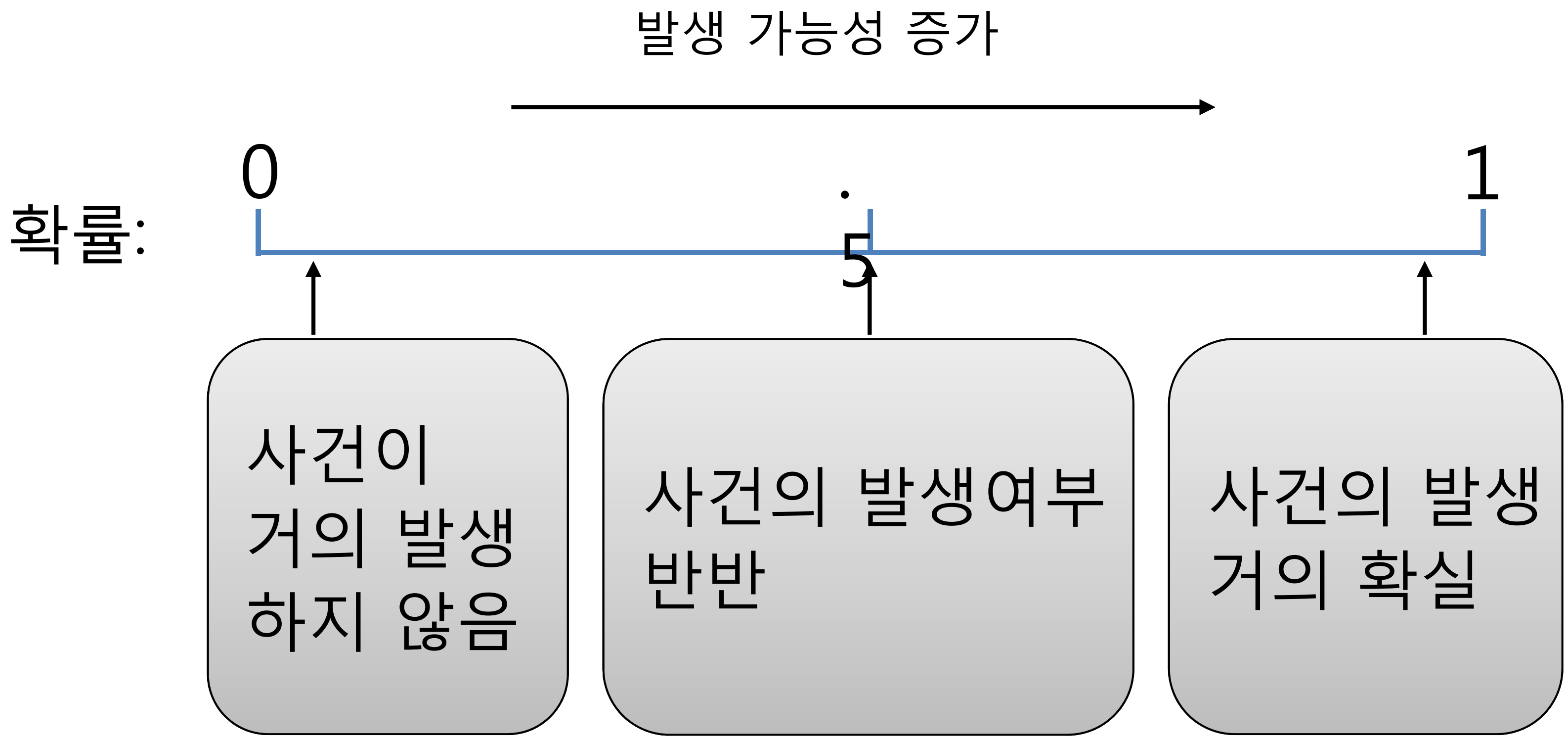
확률분포

사건의 발생 가능성에 대한 척도로서의 확률



실험, 표본 공간, 사건, 확률

- 실험: 정의된 결과들을 산출하는 과정
- 표본공간: 발생 가능한 모든 실험 결과(표본점)들의 집합
- 사건: 표본점들의 집합
- 각 표본점들의 확률을 더해 사건의 확률 계산

표본공간, 사건, 확률

- 주사위를 한 개 던졌을 때 짝수가 나올 확률?

- 동전을 두 개 던졌을 때 두 개 모두 앞면일 확률?

- 한 패의 카드에서 두 카드를 골라 모양만을 기록할 때 (Heart, Spade) 일 확률?

확률의 부여(assigning probabilities)

고전적 방법

결과의 발생확률이 같다는 가정하에 확률을 부여

상대도수 방법

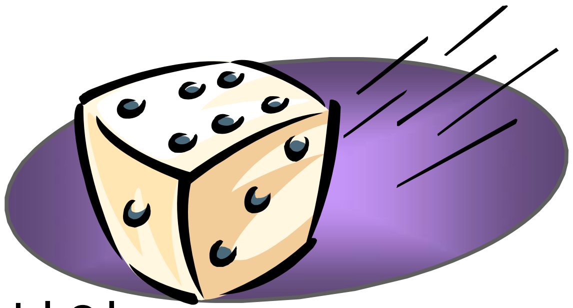
실험이나 역사적 자료에 기초하여 확률을 부여

주관적 방법

주관적 판단에 기초하여 확률을 부여

고전적(classic) 방법

만약 n 개의 가능한 실험결과가 있을 때, 각 결과에 1/n의 확률을 부여한다.



