

2022학년도 2학기

경영 통계학

담당교수: 백수정



학습 목표

1. 연속확률변수의 정의를 이해한다
2. 연속균등분포의 확률을 계산할 수 있다
3. 정규분포의 모양과 모수를 이해한다
4. 엑셀이나 확률분포표를 이용하여 주어진 z 또는 x 에 해당하는 정규분포 확률과 확률에 해당하는 값을 구할 수 있다
5. 주어진 지수 분포의 확률에 해당하는 x 값을 구할 수 있다

연속확률변수(discrete random variable)

- 연속확률변수는 다음 고객이 도착할 때까지 걸린 시간과 같은 측정에 관한 것
- 비정수(noninteger) 값을 범위로 가짐
- 확률은 확률분포함수(probability density function, PDF) 곡선의 아래 면적으로 정의함
- 연속확률변수의 확률은 $P(53.5 \leq X \leq 54.5)$ 또는 $P(X < 54)$ 또는 $P(X \geq 53)$ 와 같이 구간에 대해 정의함

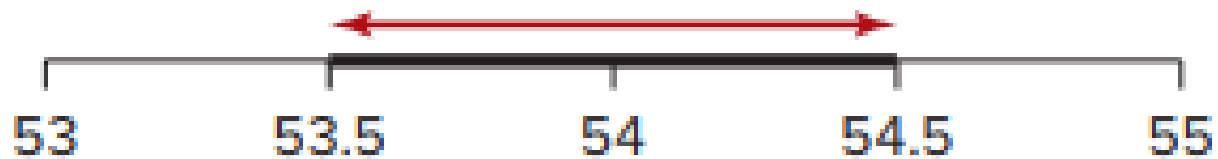
연속확률변수(discrete random variable)

- 이산확률변수와 연속확률변수의 차이

이산형 변수: 각 점에서 정의된다

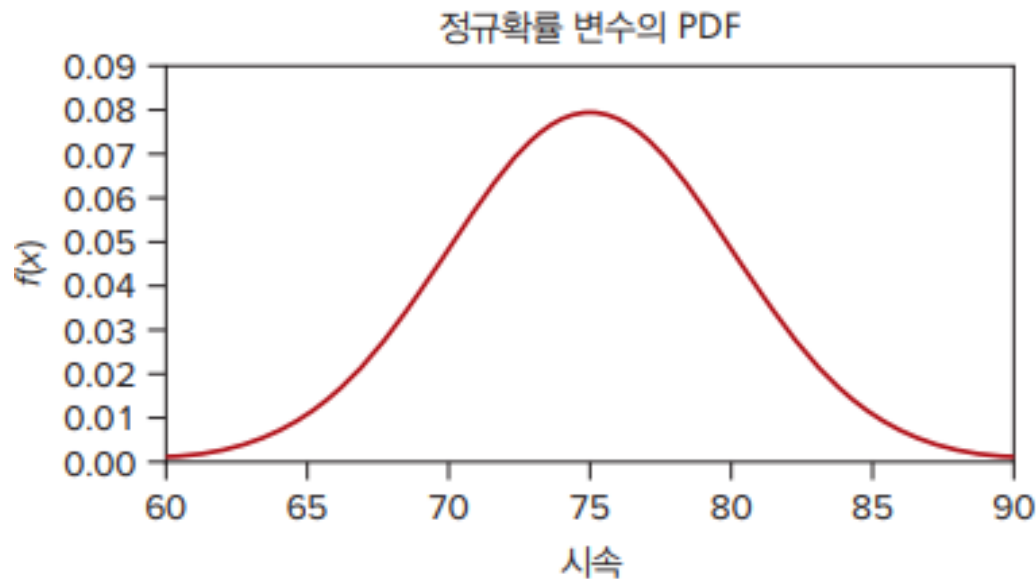
0 1 2 3 4 5

연속형 변수: 특정 구간에 대해서 정의된다



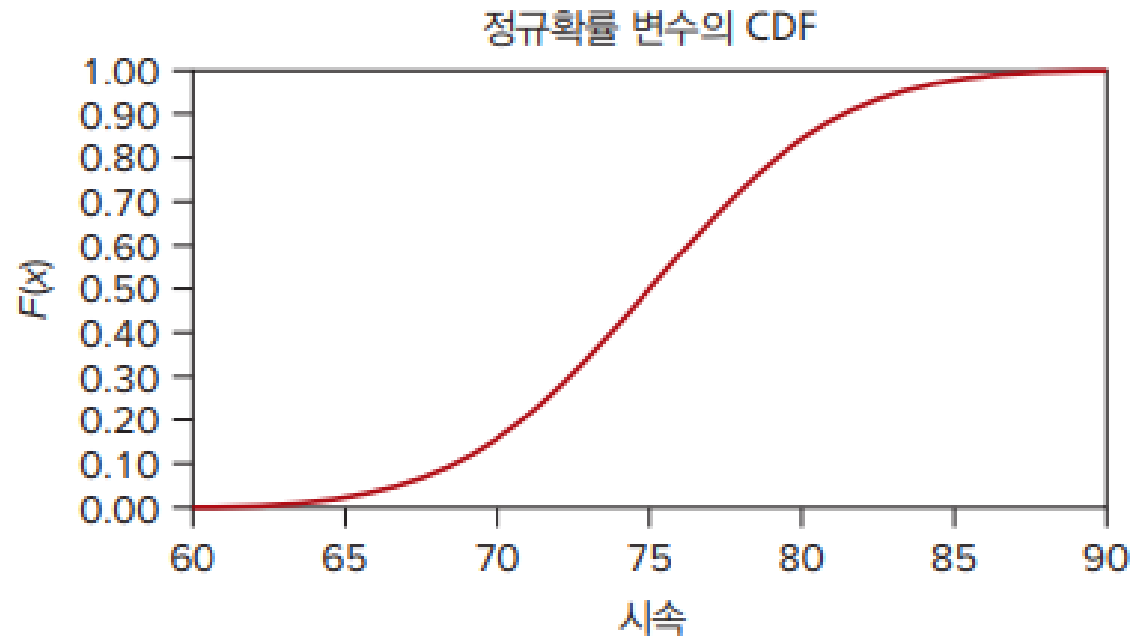
확률분포함수(PDF)

- 연속 확률분포함수는 $f(x)$ 로 나타냄
- 연속 PDF는 음의 값을 갖지 않음
- PDF 아래 면적의 합은 반드시 1
- 평균, 분산, 분포의 모양은 PDF 모수에 의존함



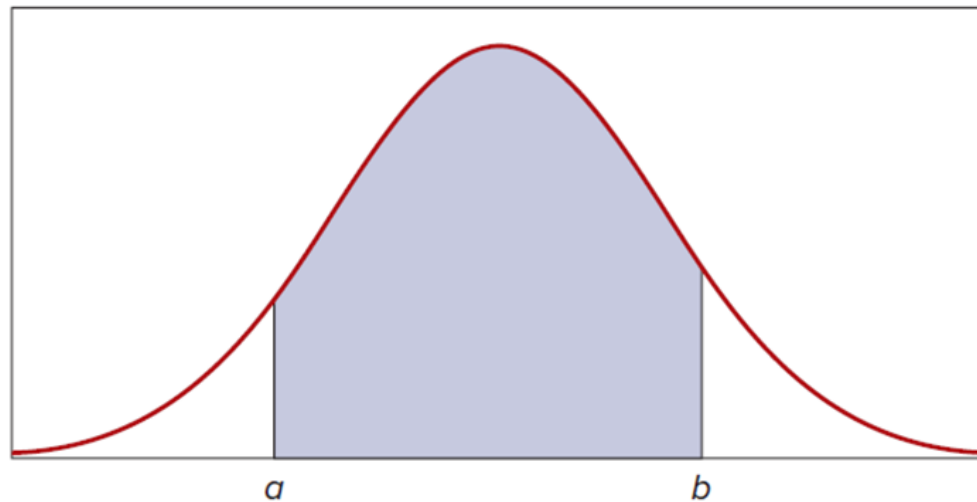
누적확률분포함수

- 연속 누적확률분포함수는 $F(x)$ 로 나타냄
- 누적면적인 $P(X \leq x)$ 를 의미함
- 확률 계산 시 유용함



면적으로서의 확률

- 연속확률분포함수: 이산확률분포와는 달리, 특정한 점에서의 확률은 0이 됨
- 어떤 PDF에서도 곡선 아래 면적의 합은 1이 됨
- 평균은 전체 분포에 대한 균형점



- 곡선 아래의 구간 a에서 b까지의 구간을 $P(a < X < b)$ 로 나타냄

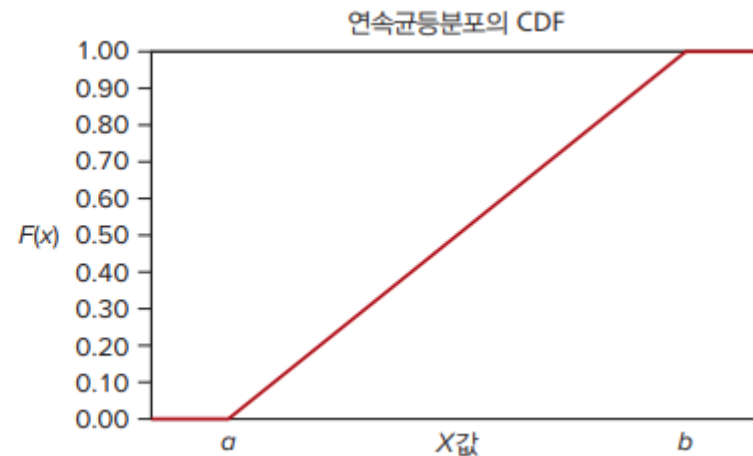
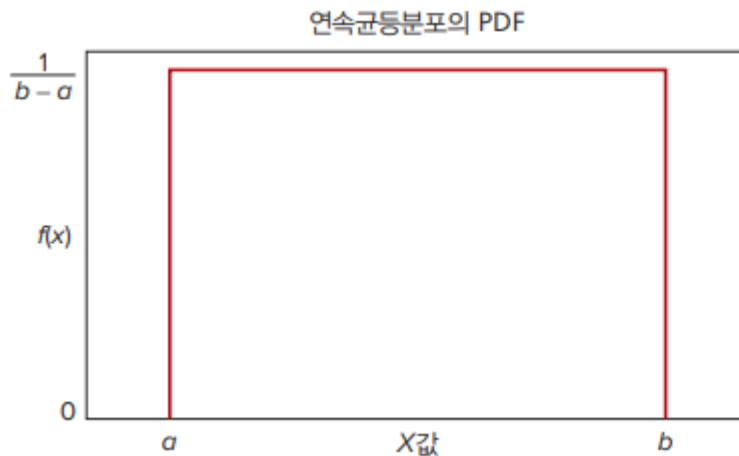
기댓값과 분산

