

$$\mu=np=100\times 0.15=15(\text{개})$$
$$\sigma^{2}=npq=np(1-p)=100\times 0.15\times 0.85=12.75$$

이항분포표를 이용한 확률계산방법 : (교재 P.86) 예제 6-2 참조.

ex) 한국의 전통 윷놀이 경우

X~B(n, p)

이항분포표

[이항분포표] $P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$

n	x	P								
		0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
	2	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
	5	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313

이산형 확률분포(초기하분포 : 평균, 분산)

앞의 이항분포는 각 시행의 결과가 다른 시행의 결과에 영향을 미치지 않으므로 각 시행의 성공확률이 항상 동일하지만(즉, 복원추출), 초기하분포는 각 시행의 결과가 다른 시행의 결과에 영향을 미쳐서 성공확률이 달라지는 경우(즉, 비 복원추출)에 적용되는 확률분포.

$$X \sim H(x\,;N, n, D),\quad x = 0, 1, \cdots, \min(n, D)$$

초기하분포(Hypergeometic Distribution)

N=모집단의 크기, D=모집단에서 속성 A를 갖는 것의 개수

n=표본의 크기, X= 표본에서 속성 A를 갖는 것의 개수

확률변수 X는 초기하분포를 따르며 다음과 같은 확률분포를 갖는다.

$$P(X=x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}},\quad x = 0, 1, \cdots, n(\text{단},\ n \le D,\quad n \le N-D$$

ex) 우리나라Lotto-당첨확률?

이산형 확률분포(초기하분포 : 예제)

[예제] 흰공 3개와 빨간공 2개가 들어있는 주머니에서 임의로 2개를 꺼냈을 때, X를 빨간공의 개수인 확률변수로 정의하자. X가 0, 1, 2의 값을 취할 확률은?

[풀이] X는 $N=5$, $D=2$, $n=2$인 초기하분포를 따른다. X는 0, 1, 2의 값을 취할 수 있으며 각각의 확률은 다음과 같다.

$$P(X=0)=\frac{\binom{2}{0}\binom{3}{2}}{\binom{5}{2}}=\frac{3}{10}=0.3$$

$$P(X=1)=\frac{\binom{2}{1}\binom{3}{1}}{\binom{5}{2}}=\frac{6}{10}=0.6$$

$$P(X=2)=\frac{\binom{2}{2}\binom{3}{0}}{\binom{5}{2}}=\frac{1}{10}=0.1$$

(교재 P.83) 예제 6-1

이산형 확률분포(포아송분포: 평균,분산)

포아송분포(Poisson Distribution)

확률변수 X를 단위시간이나 단위공간에서 희귀하게 일어나는 사건의 횟수라 하면 X는 포아송분포를 따르며 기호로 $X \sim P(\lambda)$이라 표현한다.

$$f(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad e = 2.718 \quad , \quad x = 0, 1, 2, \cdots$$

λ: 사건의 평균발생횟수

- 평균 : $E(X) = \lambda$
- 분산 : $Var(X) = \lambda$

[EX] 어느 대학의 전화교환수에게 오전 9시부터 10시 사이에 1분당 평균 2회의 전화가 걸려온다고 한다. 전화가 걸려오는 횟수가 포아송 분포의 특성을 갖는다고 하면 1분당 4회의 전화가 걸려올 확률은 얼마? (교재 P.88 예제6-4 참조)

$$f(X=4) = \frac{e^{-2}\,2^{4}}{4!} = 0.0902$$

이항분포의 포아송 근사

시행횟수(=n)이 충분히 크고 성공확률(=p)이 작은(거의 0에 가까운 경우), 이항분포 $B(n,p)$을 따르는 확률변수 X에 대해 $np=\lambda$인 포아송 분포와 거의 유사한 분포를 이룬다. 일반적으로 $np<5$인 이항분포는 아래의 관계가 성립한다.

$$\binom{n}{x}p^x(1-p)^{n-x} \approx f(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

의 근사관계가 성립된다. 따라서 이항분포의 근사분포로서 포아송분포를 정의할 수 있다.