실험: 주사위 던지기

표본 공간: $S = \{1, 2, 3, 4, 5, 6\}$

확률: 각 표본점은 1/6의 발생가능성이 있다.

상대도수(relative frequency) 방법

- 예 : Lucas Tool Rental(공구 대여)
  Lucas Tool Rental 은 하루에 대여되는 자동차 광택기의 대수에 대하여 확률을 부여하려고 한다. 지난 40일간 일 대여 대수의 기록이 아래와 같다.

대여 대수	대여일수
0	4
1	6
2	18
3	10
4	2

상대도수 방법

각 부여된 확률은 총도수(총 일수)로 해당 도수 (해당 일수)를 나눈 값이다.

대여 대수	일수	확률
0	4	.10
1	6	.15
2	18	.45
3	10	.25
4	2	.05
	40	1.00

4/40

주관적(subjective) 방법

경제적 상황이나 기업환경이 빠르게 변할 때 확률 부여를 역사적 자료에만 의존하는 것은 적절하지 못하다.

경험이나 직관과 같은 다른 이용 가능한 자료도 사용할 수 있다. 그러나 확률값은 궁극적으로 실험의 결과가 일어날 것이라는 믿음의 정도를 표시해야 한다.

가장 좋은 확률 추정값은 고전적 방법이나 상대도수 방법에 의한 추정값에 주관적 방법의 추정값을 결합하여 구해질 수 있다.

주관적 방법

한 분석가가 주관적 방법을 적용하여 아래와 같이 확률을 부여하였다.

실험결과	순 손익	확률
(10, 8)	$18,000 Gain	.20
(10, -2)	$8,000 Gain	.08
(5, 8)	$13,000 Gain	.16
(5, -2)	$3,000 Gain	.26
(0, 8)	$8,000 Gain	.10
(0, -2)	$2,000 Loss	.12
(-20, 8)	$12,000 Loss	.02
(-20, -2)	$22,000 Loss	.06

확률변수 (Random Variables)

- 하나의 실험에서 나타나는 결과를 수치로 나타낸 것
- 확률변수는 각 표본점들과 실수를 연관짓는 함수

Examples: 확률변수 X (or Y, or...)

1. W = 주사위를 두 번 던졌을 때 두 눈의 합
2. Q = 100 명의 사람 중 파란 눈인 사람의 명수
3. Y = 무작위로 선정된 사람의 몸무게

이산 vs. 연속 확률변수

- 이산확률변수
  유한한 수의 값을 갖거나 무한수열의 값을 갖는 확률변수

- 연속확률변수
  일정한 구간 또는 구간들의 집합에서 어떠한 수치적 값을 갖는 확률변수

이항분포(binomial distribution)

이산확률분포의 특수한 경우

- 이항실험의 네 가지 속성

1. 실험은 *n*개의 연속된 동일한 시행으로 구성된다.
(n이 사전에 정해진다)

2. 각 시행에서 두 개의 결과가 가능하다. 결과 중 하나는 성공(success), 다른 하나는 실패(failure)라고 부른다.

3. 성공 확률을 *p*로 표시하는데 이 확률은 시행에 따라 변하지 않는다

시행마다
동일한 가정

4. 각 시행은 독립적이다.

이항 분포

이항확률 함수

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

여기서 :

f(x) = n회 시행 중 x회 성공 확률

n = 시행 횟수

p = 하나의 시행에서 성공 확률

```
dbinom(x, size, prob, log = FALSE)
rbinom(n, size, prob)
```

Example : 동전던지기

- 동전을 2번 던질 때 확률변수 X를 앞면의 개수로 정의하자.
- Q: 이항확률분포의 4가지 속성을 만족하는가?

Event:	TT	TH	HT	HH
X:				
Prob:				

Sample Space

Prob. Dist. of x

x			
f(x)			

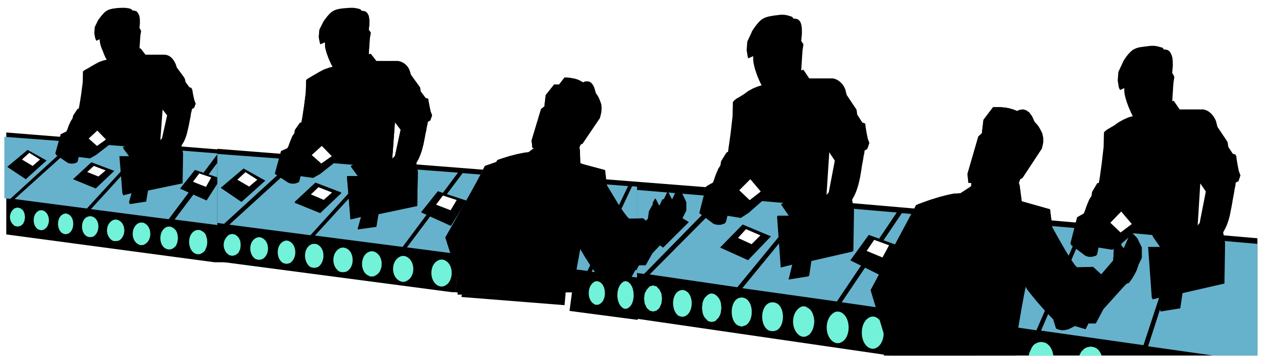
```
> dbinom(0:2,2,0.5)
[1] 0.25 0.50 0.25
```