- 연속확률변수의 평균과 분산은 이산확률변수의 $E(X)$ 및 $\text{Var}(X)$ 와 유사함
단, 적분기호 \int 이 \sum 부호를 대신함. 적분은 X 의 전체 범위에 대해 행해 짐
- 평균은 여전히 전체 분포에 대한 균형점이며 분산은 평균과의 편차 제곱을
가중평균한 것임

	연속확률변수	이산확률변수
기댓값	$E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$	$E(X) = \mu = \sum_{\text{모든 } x} xP(x)$
분산	$\text{var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$	$\text{var}(X) = \sigma^2 = \sum_{\text{모든 } x} [x - \mu]^2 P(x)$

연속균등분포

- X 가 구간 a 와 b 사이에 균등하게 분포하는 확률변수라면, 아래 그림에서 보듯이 PDF는 높이가 일정함
- 연속균등분포(uniform continuous distribution)는 $U(a, b)$ 로 표기됨
- 아래 면적의 합이 밑변 $(b - a)$ 에 높이 $1/(b - a)$ 를 곱하면 1이 됨



예제: 마취제의 효과

- 치과의사는 치아를 뽑기 전에 마취제를 주사한다. 환자의 다양한 특성이 주어졌을 때, 치과의사는 마취 효과가 유지되는 시간을 15분에서 30분 사이의 값을 취하는 균등분포로 간주한다. 간단한 표기로 $X \sim U(15, 30)$ 를 쓰자.
- $a = 15$ 와 $b = 30$ 으로 정하면 우리는 평균과 표준편차를 아래와 같이 계산할 수 있다.

$$\mu = \frac{a + b}{2} = \frac{15 + 30}{2} = 22.5 \text{ 분}$$

$$\sigma = \sqrt{\frac{(b - a)^2}{12}} = \sqrt{\frac{(30 - 15)^2}{12}} = 4.33 \text{ 분}$$

예제: 마취제의 효과

- 사건의 확률은 단지 전체에서 차지하는 비중으로 나타나는 구간 넓이이다. 따라서 c에서 d분 사이의 값을 취할 확률은 다음과 같다.

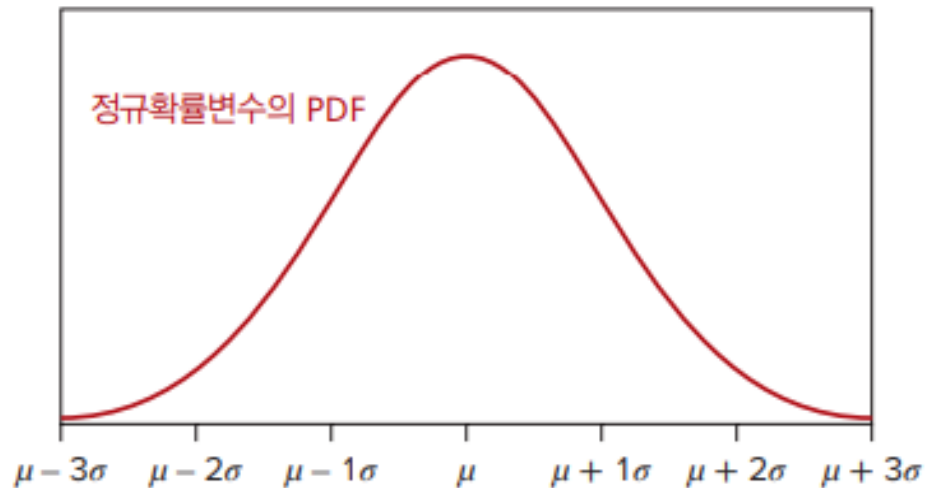
$$P(c < X < d) = (d - c) / (b - a) \quad (\text{균등분포 모형에서 c에서 d 사이의 면적})$$

예를 들어, 마취효과가 20분에서 25분 유지될 확률은 다음과 같다.

$$P(20 < X < 25) = (25 - 20) / (30 - 15) = 5 / 15 = 0.3333, \text{ or } 33.33\%.$$

정규분포

- * 정규(normal) 또는 가우시안 분포(Gaussian distribution)의 특징
 - 두 모수인 μ 와 σ 에 의해서 정의되며, $N(\mu, \sigma)$ 로 표기
 - 정규확률변수 X 의 범위는 $-\infty < x < +\infty$
 - 실제로 대부분(99.7%)의 값은 $[\mu - 3\sigma, \mu + 3\sigma]$ 에 존재함
 - 정규분포는 대칭적인 분포이며 평균주변에서 단봉 분포(unimodal)임



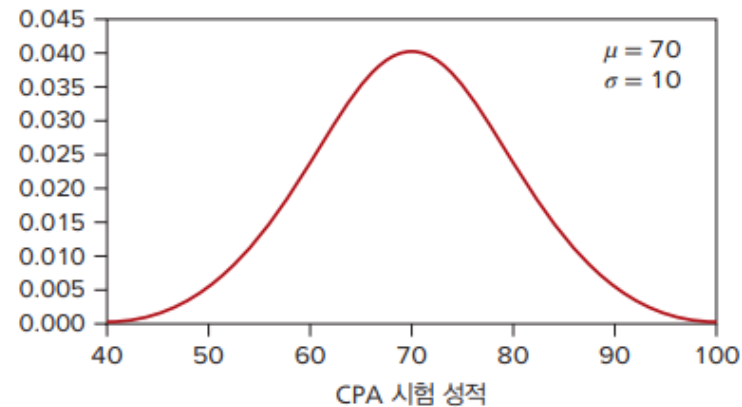
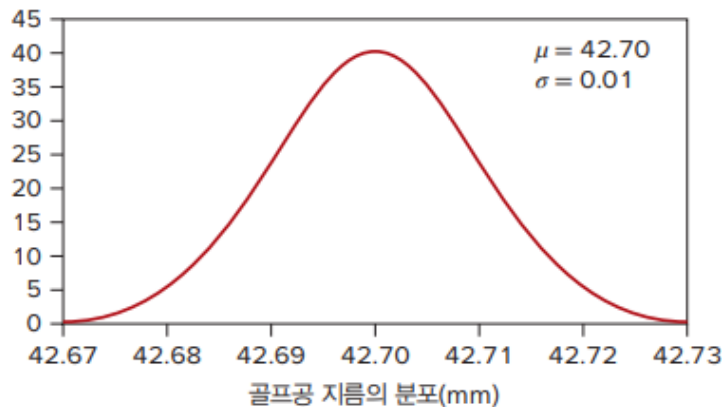
정규분포의 특징

모수	μ = 모평균 σ = 모표준편차
PDF	$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$
정의역	$-\infty < X < +\infty$
평균	μ
표준편차	σ
분포 모양	대칭적이고 종 모양의 대형분포
엑셀에서 PDF*	=NORM.DIST($x, \mu, \sigma, 0$)
엑셀에서 CDF*	=NORM.DIST($x, \mu, \sigma, 1$)
엑셀에서 난수 생성	=NORM.INV(RAND(), μ, σ)

정규분포

- 모든 정규분포는 축의 눈금만 달라질 뿐 동일한 모양을 가짐
- 아래 왼쪽 그래프는 생산되는 골프 공 지름에 대한 분포. 골프공 지름은 평균 $\mu = 42.70$ mm와 표준편차 $\sigma = 0.01$ mm를 가진 $N(42.70, 0.01)$ 인 정규분포를 나타냄
- 오른쪽 그래프는 CPA시험 성적의 분포를 보여주는데, 시험성적은 평균 $\mu = 70$ 와 표준편차 $\sigma = 10$ 을 가진 $N(70, 10)$ 인 정규분포를 따름

척도를 제외하고 모든 정규분포는 거의 유사한 모양



정규분포

* 확률변수가 정규성을 갖기 위한 특징

- 연속적인 눈금으로 측정되어야 함
- 분명한 중심 경향을 가지고 있어야 함
- 단봉(single peak)을 가지고 있어야 함
- 점차 줄어드는 꼬리 모양을 가지고 있어야 함
- 평균을 중심으로 대칭임(양쪽 꼬리 모양이 동일)

* 정규분포를 따를 것으로 예상되는 변수들

- X = 2리터 다이어트 펩시콜라의 양
- X = 보잉 777기의 운항 중 조종사의 왼쪽 귀에서 들리는 조종실의 소음 수준
- X = 제조된 볼베어링의 지름(단위: 밀리미터)

표준정규분포(standard normal distribution)

- μ 와 σ 값에 따라 다양한 종류의 정규분포가 존재할 수 있는데, 평균을 빼고 표준편차로 나누어 줌으로써 표준화된 확률변수로 변환 가능

$$z = \frac{x - \mu}{\sigma} \quad (\text{각 } x\text{값을 } z\text{값으로 변환})$$

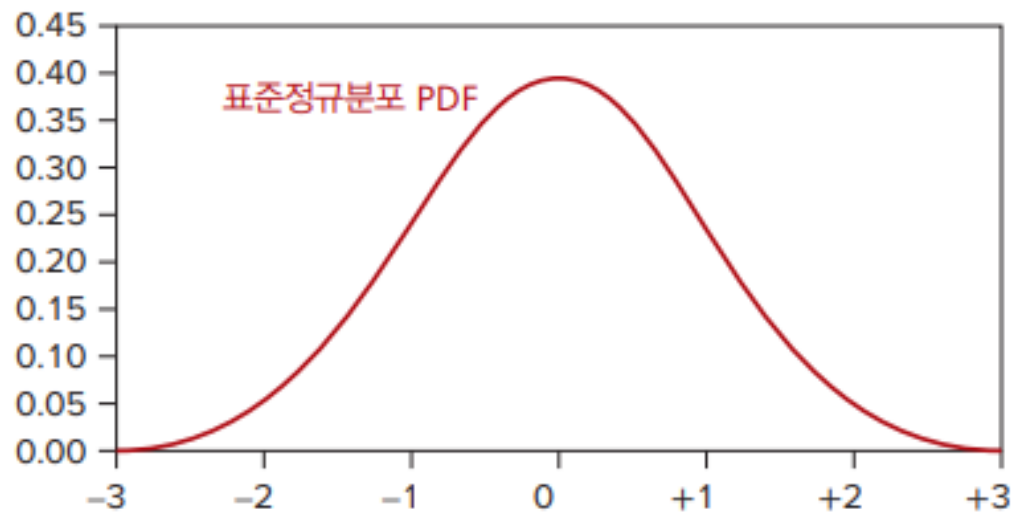
- 만약 X 가 정규분포를 따른다면 변환된 Z 는 평균이 0이고 표준편차가 1인 표준정규분포(standard normal distribution)를 따르게 된다.
- 그리고 $Z \sim N(0, 1)$ 으로 표기
- $f(z)$ 의 모양은 0(평균)에서 가장 높고 변곡점은 ± 1 (표준편차)