[EX] 불량률이 2%인 생산공정에서 100개의 표본을 임의로 추출할 경우 불량률이 하나도 없을 확률은 얼마인가? (교재 P.88 예제6-5 참조)

(1) 이항분포 : $f(X=0)=\binom{100}{0}(0.02)^0 0.98^{100} = 0.1326$

(2) 포아송분포 : $np=\lambda=2$인 $f(0)=\dfrac{e^{-2}2^0}{0!}=0.1353$

연속형 확률분포-정규분포와 특징-1

➢ C. Gauss 의 정규분포 (Normal distribution), bell distribution

C. Gauss(1777~1855) 는 각종의 물리학 실험을 행할 때 수반되는 계측오차에 대한 확률분포 로서 , 오늘날 가우스분포 (Gauss distribution) 라 알려져 있는 연속확률분포 제시 . 그 후로 이 분포는 물리학 뿐만 아니라 다른 모든 학문 분야에서도 널리 사용

▶ 정규분포: $X \sim N(\mu, \sigma^2)$

X의 확률밀도 함수가

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

이면, 연속확률변수 X가 모수 $\mu(-\infty < x < \infty)$와 $\sigma^2(0 < \sigma^2 < \infty)$를 갖는 정규분포라 한다.

– 정규분포의 성질

① $\int_{-\infty}^{\infty} f(x;\mu,\sigma^2)dx = 1$

② $E[X] = \mu$

③ $Var(X) = \sigma^2$

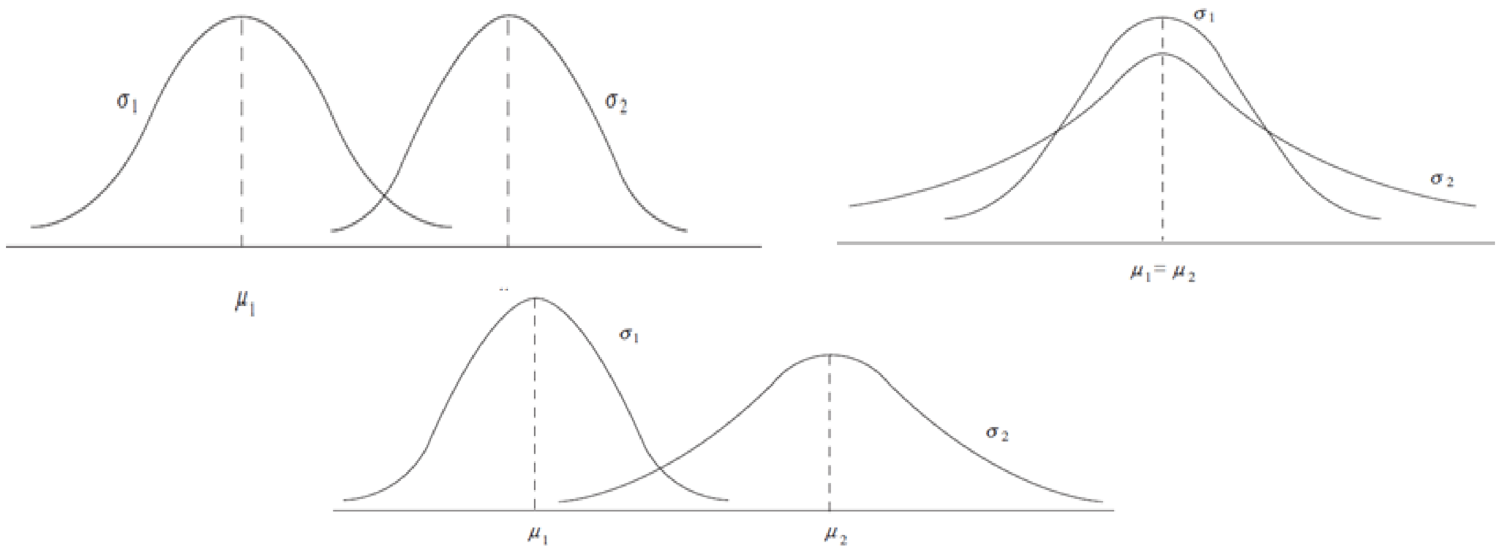
연속형 확률분포-정규분포와 특징-2

정규분포 특징

① 평균 μ대해 대칭이고 종 모양. 즉 종의 중심이 평균이고 중앙값

② 표준편차 σ는 평균 μ에 곡선의 변곡점까지 거리

- σ가 크면, μ에 대해 퍼진 그래프
- σ가 작으면, μ에 대해 뾰족한 그래프

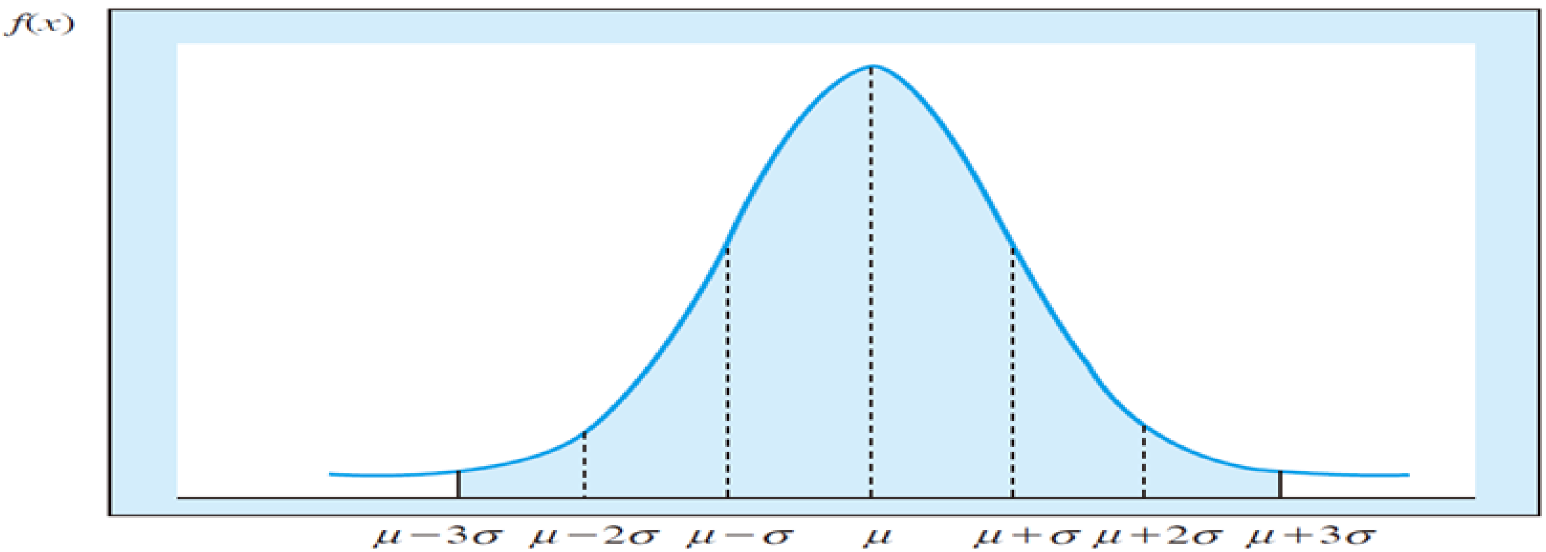


연속형 확률분포-정규분포와 특징-3

– 확률분포 X가 (근사적으로) 정규분포 $N(\mu, \sigma^2)$을 따를 때

$$P(\mu-\sigma<X<\mu+\sigma)=0.683$$
$$P(\mu-2\sigma<X<\mu+2\sigma)=0.954$$
$$P(\mu-3\sigma<X<\mu+3\sigma)=0.997$$