이항 분포

- 예: Evans Electronics

  Evans회사는 근로자들에 대한 낮은 근속율을 걱정하고 있다. 최근, 경영진은 시간제 근로자 중 연간 10%가 이동한 것으로 알고 있다.

  그래서, 경영진은 무작위로 선정한 시간제 근로자에 대하여, 그 근로자가 내년에 회사에서 일하지 않을 확률이 0.1 이라고 추정하였다.



이항 분포

- 이항확률 함수 이용

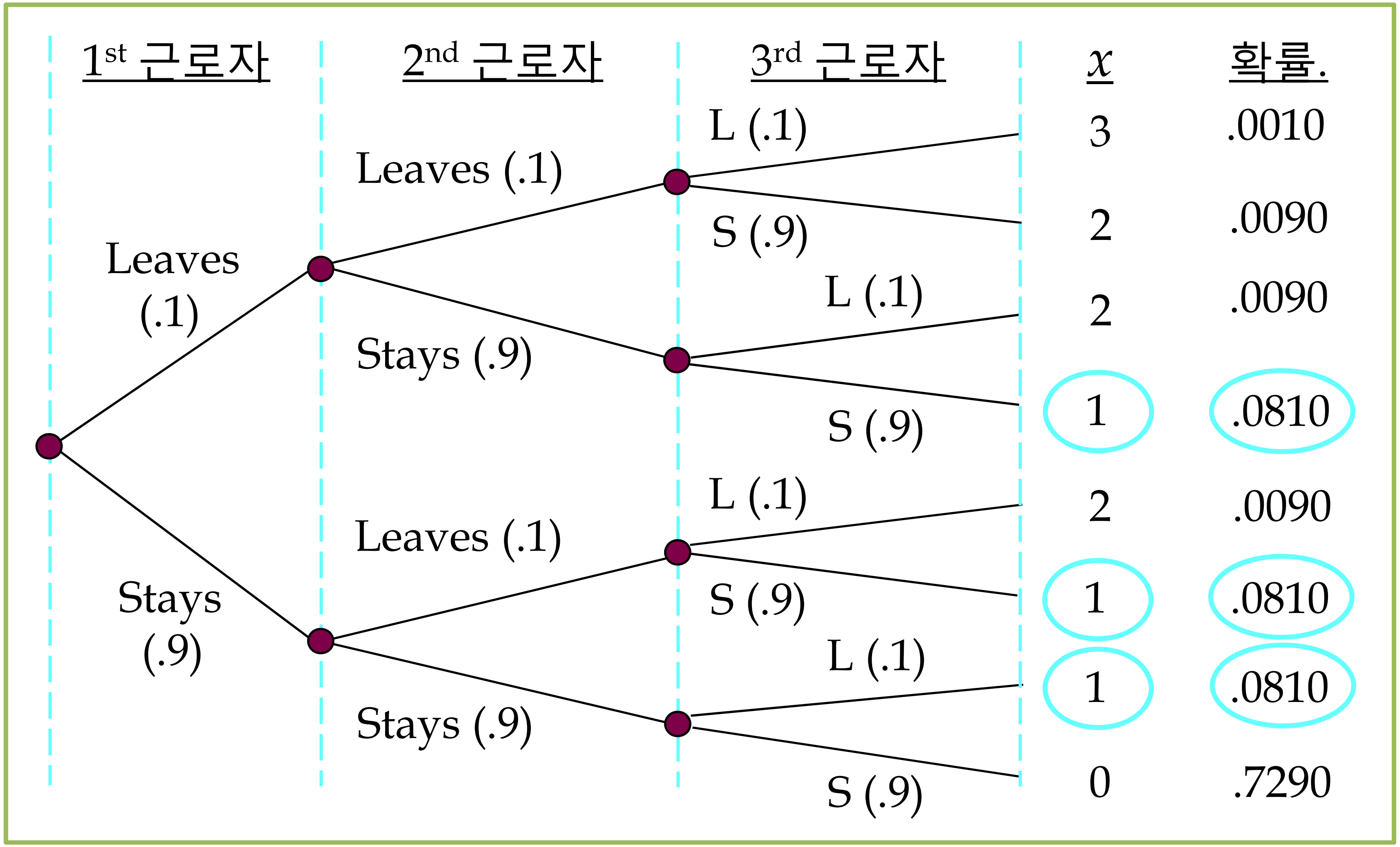
시간제 근로자 3명을 무작위로 뽑을 경우, 그 중에 한 명이 금년에 회사를 떠날 확률이 얼마인가?