표준정규분포(standard normal distribution)

모수	μ = 모평균 σ = 모표준편차
참고	정규분포 CDF는 간단한 공식이 없기 때문에 면적을 계산하기 위해서는 엑셀이나 표준정규분포 표를 이용해야 한다.
PDF	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ 여기서 $z = \frac{x - \mu}{\sigma}$
정의역	$-\infty < Z < +\infty$
평균	0
표준편차	1
분포모양	대칭적이고 종 모양의 대형분포
엑셀에서 PDF*	=NORM.S.DIST(z,0)
엑셀에서 CDF*	=NORM.S.DIST(z,1)
엑셀에서 난수 생성	=NORM.S.INV(RAND())
참고	정규분포 면적 계산에는 간단한 공식이 없기 때문에 엑셀이나 표준정규분포 표를 이용해야 한다.

정규분포

- 변환된 모든 정규분포는 같은 모양을 가지고 있기 때문에 축의 눈금도 서로 같음
- 아래 그림에서 가로축은 -3 에서 $+3$ 까지이며, $f(z)$ 는 확률분포함수이기 때문에 곡선 아래 면적은 반드시 그 합이 1이 됨



정규분포

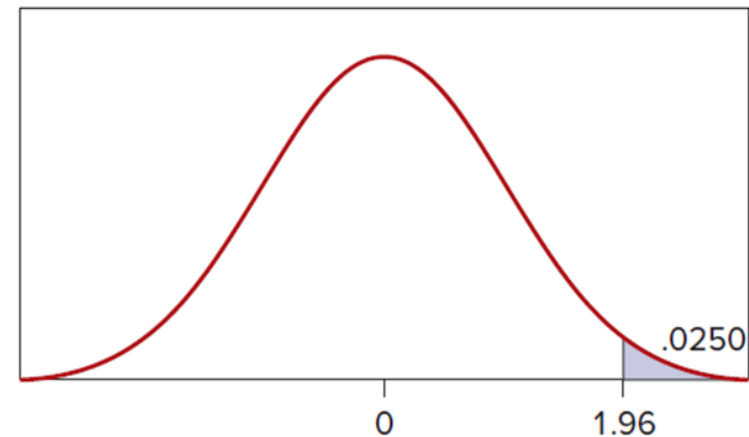
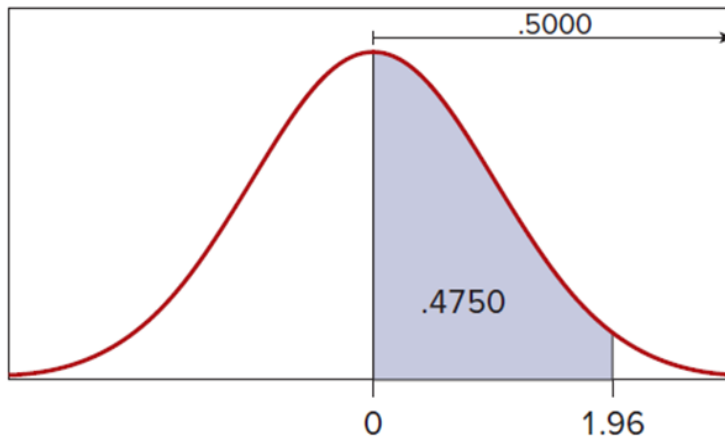
- 부록 C-1는 0부터 특정한 값 z 까지 0.01씩 증가하면서 면적을 나타냄
- 예를 들어, $P(0 < Z < 1.96)$ 을 찾기 위해서는 행에서 $z = 1.9$ 를 택하고 열에서 0.06을 선택($1.96 = 1.9 + 0.06$)
- 이 열과 행은 다음 페이지의 표에서 짙게 표시되어 있다. 표시된 행과 열의 교차점에서 우리는 $P(0 < Z < 1.96) = 0.4750$ 임을 알 수 있음

정규분포

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.6	.49984	.49985	.49985	.49986	.49986	.49987	.49987	.49988	.49988	.49989
3.7	.49989	.49990	.49990	.49990	.49991	.49991	.49992	.49992	.49992	.49992

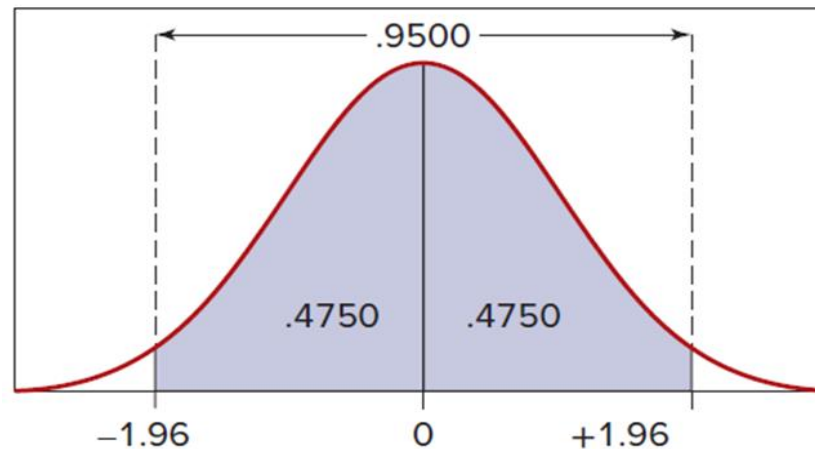
정규분포

- 그 면적은 아래 그림과 같음. 전체 면적의 절반(즉 0.5)이 평균의 오른쪽에 있기 때문에, 오른쪽 꼬리 부분의 면적을 다음과 같이 계산할 수 있음
- 예를 들어, $P(Z > 1.96) = .5000 - P(0 < Z < 1.96) = .5000 - .4750 = .0250$.



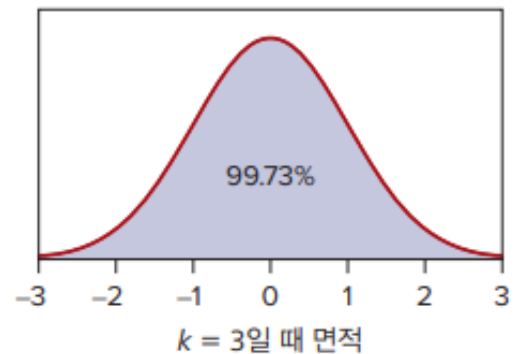
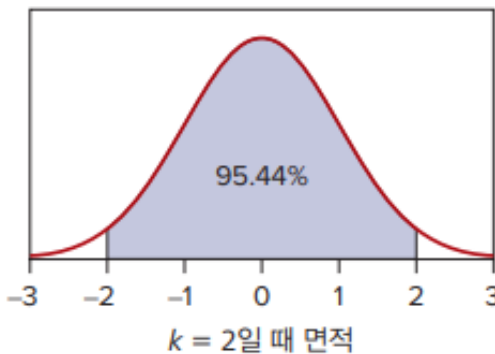
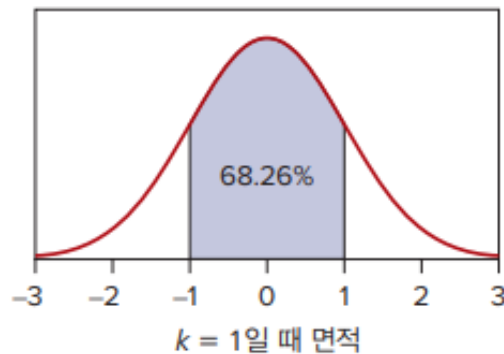
정규분포

- 예: $P(-1.96 < Z < 1.96)$ 과 같은 중간에 있는 면적을 계산하고자 한다고 가정하자. 정규분포는 대칭적이며, 또한 $P(-1.96 < Z < 0) = 0.4750$ 임을 알고 있다.
- 따라서 구간 $-1.96 < Z < 1.96$ 은 정규분포 곡선 아래 전체 면적의 약 95%를 포함하고 있는 것을 알 수 있다. 아래 그림은 이러한 계산을 보여준다.