$\boxed{6\sigma-\text{운동}}$



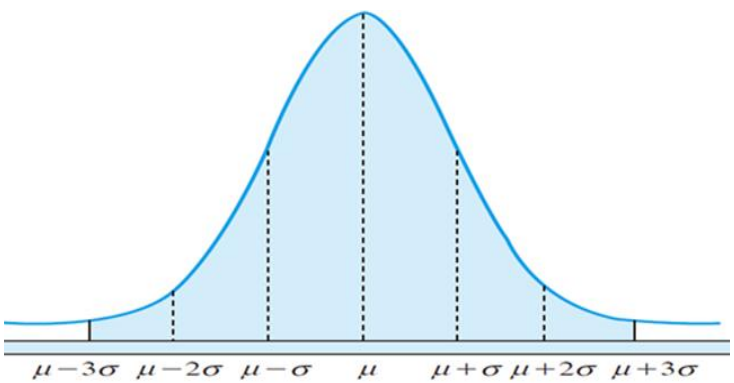
평균으로부터 표준편차의 3배 이내에 있는 면적

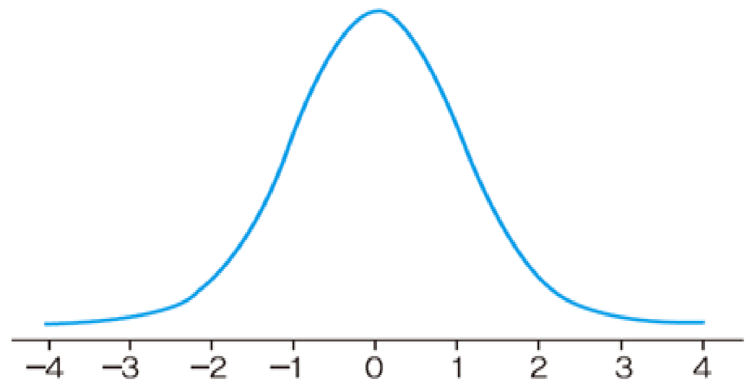
표준정규분포의 특징-1

$X \sim N(\mu, \sigma^2)$ 확률변수

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

$Z \sim N(0,1)$





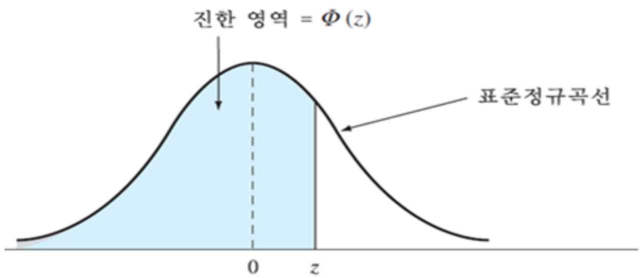
$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

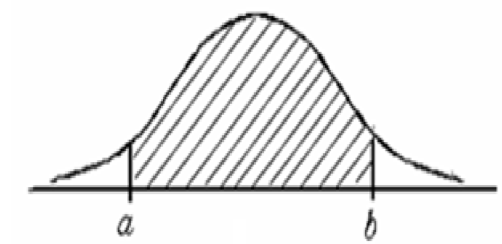
$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

표준정규분포의 특징-2

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} f(y; 0, 1)\,dy$$

$$P(a < X < b) = P(a \leqq X < b) = P(a < X \leqq b) = P(a \leqq X \leqq b)$$





(EX) 서울의 강우량은 평균이 46.0cm이고 표준편차가 15.2cm인 정규분포를 이룬다면, 강우량이 9.2cm보다 적을 확률은?