Let: $p = .10,\ n = 3,\ x = 1$

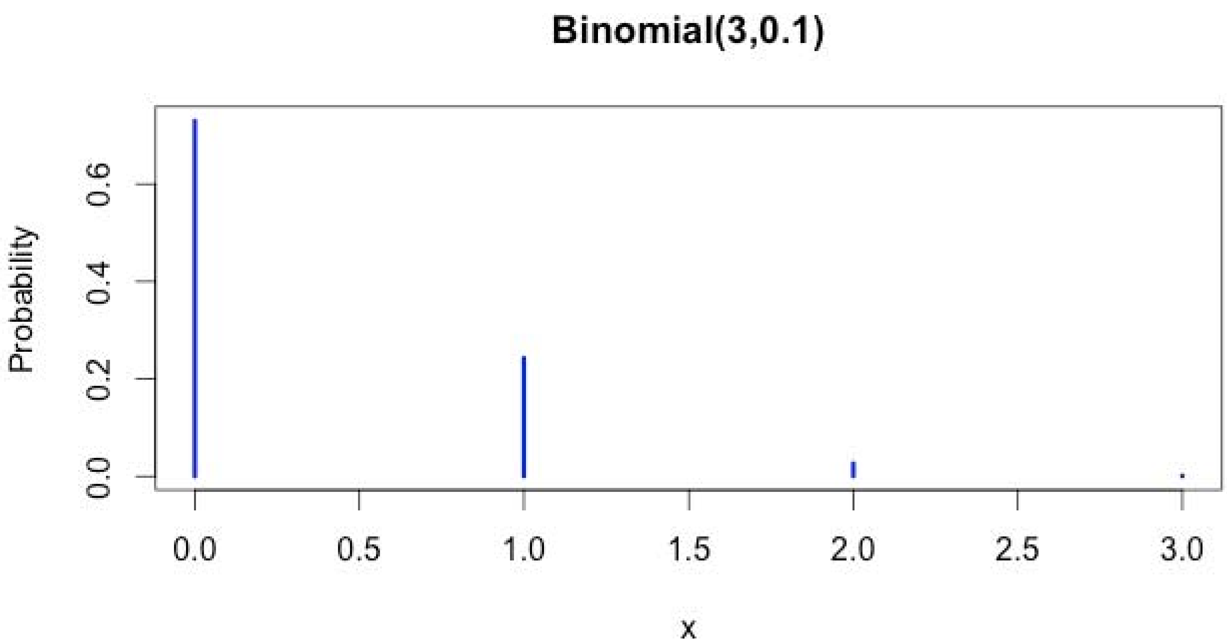
$$f(x) = \frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x}$$

이항 분포

- 계통도(tree diagram)



이항 분포




```
> x=0:3
> plot(x,dbinom(x,3,0.1),type="h", xlab="x",ylab="Probability",main="Binomial(3,0.1)",col=4,lwd=3)
```

이항 분포

기대값 $$E(x) = \mu = np$$

분산 $$\mathrm{Var}(x) = \sigma^2 = np(1-p)$$

표준편차 $$\sigma = \sqrt{np(1-p)}$$

```
> 3*0.1
[1] 0.3
> 3*0.1*0.9
[1] 0.27
> sqrt(3*0.1*0.9)
[1] 0.5196152
```

```
> x=rbinom(100000,3,0.1)
> mean(x)
[1] 0.30338
> var(x)
[1] 0.2711233
> sd(x)
[1] 0.520695
```

연속확률분포

연속확률변수는 연속된 어떤 구간이나 구간들의 집합에 있는 값을 취할 수 있다.

예:

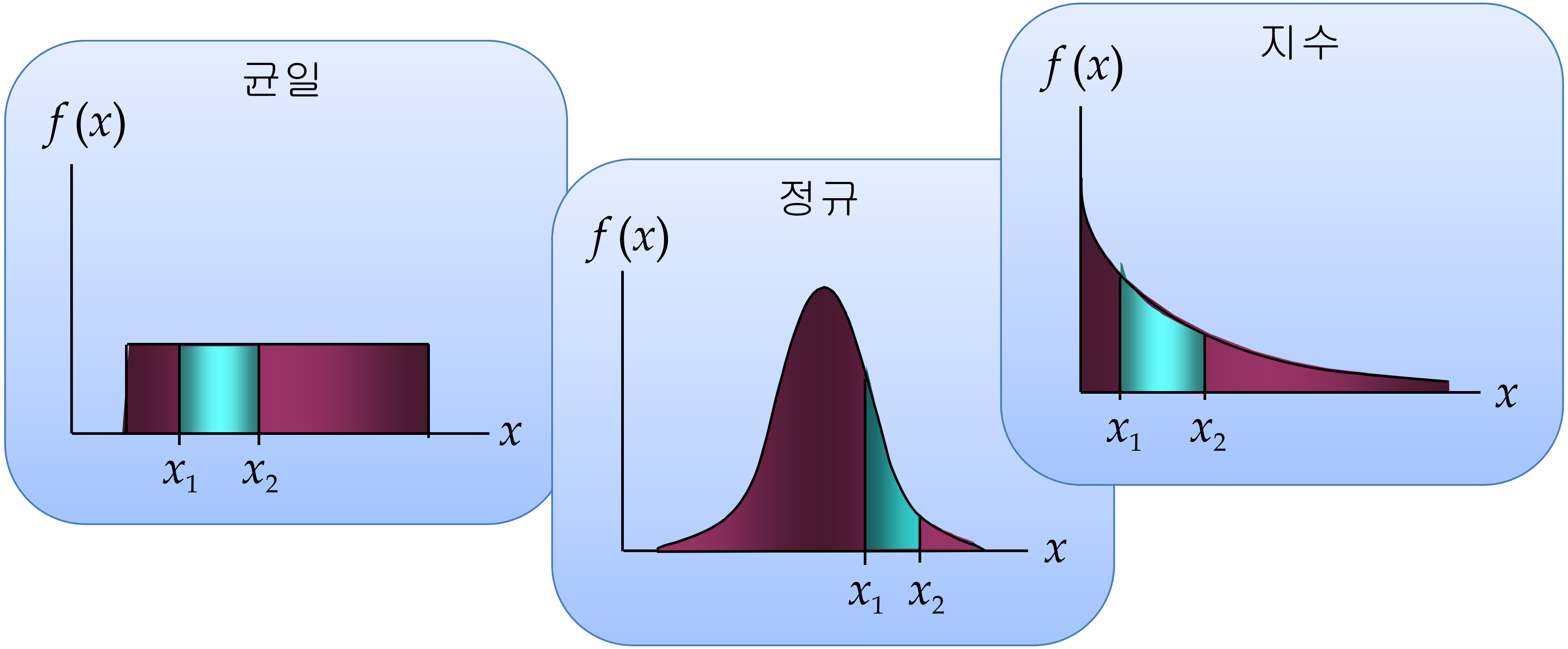
1. 버스가 도착할 때 까지 기다리는 시간
2. 무작위로 뽑은 사람의 몸무게
3. 무작위로 뽑은 사람의 콜레스테롤 수치

특정한 값을 가진 확률변수의 확률값은 나타낼 수 없다.

대신에, 확률변수가 주어진 **구간** 내에 있을 확률을 계산한다.

연속확률분포

주어진 x_1 에서 x_2까지 구간에서 값을 가질 수 있는 확률변수의 확률은 확률밀도함수의 x_1 에서 x_2까지 구간에서 그래프 아래 부분의 면적이 된다.



정규확률분포(normal probability distribution)

- 정규확률분포는 연속확률분포를 기술하는 가장 중요한 분포이다.
- 통계적 추론에 폭넓게 사용된다.

정규확률분포

- 아주 다양한 분야에 이용되어 왔다.

사람들의 신장(키)

과학적 측정

정규확률분포

- 아주 다양한 분야에 이용되어 왔다.

시험점수

강수량

정규확률분포

- 정규확률밀도함수(normal probability density function) ; $N(\mu, \sigma^2)$

$$\boldsymbol{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}}$$

여기서:

μ = 평균

σ = 표준편차

π = 3.14159

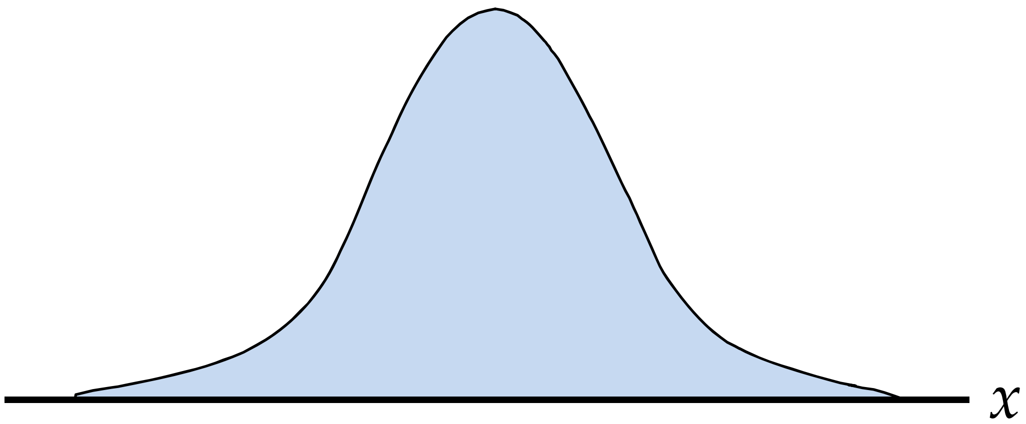
e = 2.71828

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

정규확률밀도함수

- 특성

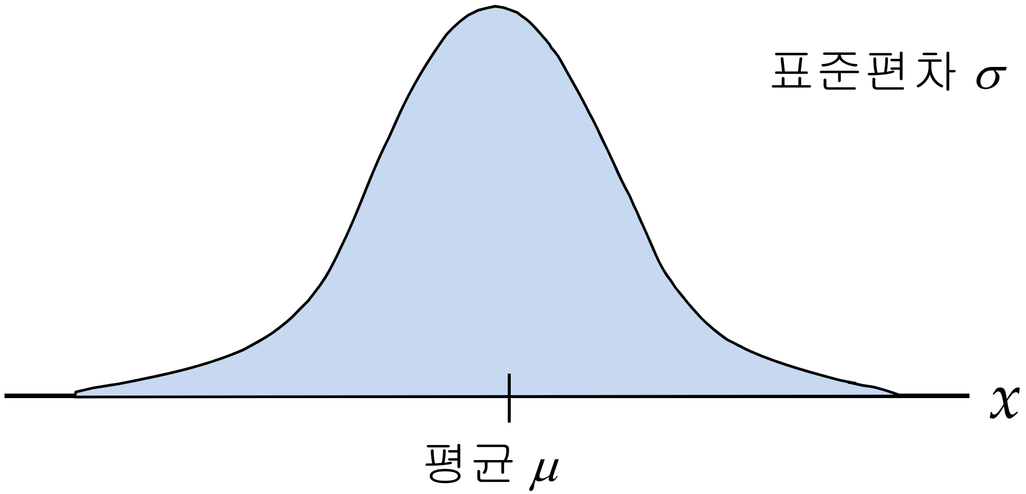
대칭 분포 (종형 분포)



정규확률분포

- 특성

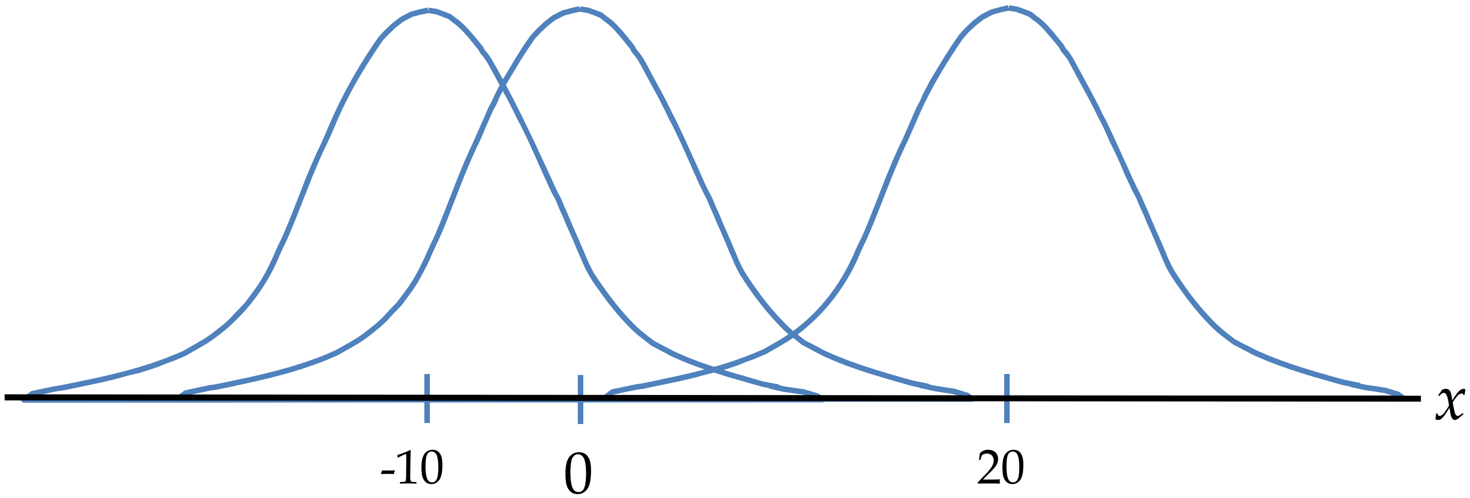
모든 정규확률분포는 평균 μ 와 분산 σ 에 의해 정의될 수 있다.



정규확률분포

- 특성

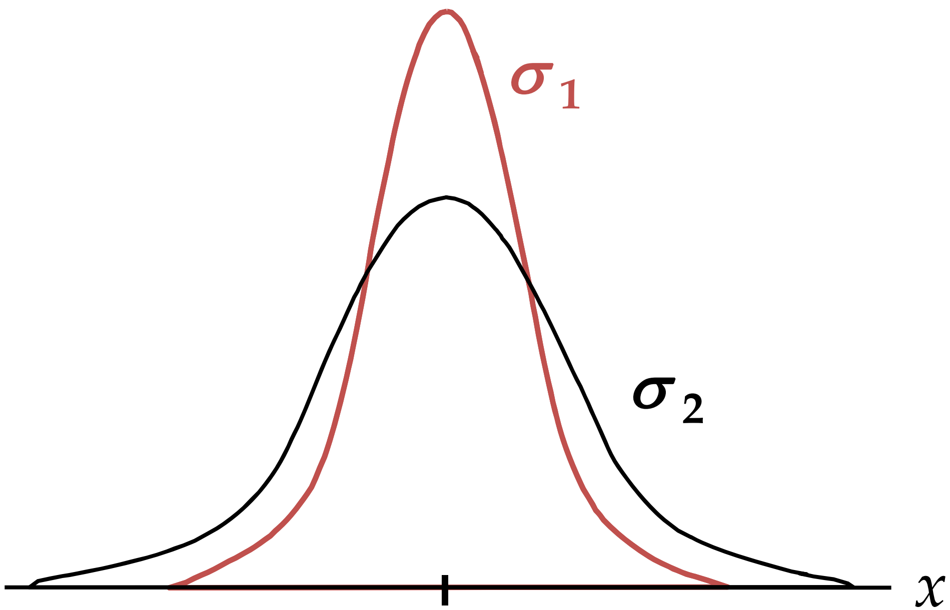
평균은 어떤 수치(음수, 0 , 양수)라도 될 수 있다.