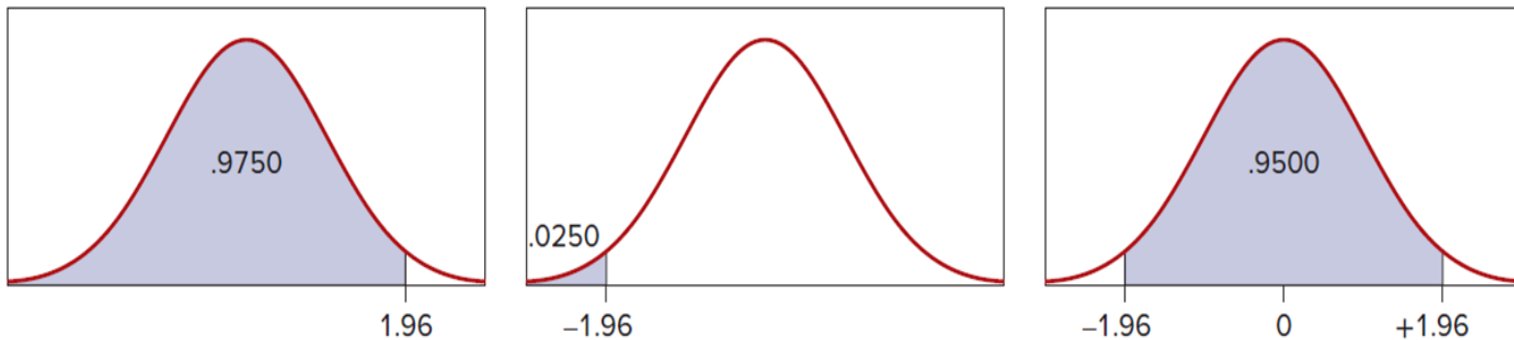
정규분포

- 1표준편차($\pm 1\sigma$) 구간은 전체 면적의 약 68%
- 2표준편차($\pm 2\sigma$) 구간은 전체 면적의 약 95%
- 3표준편차($\pm 3\sigma$) 구간은 전체 면적의 약 99%



정규분포

- 책의 부록 C-1에서 우리는 0부터 z 정규분포 곡선 아래 전체 면적을 구할 수 있음
- 책의 부록 C-2는 왼쪽 끝에서부터 특정한 값 z 까지 누적확률을 나타냄 (엑셀과 유사)

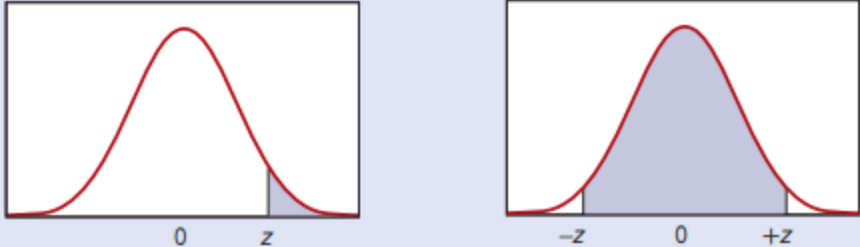


면적이 주어졌을 때 z값 찾기

- 주어진 면적에 해당하는 z-값을 찾는 데 부록 C-1과 C-2의 표를 활용할 수 있음
- 예를 들어 정규분포의 최상위 1%의 면적에 해당하는 z-값은 얼마인가?
- 0.4900에 해당하는 면적을 부록 C-1에서 확인하면 $z = 2.33$ 이 면적 0.4901으로 49%의 값과 거의 유사함을 알 수 있음

면적이 주어졌을 때 z 값 찾기

- 비슷한 방법으로 몇 가지 중요한 면적을 찾을 수 있음. 상위 25%, 10%, 5%, 1% 또는 중간 50%, 중간 90%, 중간 99% 등에 관심이 있기 때문에 자주 사용되는 값을 기억하는 것은 편리함

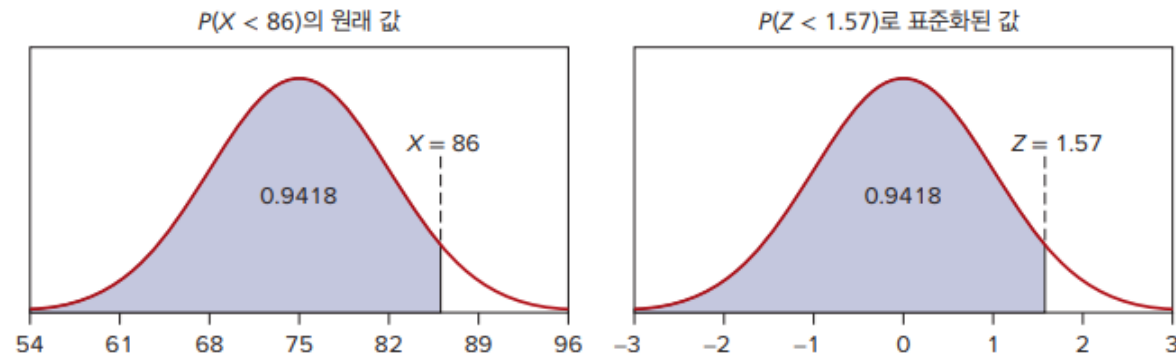
z		
	오른쪽 꼬리면적	중간 면적
0.675	0.255	0.50
1.282	0.105	0.80
1.645	0.055	0.90
1.960	0.025	0.95
2.326	0.015	0.98
2.576	0.005	0.99

면적이 주어졌을 때 z값 찾기

- 존(John)은 경제학 시험에서 86점을 받았다. 반 평균은 75점이고 표준편차는 7점이다. 존의 상대적 위치는? 즉, $P(X < 86)$ 을 구하는 것과 같다. 먼저 존의 표준화된 Z값을 계산할 필요가 있다.

$$z_{\text{John}} = \frac{x_{\text{John}} - \mu}{\sigma} = \frac{86 - 75}{7} = \frac{11}{7} = 1.57$$

- 이 결과는 존의 성적이 평균보다 1.57 표준편차만큼 높다는 의미이다. 부록 C-2표에서 우리는 $P(X < 86) = P(Z < 1.57) = 0.9418$ 임을 알 수 있다. 따라서 존은 대략 94% 정도에 위치한다. 그의 성적이 전체 반 학생수의 94%보다 좋은 성적임을 의미한다.



면적이 주어졌을 때 z값 찾기

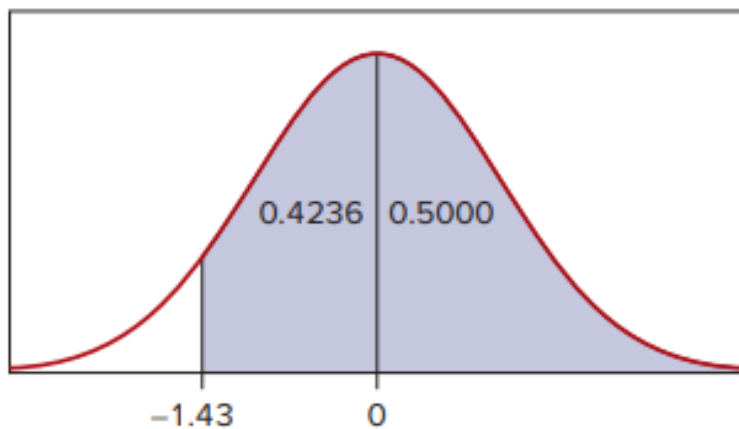
- 이 시험에서 무작위로 추출된 학생 한 명의 점수가 최소 65점일 확률은 얼마인가?
표준화를 통해서 그 확률을 계산할 수 있다

$$z = \frac{x - \mu}{\sigma} = \frac{65 - 75}{7} = \frac{-10}{7} = -1.43$$

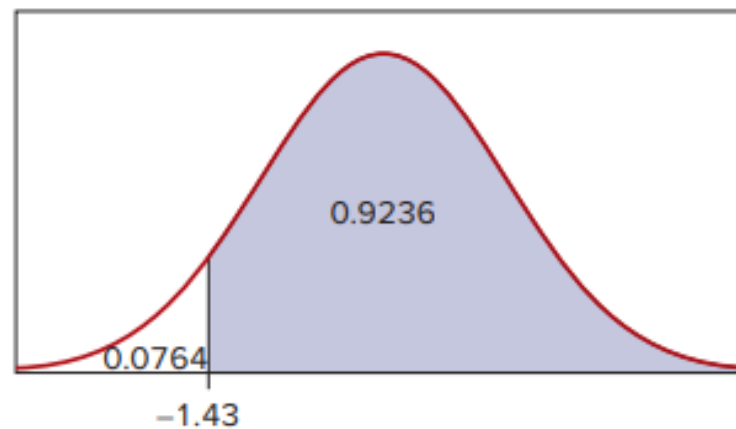
- 부록 C-1표를 이용하여 $P(X \geq 65) = P(Z \geq -1.43)$ 임을 알고 있다. 따라서
 $P(Z \geq -1.43) = P(-1.43 < Z < 0) + 0.5000 = 0.4236 + 0.5000 = 0.9236$ 또는 92.4%
- 부록 C-2표를 이용하면 $P(X \geq 65) = P(Z \geq -1.43)$ 을 다음과 같이 계산가능하다.
 $P(Z \geq -1.43) = 1 - P(Z < -1.43) = 1 - 0.0764 = 0.9236$ 또는 92.4%

면적이 주어졌을 때 z값 찾기

부록 C-1을 이용하여 면적 계산



부록 C-2를 이용하여 면적 계산



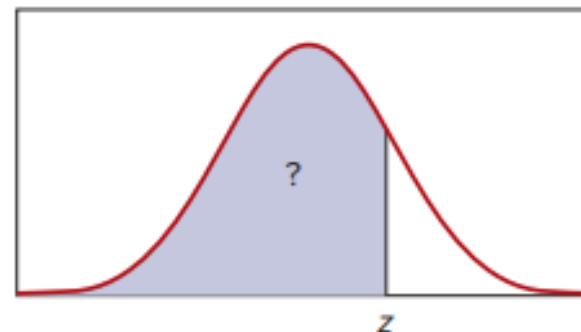
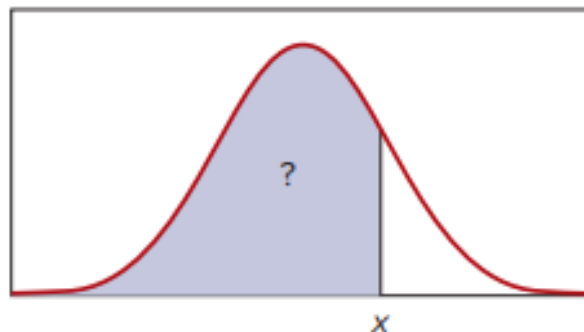
면적이 주어졌을 때 z값 찾기

- 어떻게 역 정규분포(inverse normal)로 알려진 정규분포의 확률위치(5%, 10%, 25%, 75%, 90%, 95% 등)를 어떻게 찾아낼 수 있을까?
- 표준화 공식을 이용함

$$x = \mu + z\sigma \quad \left(z = \frac{x - \mu}{\sigma} \text{를 } x \text{에 대해 풀어준다} \right)$$

백분위수	z	$x = \mu + z\sigma$	x(가장 가까운 정수)
95분위수(상위 5%)	1.645	$x = 75 + (1.645)(7)$	86.52 또는 87 (반올림 됨)
90분위수(상위 10%)	1.282	$x = 75 + (1.282)(7)$	83.97 또는 84 (반올림 됨)
75분위수(상위 25%)	0.675	$x = 75 + (0.675)(7)$	79.73 또는 80 (반올림 됨)
25분위수(하위 25%)	-0.675	$x = 75 - (0.675)(7)$	70.28 또는 70 (반올림 됨)
10분위수(하위 10%)	-1.282	$x = 75 - (1.282)(7)$	66.03 또는 66 (반올림 됨)
5분위수(하위 5%)	-1.645	$x = 75 - (1.645)(7)$	63.49 또는 63 (반올림 됨)

면적이 주어졌을 때 z값 찾기



함수문법:

`=NORM.DIST(x,μ,σ, cumulative)`

`=NORM.S.DIST(z,1)`

예시:

`=NORM.DIST(80,75,7,1) = 0.762475`

`=NORM.S.DIST(1.96,1) = 0.975002`

무엇을 계산:

주어진 μ 와 σ 하에서 x 의 왼쪽 면적
 $\mu = 75$ 그리고 $\sigma = 7$ 에서 시험응시자의
 점수가 80점 이하일 확률은 76.25%이다.