(P.315) 에 표준정규분포에 대한 누적확률 $\Phi(z) = P(Z \le z)$

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right), \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$= P(z_1 < Z < z_2), \quad Z = \frac{X-\mu}{\sigma}, z_1 = \frac{a-\mu}{\sigma}, z_2 = \frac{b-\mu}{\sigma}$$

$$= \Phi(z_2) - \Phi(z_1)$$

$$Z = \frac{X-\mu}{\sigma} = \frac{9.2-46.0}{15.2} = -2.42$$

$$P(X \leqq 9.2) = P(Z \leqq -2.42)$$

$$= 0.0078$$

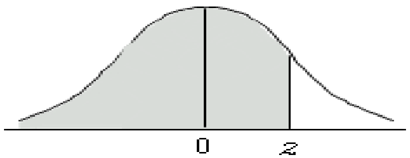
$$\Phi(0) = P(Z < 0) = 0.5$$

$$\Phi(-z) = P(Z < -z) = 1 - \Phi(z) = 1 - P(Z \le z)$$

표준 정규분포표

[표준정규분포표] $P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$

표의 숫자는 z보다 작거나 같을 확률을 나타낸다.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5000	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5398	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5793	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6179	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6554	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6915	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7257	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7580	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7881	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8159	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8413	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8643	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8849	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9032	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9192	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9332	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9452	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9554	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9641	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9713	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9772	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9821	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9861	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9893	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9918	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9938	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9953	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9965	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9974	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9981	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9990	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

정규분포의 확률계산-예제

● [예 제] 평균이 18이고 표준편차가 3인 정규분포를 따르는 확률변수 X에 대하여 다음을 구하여라.

(1) $P(X<15)$

(2) $P(12<X<21)$

(3) $P(X<k) = 0.3050$ 을 만족하는 k 값

풀이 $X \sim N(18, 3^2)$ 이므로 $\mu = 18$, $\sigma = 3$ 이다.

$$(1)\ P(X<15) = P\left(Z<\frac{15-18}{3}\right) = P(Z<-1)$$

$$= \Phi(-1) = 0.1587$$

$$(2)\ P(12<X<21) = P\left(\frac{12-18}{3}<Z<\frac{21-18}{3}\right) = P(-2<Z<1)$$

$$= \Phi(1) - \Phi(-2) = 0.8413 - 0.0228 = 0.8185$$

1−0.305=0.695

(3) $P(X<k) = P\left(Z<\dfrac{k-18}{3}\right) = 0.3050$이므로 [부록 표 3]에서 확률값 0.3050값을 갖는 z값은 $z = -0.51$이다. 즉, $\dfrac{k-18}{3} = -0.51$이므로 $k = 16.47$ 이다.

이항분포의 정규분포근사 1

이항분포의 확률을 구하는 산술적 계산은 n이 어느 정도 커지면 매우 복잡하고 시간이 많이 걸림.

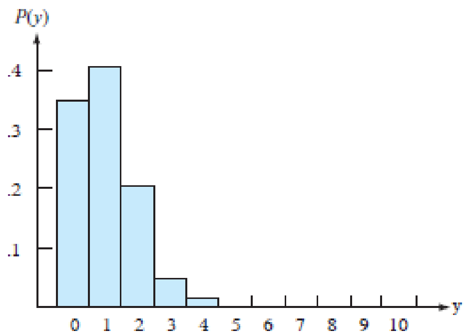
예를 들어, $n=20, p=0.4$일 때만 해도 $P(X \geq 10)$을 구하려면,

$$\begin{aligned} P(X \geq 10) &= P(X=10)+P(X=11)+\cdots+P(X=20) \\ &= \frac{20!}{10!10!}(0.4)^{10}(0.6)^{10}+\frac{20!}{11!9!}(0.4)^{11}(0.6)^{9} \\ &+\cdots+\frac{20!}{20!0!}(0.4)^{20}(0.6)^{0} \end{aligned}$$

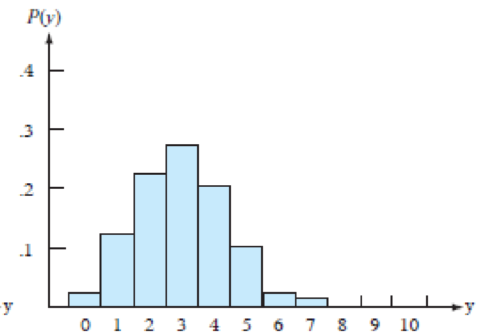
이항분포의 확률은 시행횟수 n이 크고 각 시행의 성공확률 P가 ½ 에 가까운 값일 때 정규분포를 이용하여 근사적으로 계산할 수 있다.