정규확률분포

- 특성

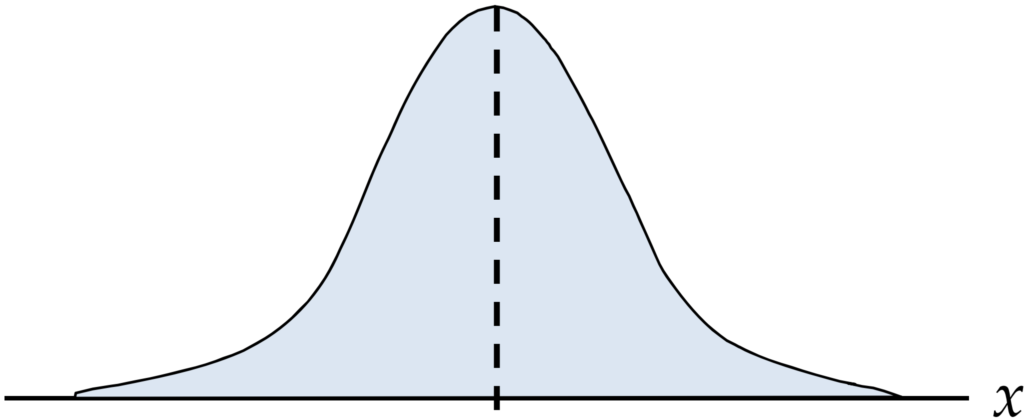
  표준 편차는 곡선의 넓이를 결정한다: 표준편차가 클수록 곡선은 넓어지고 평평해진다.