표준정규분포에서 z 의 왼쪽 면적
 $z = 1.96$ 의 왼쪽 면적은 97.50%이다.

예제: 시험 점수

- John이 수강하는 경제학 과목 교수는 10백분위 이하의 학생들은 재시험을 치러야 한다고 하였다. 시험점수는 정규분포이고 $\mu = 75$ 이고 $\sigma = 7$ 이다. 재시험을 치르는 학생의 시험점수는 몇 점 이하인가?
- $P(X < x) = 0.10$ 을 만족하는 x 값 찾기(10백분위에 해당하는 값은 $z = -1.28$)

* 구하는 방법

$P(Z < -1.28) = 0.10$ 을 만족하는 $z = -1.28$ 을 엑셀 또는 부록 C를 이용해서 찾을 때 $z = (x - \mu) / \sigma$ 공식을 이용하여 $-1.28 = (x - 75) / 7$ 을 계산하기 위해서 주어진 정보 대입 $x = 75 - (1.28)(7) = 66.04$ 에 대해 위 식을 풀어(또는 반올림해서 66점) x 를 구함
→ 경제학 시험에서 66점 이하를 받는 학생들은 재시험을 치러야 한다.

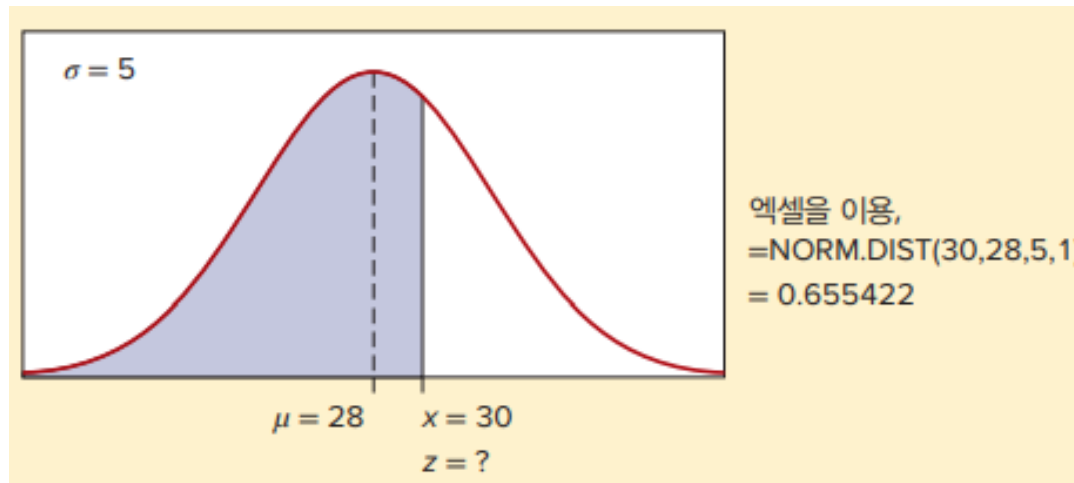
예제: 엔진오일 교환 시간

- 엔진오일 교환과정을 연구한 후 정비공장 매니저는 서비스 시간 X 가 정규분포로서 평균이 $\mu = 28$ 분이고 표준편차가 $\sigma = 5$ 분이라는 것을 발견하였다. 즉, $X \sim N(28, 5)$ 이다. 이러한 정보는 정규분포 확률을 구하는 데 사용될 수 있다.
- 정규분포에서 확률을 구하는 문제에 답하기 위해서는 다음 몇 가지 단계를 따르면 유용하다
 - ① 당신이 알고 있는 정보를 가지고 그림을 그리고 그림에 표시를 하라.
 - ② 답에 해당하는 면적을 표시하라.
 - ③ 확률변수를 표준화 하라.
 - ④ 분포표나 엑셀을 이용하여 면적을 계산하라.

예제: 엔진오일 교환 시간

질문 #1: 30분 이내로 작업이 끝나는 자동차의 비율은?

- 1단계와 2단계: 그림을 그리고 원하는 면적(30이하)을 표시하라.

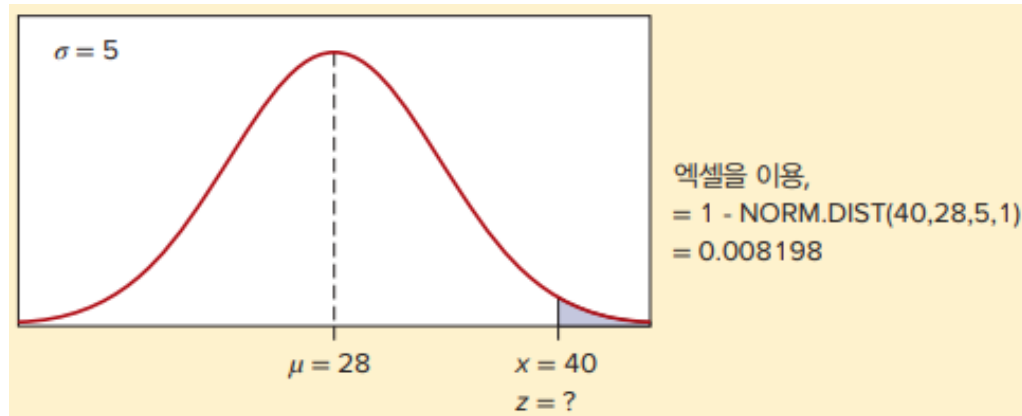


- 3단계: $z = \frac{30 - 28}{5} = 0.40$
- 4단계: 부록 C-2나 엑셀을 이용하여 $P(Z < 0.40) = 0.6554$ 임을 찾는다

예제: 엔진오일 교환 시간

질문 #2: 임의로 선택된 자동차가 40분 이상 걸릴 확률은?

- 1단계와 2단계: 그림을 그리고 원하는 면적(x 가 40이상)을 표시하라.



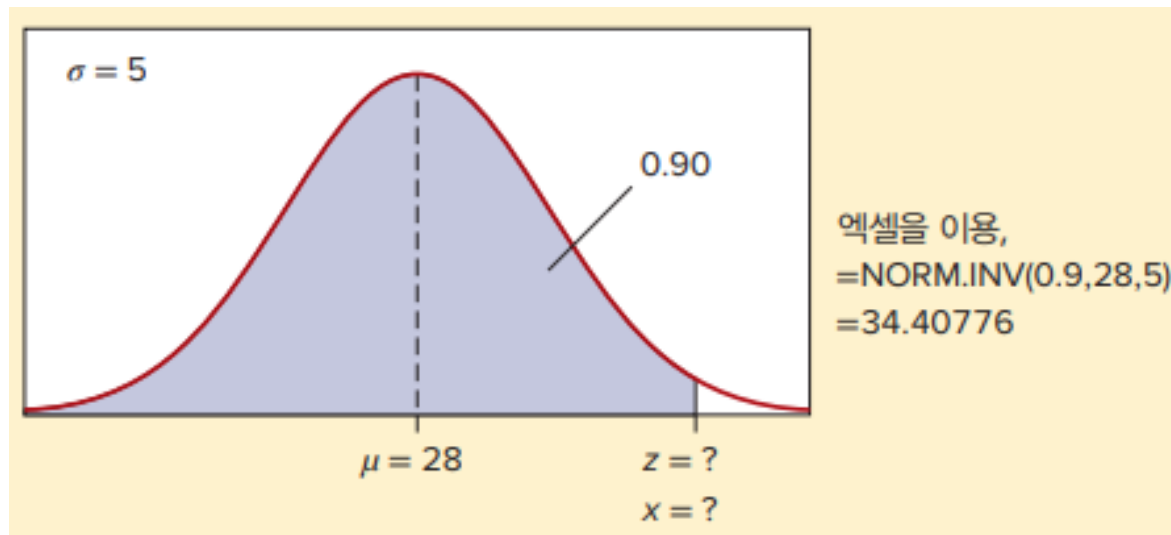
- 3단계: $z = \frac{40 - 28}{5} = 2.4$
- 4단계: 부록 C-2 또는 엑셀을 이용하여 아래 확률을 구한다.

$$P(Z > 2.4) = 1 - P(Z \leq 2.4) = 1 - .9918 = .0082.$$

예제: 엔진오일 교환 시간

질문 #3: 90백분위수에 해당하는 교환시간은 얼마인가?

- 1단계와 2단계: 그림을 그리고 원하는 면적을 표시하라.



이 질문의 경우 3단계와 4단계의 순서를 바꾸어 풀 수 있다.