(교재 P. 96 의 {그림 6-4}참조.)

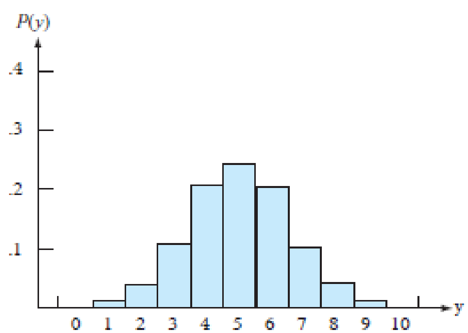
이항분포의 정규분포근사 2



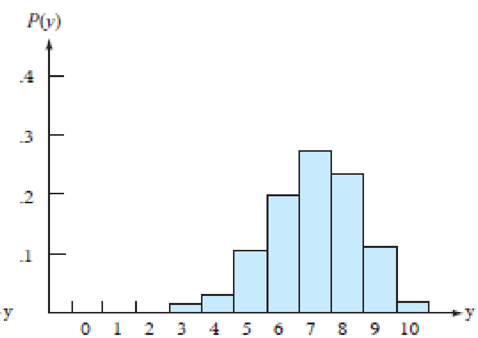
(a) $p = 0.1$



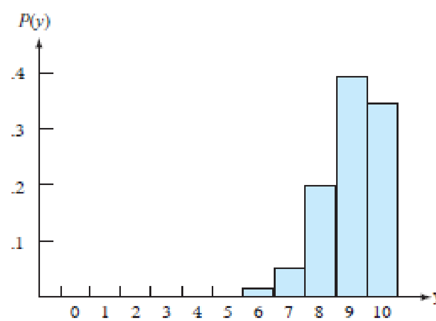
(b) $p = 0.3$



(c) $p = 0.5$



(d) $p = 0.7$



(e) $p = 0.9$

n=10일 때 p=0.1~0.9에 따른 이항분포 ➔ n(커지고), p~1/2에 가까우면 정규분포에 가까워짐.

이항분포의 정규분포근사 3

Laplace의 극한 정리

충분히 큰 n과 0과 1에 가깝지 않는 p에 대하여, 이항분포의 확률모형

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (0 < p < 1,\ x = 0, 1, 2, \cdots, n)$$

은 근사적으로 정규분포곡선

$$\frac{1}{\sqrt{2\pi}\ \sqrt{np(1-p)}} exp\left[-\frac{1}{2}\left(\frac{x-np}{\sqrt{np(1-p)}}\right)^2\right]$$

과 일치한다. 이 정리를 De Moivre-Laplace의 극한정리라 한다.

$X \sim B(n, p)$이면, n이 큰 경우에는

$$P(a < X < b) = P(a - np < X - np < b - np)$$

$$= P\left(\frac{a-np}{\sqrt{np(1-p)}} < \frac{X-np}{\sqrt{np(1-p)}} < \frac{b-np}{\sqrt{np(1-p)}}\right)$$

$$\simeq \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right)$$

이항분포의 정규분포근사 4

(EX1) 어떤 학생이 10개의 OX문제를 단순히 추측에 의해 풀 경우 일곱 문제 또는 여덟 문제를 맞힐 확률을 구해보자.