예제: 엔진오일 교환 시간

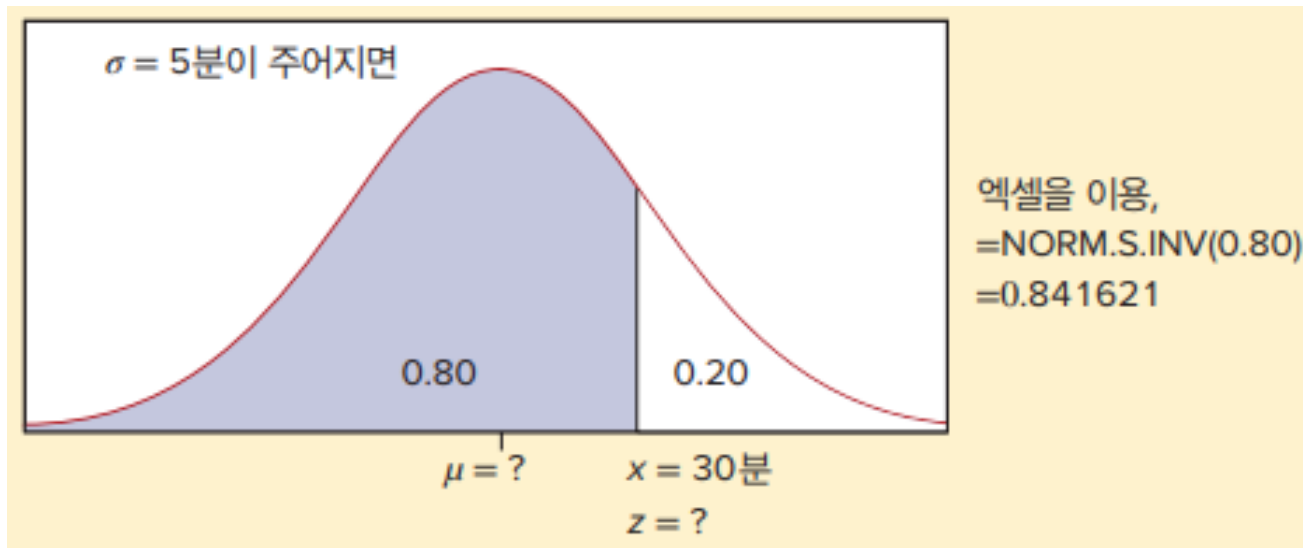
- 3단계: 표나 엑셀을 이용하여 $z = 1.28$ 을 찾는다
- 4단계: $1.28 = \frac{x - 28}{5}$, $x = 28 + 5(1.28) = 34.4$ 분
→ 90%의 자동차는 34.4분 이내에 교환을 마칠 것이다.

예제: 엔진오일 교환 시간

질문 #4: 매니저는 전체 차량의 80%가 30분 이내 마치기를 원한다.

이 목표를 달성하기 위해서 평균 서비스 시간은 얼마여야 하는가?

- 1단계와 2단계: 그림을 그리고 원하는 면적을 표시하라.



예제: 엔진오일 교환 시간

- 3단계: 표 또는 엑셀을 이용하여 상위 20%(왼쪽 꼬리면적이 80%)에 해당하는 $z = 0.84$ 를 찾는다.

- 4단계: $z = \frac{x - \mu}{\sigma}$ 공식에서 $0.84 = \frac{30 - \mu}{5}$ 를 만들고 이 식을 풀어

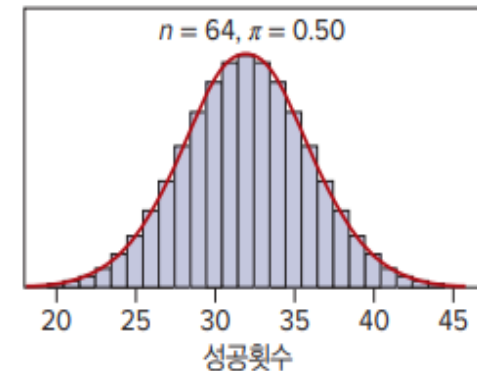
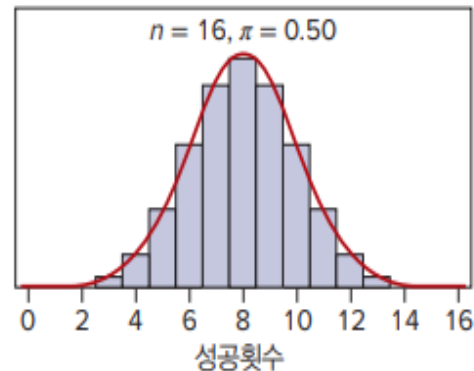
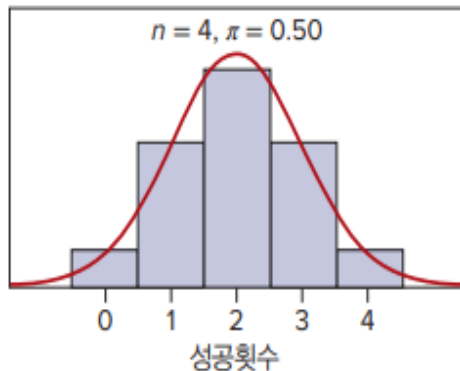
$$\mu = 30 - 0.84(5) = 25.8 \text{ 를 구한다.}$$

→ 평균 서비스 시간이 25.8분이 되면 전체 차량의 80%가 30분 이내에 오일교환 서비스를 마칠 수 있다.

이항분포의 정규분포 근사

- 이항분포의 경우 n 이 크면 계산과정이 매우 복잡하므로 정규분포 근사법(approximation)을 대신 이용
이러한 근사법의 원리는 n 이 커질수록 이항분포의 막대그래프가 점점 매끈해지고, 연속적인 정규분포 모양을 닮아가기 때문임
- 아래 그림은 n 번(4번, 16번, 64번) 동전을 던졌을 때 앞면의 횟수를 x 로 정의할 때의 확률분포
- 표본의 크기가 커짐에 따라 아래 그림에 나와 있는 막대그래프처럼 매끈하고 종 모양의 분포가 되는 것을 시각적으로 알 수 있음

n 이 커짐에 따라 이항분포는 정규분포에 근사



이항분포의 정규분포 근사

- 경험법칙에 따르면 $n\pi \geq 10$ 이고, $n(1 - \pi) \geq 10$ 일 때, 이항분포를 정규분포로 근사해도 무리가 없음
- 정규분포의 평균 μ 와 표준편차 σ 는 이항분포의 평균과 표준편차와 같음

$$\mu = n\pi$$

$$\sigma = \sqrt{n\pi(1 - \pi)}$$

예제: 동전 던지기

- 공정한 동전을 32번 던졌을 때 앞면이 17번을 초과하여 나올 확률은?

- 이항분포의 경우

$$P(X \geq 18) = P(18) + P(19) + \dots + P(32),$$

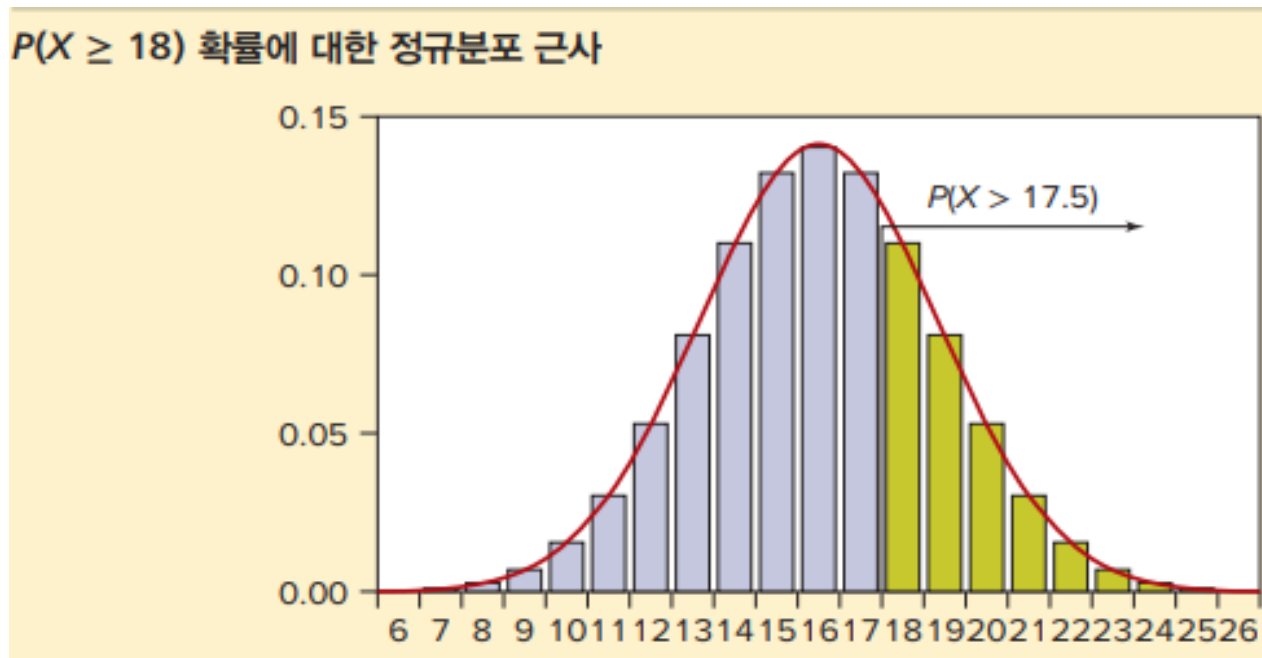
이고, 표를 이용하여 계산해도 계산이 지루해짐

- 이 때, 정규분포 근사법을 이용하면?

$n = 32$ 와 $\pi = 0.50$ 은 $n\pi \geq 10$ 이고 $n(1 - \pi) \geq 10$ 조건을 만족시킨다

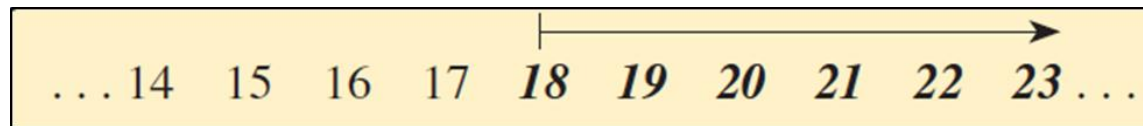
예제: 동전 던지기

- 그러나 이산형 변수가 연속형 변수로 간주되기 위해서는 개별 값에 주의하여야 한다.
'17번 초과' 사건의 그림 7.21에 보듯이 이산형 눈금의 17과 18의 중간보다 큰 값에 해당된다.



예제: 동전 던지기

- 데이터의 전체 분포를 그릴 필요는 없지만, ‘17번 초과’라는 사건을 시각적으로 보여주기 위해 그림을 그려볼 수도 있다.



- 위와 같이 그림을 그려보면, 올바른 절단점을 찾을 수 있다. ‘17 초과’ 사건에 대한 적절한 절단점은 17과 18의 중간 정도로 보이며, 정규근사법을 이용한다면 $P(X > 17.5)$ 의 확률을 구하면 된다. X 에 추가적으로 더해진 0.5를 연속성 수정(continuity correction)이라고 부른다. 정규분포의 모수들은 다음과 같다.

$$\mu = n\pi = (32)(0.5) = 16$$

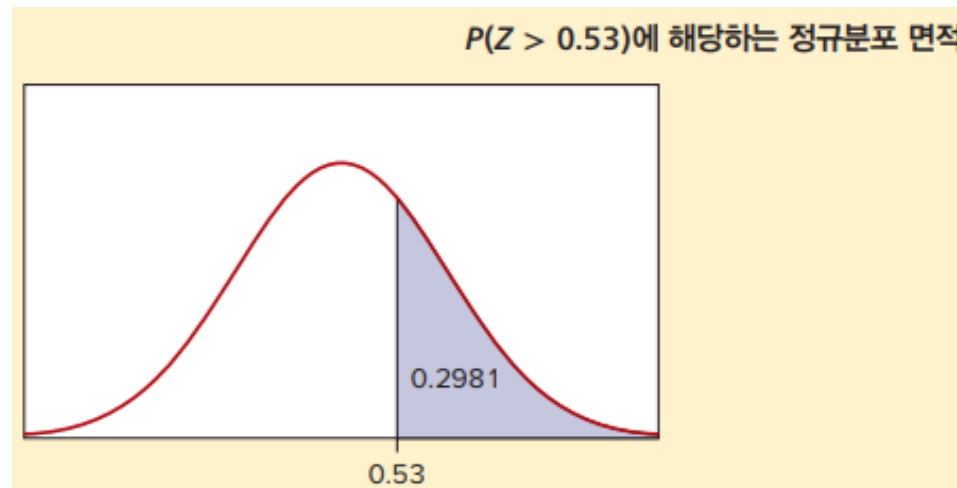
$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(32)(0.5)(1-0.5)} = 2.82843$$

예제: 동전 던지기

- 연속성 수정된 X값을 가지고 표준화를 다음과 같이 실시한다.

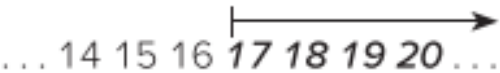
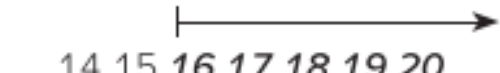
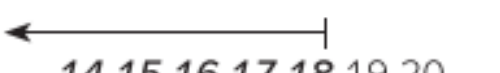
$$z = \frac{x - \mu}{\sigma} = \frac{17.5 - 16}{2.82843} = .53$$

- 부록 C-1표에서 우리는 $P(Z > 0.53) = 0.5000 - P(0 < Z < 0.53) = 0.5000 - 0.2019 = 0.2981$ 을 계산할 수 있다. 또 다른 방법으로 부록 C-2표를 이용하면 $P(Z > 0.53)$ 은 정규분포의 대칭성을 이용하여 $P(Z < -0.53) = 0.2981$ 이다.



이항분포의 정규분포 근사

- 연속성 수정을 이해하기 위해 아래 표에서 제공하는 사건을 고려하자.
- 이산형 모형을 연속형 모형으로 바꾸는 데 있어서 적절한 절단점을 찾는 방법을 간략한 그림으로 제공하였다.

사건	적절한 값	정규분포 절단점
적어도 17		$x = 16.5$ 를 사용
15 초과		$x = 15.5$ 를 사용
19 미만		$x = 18.5$ 를 사용

포아송분포의 정규분포 근사

- 포아송분포에 대한 정규분포 근사는 λ 값이 클 때 적절한 방법(예를 들어, 부록 B 표에서 λ 값을 찾지 못한다면 정규분포 근사 활용 가능)
- 포아송분포를 정규분포로 근사하여 사용하기 위해서는 정규분포의 모수인 μ 와 σ 를 포아송분포의 평균과 표준편차와 같다고 설정해야 함

$$\mu = \lambda$$

$$\sigma = \sqrt{\lambda}$$

예제: 전기요금 문의

* 수요일 오전 10시부터 12시까지 전기요금 고지서에 대한 문의는 평균적으로 1시간에 42건이다. 50건을 초과하여 문의를 받을 확률은?

- 평균 $\lambda = 42$ 는 부록 B를 사용하기에는 너무 큰 값이다.
- 정규분포 근사를 이용하면 간단하게 계산이 가능하다.

$$\mu = \lambda = 42$$

$$\sigma = \sqrt{\lambda} = \sqrt{42} = 6.48074$$

→ $X \geq 51$ 에 대한 연속성 수정값은 $X = 50.5$ (50과 51의 중간값)이다.



예제: 전기요금 문의

- ‘50 초과’에 대한 표준화된 Z값을 아래와 같이 계산하면 $P(X > 50.5) = P(Z > 1.31)$

$$z = \frac{x - \mu}{\sigma} = \frac{50.5 - 42}{6.48074} \cong 1.31$$

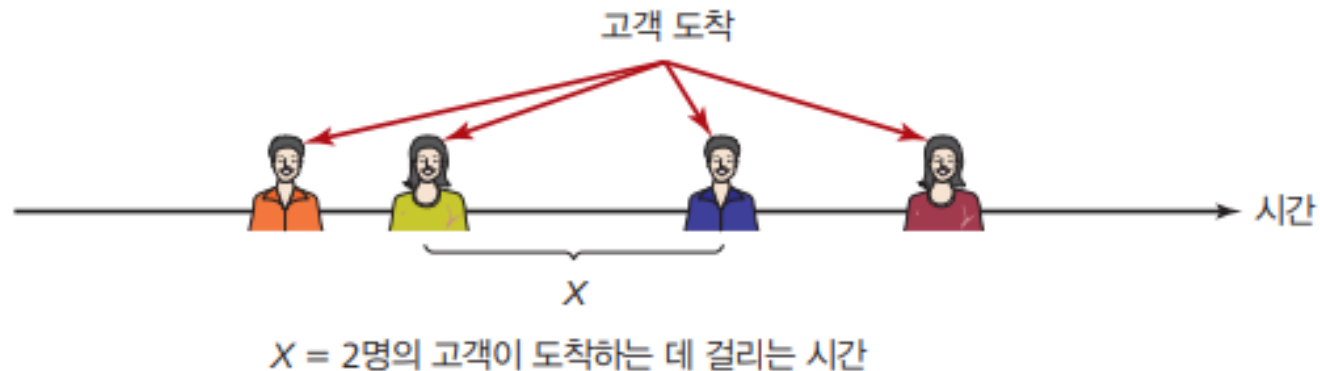
- 부록 C-2표를 이용하면 정규분포는 대칭이기 때문에 우리는 $P(Z < -1.31) = 0.0951$ 이고 이확률은 $P(Z > 1.31)$ 와 같다. 실제 포아송 확률을 엑셀 함수 =POISSON.DIST(50, 42, 1)로 계산하고 1에서 그 값을 빼주면

$$P(X \geq 51) = 1 - P(X \leq 50) = 1 - .9025 = .0975$$

이 경우 정규분포 근사확률(0.0951)은 실제 포아송 결과(0.0975)와 매우 유사하다.
이 예를 통해서 우리는 근사법에 대한 어느 정도 확신을 가질 수 있다(엑셀을 이용할 수 있다면 실제로 근사법을 사용해야 할 이유는 없다)

지수분포(exponential distribution)

- 단위시간당 사건발생횟수는 포아송분포(Poisson distribution)를 따르는 반면, 사건이 일어난 후 다른 사건이 발생하기까지 걸리는 시간은 지수분포(exponential distribution)를 따름
- 다른 사건이 발생하기까지 걸리는 시간은 연속형 변수



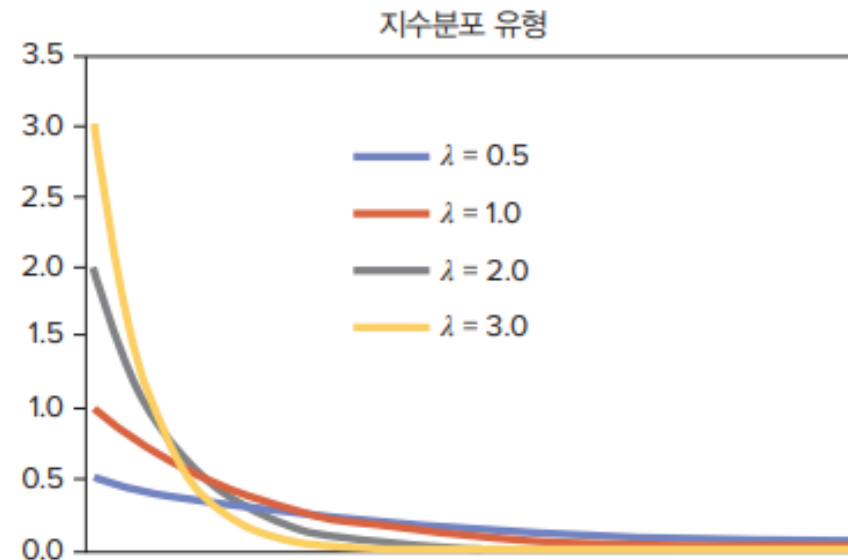
지수분포(exponential distribution)의 특징

모수	λ = 단위시간 또는 공간에서 평균 도착비율(포아송에서 평균과 같음)
PDF	$f(x) = \lambda e^{-\lambda x}$
CDF	$P(X \leq x) = 1 - e^{-\lambda x}$
정의역	$x \geq 0$
평균	$1/\lambda$
표준편차	$1/\lambda$
분포모양	항상 오른쪽 꼬리분포
엑셀에서 PDF*	=EXPON.DIST(x, λ , 0)
엑셀에서 CDF*	=EXPON.DIST(x, λ , 1)
엑셀에서 난수생성	=-LN(RAND())/ λ (1개의 값 생성)
참조	λ (평균 도착비율) 대신 $1/\lambda$ (평균 대기시간)을 알고 있는 경우도 있다.