이항분포에서의 계산 :

$n=10, p=0.5, x=$ 이항확률변수

$P(x)=\binom{10}{x}(0.5)^x(1-0.5)^{10-x}, P(x=7 \text{ 또는 } 8)=P(7)+P(8)=0.1172+0.0439=0.16$

정규분포의 근사계산 :

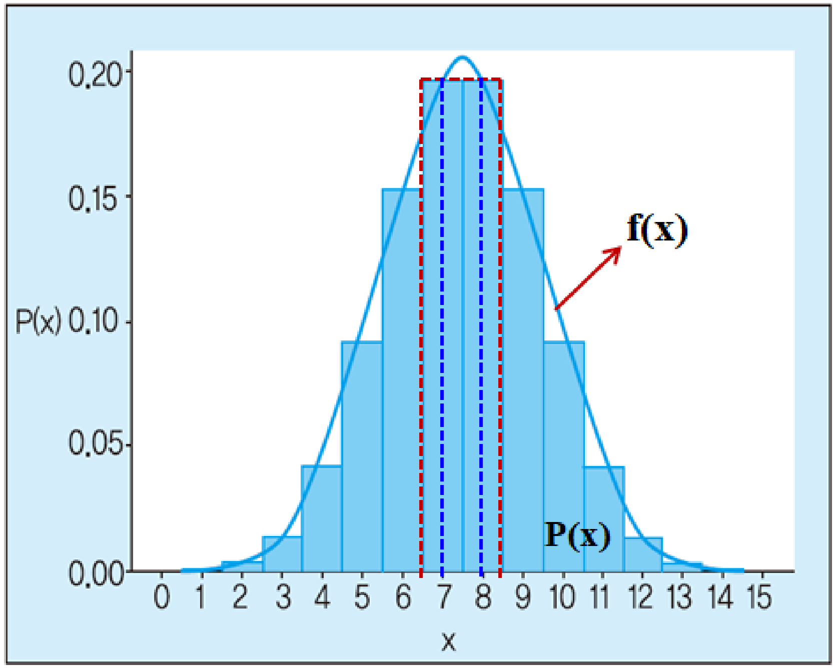
$$\mu=np=10(0.5)=5,\ \sigma=\sqrt{np(1-p)}=\sqrt{10(0.5)(1-0.5)}=1.58$$

$$P(7 \le x \le 8) \approx P\left(\frac{7-5}{1.58} \le Z \le \frac{8-5}{1.58}\right)=P(1.27 \le Z \le 1.90)$$
$$=\phi(1.9)-\phi(1.27)=0.9713-0.8980=0.0733$$

이항분포와 정규분포의 관계

표 $n=15$이고 $p=0.5$일 때의 이항확률분포표

X	$P(X=x)$
0	0.000031
1	0.000458
2	0.003204
3	0.013885
4	0.041656
5	0.091644
6	0.152740
7	0.196381
8	0.196381
9	0.152740
10	0.091644
11	0.041656
12	0.013885
13	0.003204
14	0.000458
15	0.000031



[표]의 확률분포

이항분포의 정규분포근사-연속성수정

연속성의 수정(correction for continuity) : 교재 P.97 (그림 6-5)참조.

$$P(a < X < b) = P\left(\frac{(a-0.5)-np}{\sqrt{np(1-p)}} < Z < \frac{(b+0.5)-np}{\sqrt{np(1-p)}} \right)$$

$$\simeq \Phi\left(\frac{(b+0.5)-np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{(a-0.5)-np}{\sqrt{np(1-p)}} \right)$$

$$P(7 \leqq X \leqq 8) = P\left(\frac{(7-0.5)-5}{1.58} < Z < \frac{(8+0.5)-5}{1.58} \right)$$

$$= P(0.95 < Z < 2.22) = 0.9868 - 0.8289 = 0.1579$$

R-이항분포의 정규근사

[EX] 이항분포의 정규근사 : B(n,p) → n=10, 15, 20, 30 ; p=0.2~0.5로 하여 400개의 난수를 발생시켜 히스토그램을 그려 확인하여라.

```
x1=rbinom(400, 10, 0.2)
x2=rbinom(400, 15, 0.3)
x3=rbinom(400, 20, 0.4)
x4=rbinom(400, 30, 0.5)
par(mfrow=c(2,2))
hist(x1) ; hist(x2); hist(x3); hist(x4)
```