지수분포(exponential distribution)

- 지수분포는 단 하나의 모수에 의해 결정. 즉 평균 도착비율인 λ 이며, λ 값에 따라 서로 다른 PDF를 생성함
- 그러나 지수분포 PDF는 서로 같은 모양으로 나타남

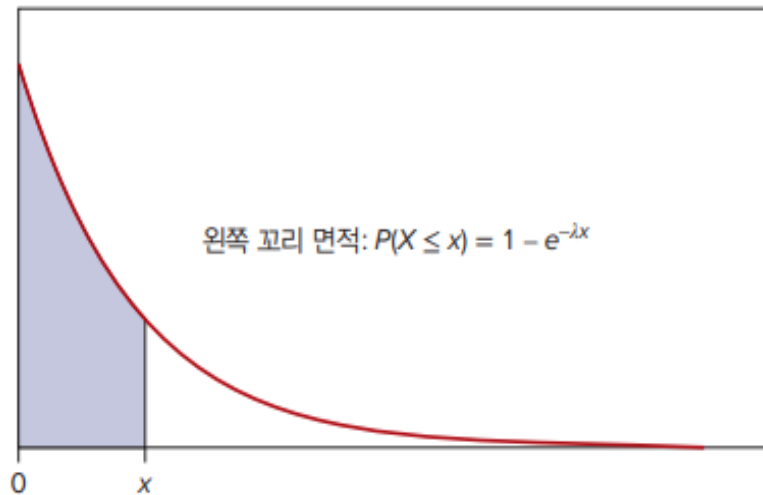
- 오른쪽 PDF에서 알 수 있듯이 $f(0) = \lambda$ 이기 때문에 수직축과 만나는 점은 항상 λ 임



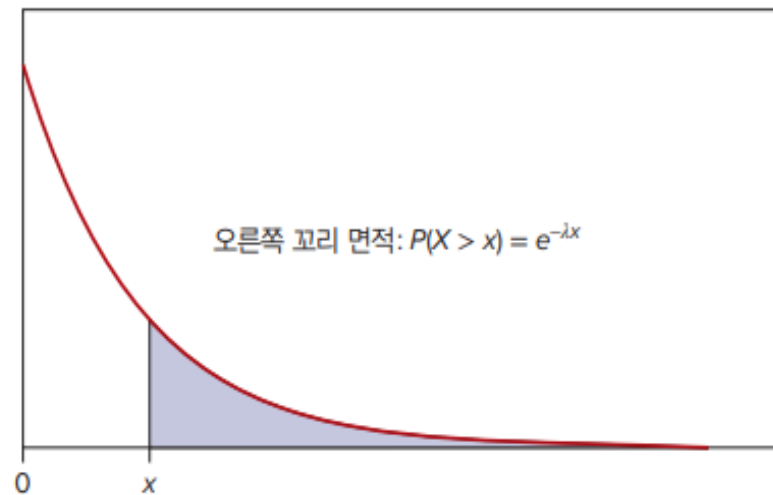
지수분포(exponential distribution)

- 지수분포의 확률함수는 x 가 증가함에 따라 0에 가까워지며 오른쪽꼬리분포임
- PDF $f(x)$ 의 높이보다 그 곡선 아래의 면적이 포인트
- 다음 사건 발생까지의 대기시간이 x 보다 더 클 확률은 $e^{-\lambda x}$ 이고 대기시간이 x 보다 작을 확률은 $1 - e^{-\lambda x}$ 임

왼쪽 꼬리 지수분포 면적



오른쪽 꼬리 지수분포 면적



예제: 고객 대기시간

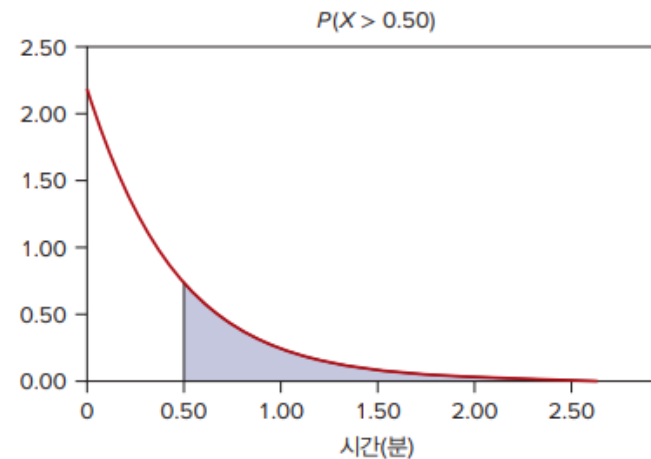
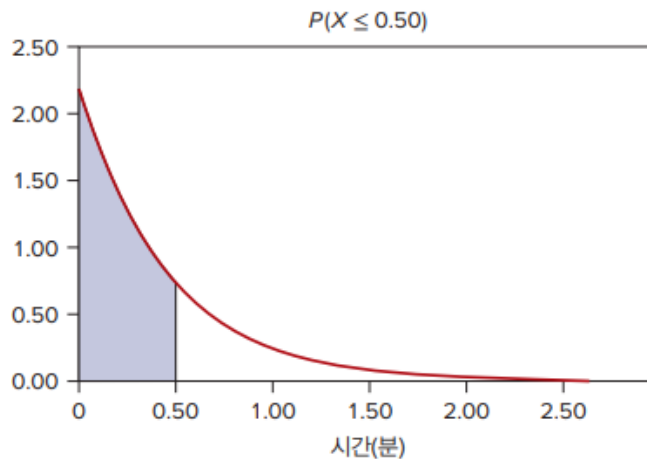
- * 블루초이스 보험회사의 경우 수요일 오후 2시부터 4시까지 고객들의 전화문의는 1분당 2.2건이다. 다음 전화문의까지 대기시간이 30초 이상일 확률은?
- 우리는 1분당 $\lambda = 2.2$ 로 놓고 $x = 0.5$ 분으로 설정한다.
- λ 가 분 단위로 되어 있으므로 30초를 0.5분으로 변환하였음에 주목하라. 단위는 서로 같게 설정하여야 한다.
- 다음과 같은 확률을 얻을 수 있다.

$$P(X > 0.50) = e^{-\lambda x} = e^{-(2.2)(0.50)} = .3329, \text{ or } 33.29\%$$

예제: 고객 대기시간

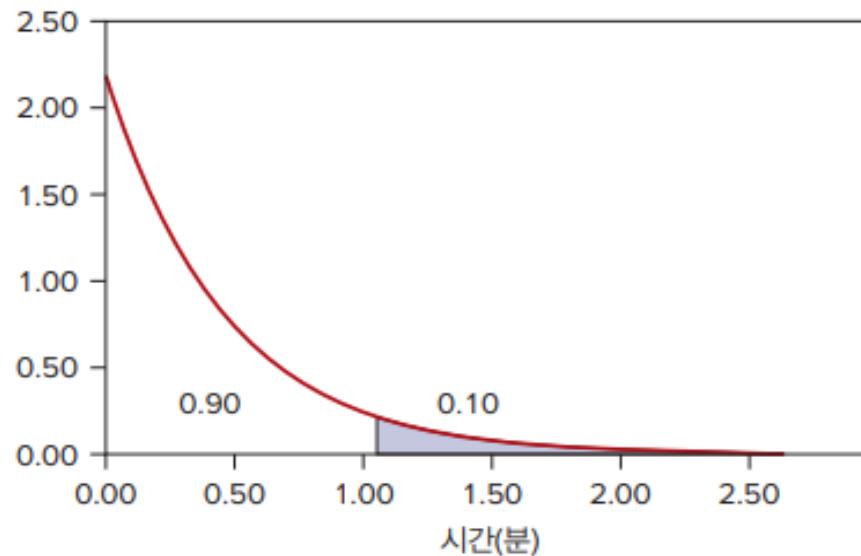
- 다음 전화문의가 올 때까지 30초 이상 기다릴 확률은 대략 33%이다. $x = 0.50$ 은 한 점이고 확률이 존재하지 않기 때문에 $P(X \geq 0.50)$ 와 $P(X > 0.50)$ 는 같은 사건을 언급하는 것이다.
- 이항분포의 경우에는 각 점에서 확률이 존재하므로 서로 다른 확률이었다. 대기시간이 30초 미만일 확률은 다음과 같다.

$$P(X \leq 0.50) = 1 - e^{-(2.2)(0.50)} = 1 - .3329 = .6671$$



예제: 고객 대기시간

- 우리는 지수분포의 면적을 계산하는 공식을 역으로 이용할 수 있다.
- 평균도착 비율이 1분에 2.2건일 때 대기시간이 90백분위인(상위 10%인) 값을 알고자 한다.
- 우리는 상위 10%를 정의하는 x값을 찾고자 한다.



예제: 고객 대기시간

- $P(X \leq x) = 0.90$ 은 $P(X > x) = 0.10$ 이기 때문에, 우리는 오른쪽 꼬리의 확률을 0.10으로 설정하였다. 양변에 로그를 취하고 x 에 대해서 풀면

$$P(X \leq x) = 1 - e^{-\lambda x} = .90$$

$$e^{-\lambda x} = .10$$

$$-\lambda x = \ln(.10)$$

$$-(2.2)x = -2.302585$$

$$x = 2.302585/2.2$$

$$x = 1.0466 \text{ 분}$$

→ 전화 대기시간 90백분위수는 1.0466분(또는 62.8초)이 된다.

사건들 간의 평균 대기 시간

- 지수분포를 따르는 대기시간은 사건들 간의 평균대기시간(mean time between events: MTBE)으로 표현되기도 하며, 이 경우 단위시간 동안의 포아송 발생횟수인 λ 대신 $1/\lambda$ 가 주어짐
- $MTBE = 1/\lambda =$ 사건들 간의 평균 대기시간(사건당 평균시간)
- $1/MTBE = \lambda =$ 단위시간당 평균사건 횟수(단위시간당 사건)
- * 예를 들어, 응급실 환자 도착의 평균시간 간격이 20분이라면 1분당 $\lambda = 1/20 = 0.05$ 건 발생(또는 1시간에 $\lambda = 3.0$ 건 발생)
- $e^{-\lambda x}$ 를 계산할 때 x 와 λ 가 같은 단위로 표현되어야 한다는 것에 주의한다면 시간 또는 분 단위 중 어느 것을 사용해도 됨
- 예를 들어, $P(X > 12\text{분}) = e^{-(0.05)(12)} = e^{-0.60}$ 과 $P(X > 0.2\text{시간}) = e^{-(3)(0.20)} = e^{-0.60}$ 은 서로 같음

예제: LCD 모니터

- 비행기 조종실에서 사용되는 NexGenCo의 컬러 LCD 모니터는 평균무고장시간(MTBE)이 22,500시간이다. 다음 10,000비행시간 동안 오작동이 발생할 확률은? 여기에서 시간당 오작동 횟수가 $\lambda = 1/22,500$ 이므로 우리는 다음과 같이 확률을 계산한다.

$$P(X < 10,000) = 1 - e^{-\lambda x} = 1 - e^{-(1/22,500)(10,000)} = 1 - e^{-0.4444} = 1 - .6412 = .3588$$

- 다음 10,000 비행시간 내에 오작동을 일으킬 확률은 35.88%이다. 이는 오작동이 포아송분포를 따른다고 가정할 때의 결과이다.