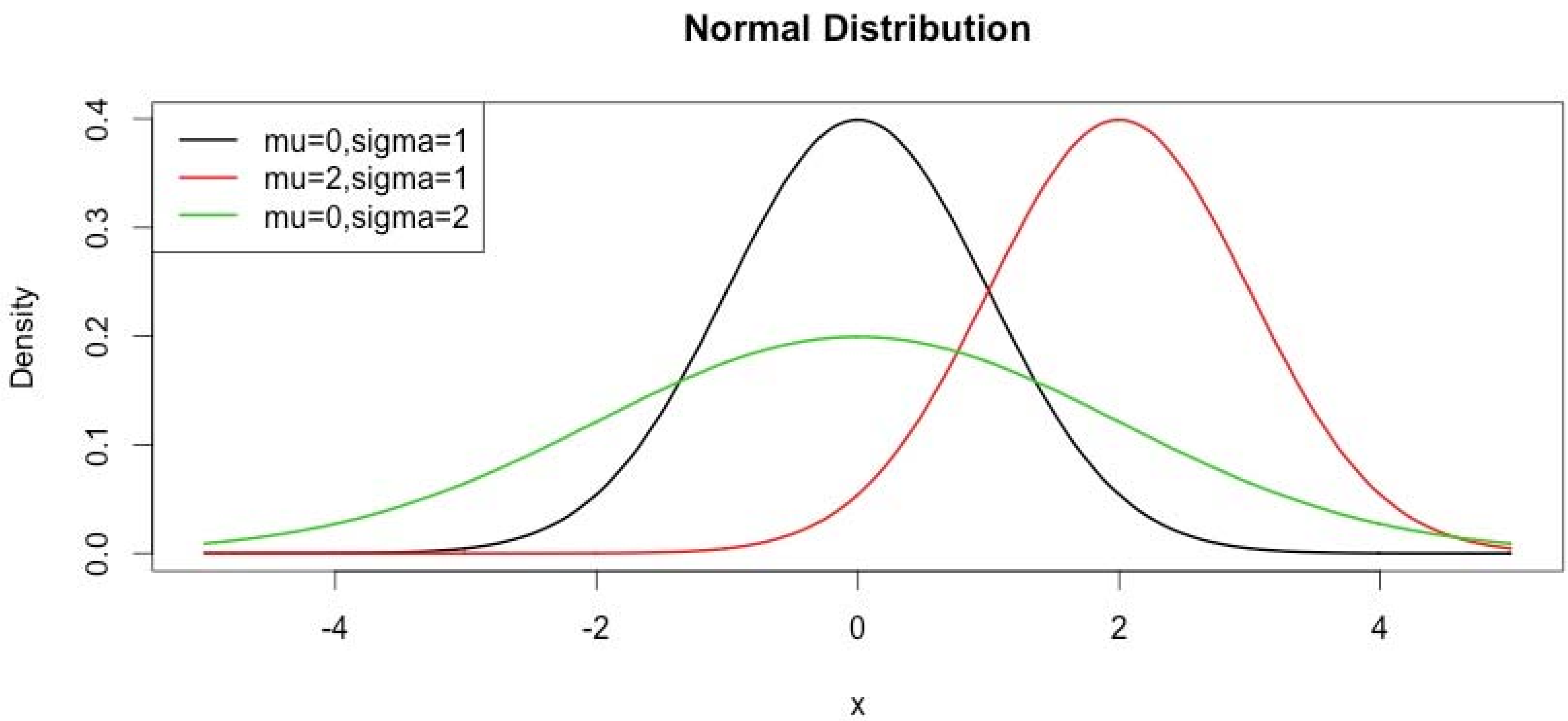
정규확률분포

■ 특성

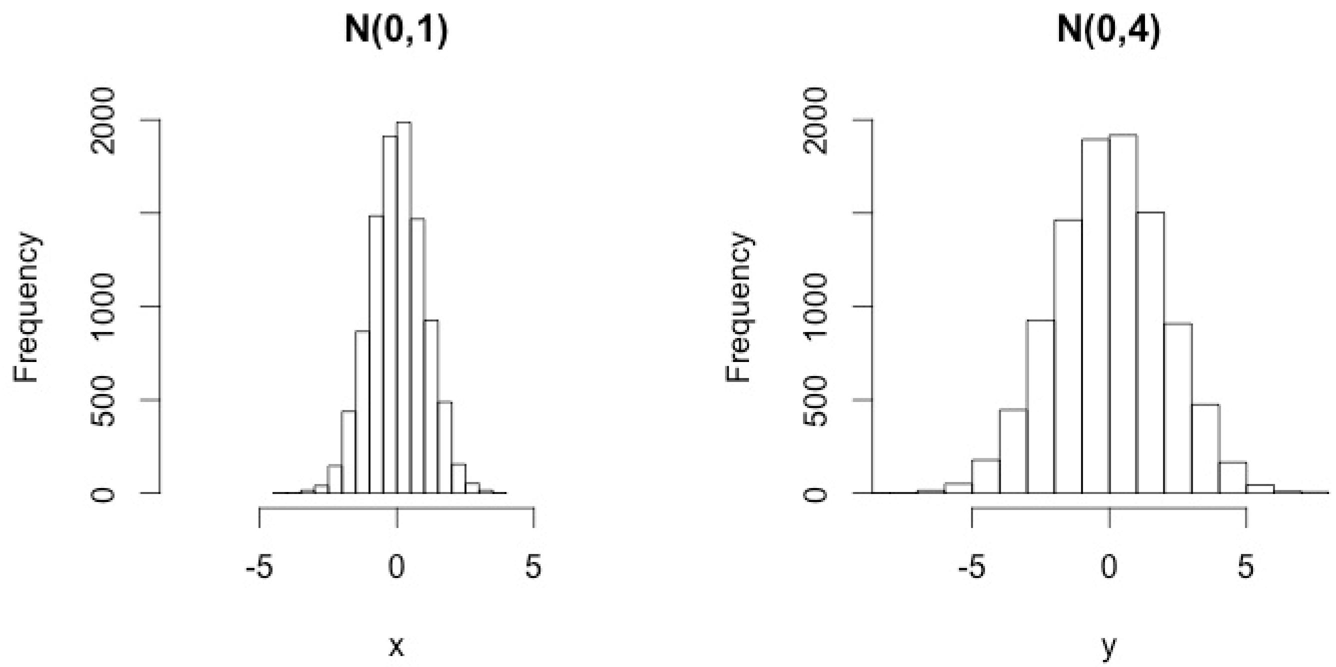
정규확률변수의 확률은 곡선 아래 부분의 면적이다.
곡선아래의 총 면적은 ______이다. (평균의 왼쪽은 ________ 그리고 평균의 오른쪽은 ______ 이다).



정규확률분포




```
plot(x,dnorm(x,0,1),'l',ylab="Density",main="Normal Distribution",lwd=2)
lines(x,dnorm(x,2,1),col=2,lwd=2)
lines(x,dnorm(x,0,2),col=3,lwd=2)
legend("topleft",c("mu=0,sigma=1", "mu=2,sigma=1","mu=0,sigma=2"),lty=1,col=1:3,lwd=2)
```




```
x=rnorm(10000,0,1)
y=rnorm(10000,0,2)
par(mfcol=c(1,2))
hist(x,xlim=c(-8,8),ylim=c(0,2000),main="N(0,1)")
hist(y,xlim=c(-8,8),ylim=c(0,2000),main="N(0,4)")
```

예: Pep Zone

Pep Zone은 여러 등급의 유명한 자동차 오일을 포함하여 자동차 부품이나 용품을 판매한다. 이 오일의 재고가 20 gallons 이하로 떨어질 경우,

보충 주문을 한다.

예: Pep Zone

관리자는 주문을 기다리는 동안 재고가 떨어져 판매를 못할 경우를 걱정하고 있다.

오일 주문에서 보충될 때 까지

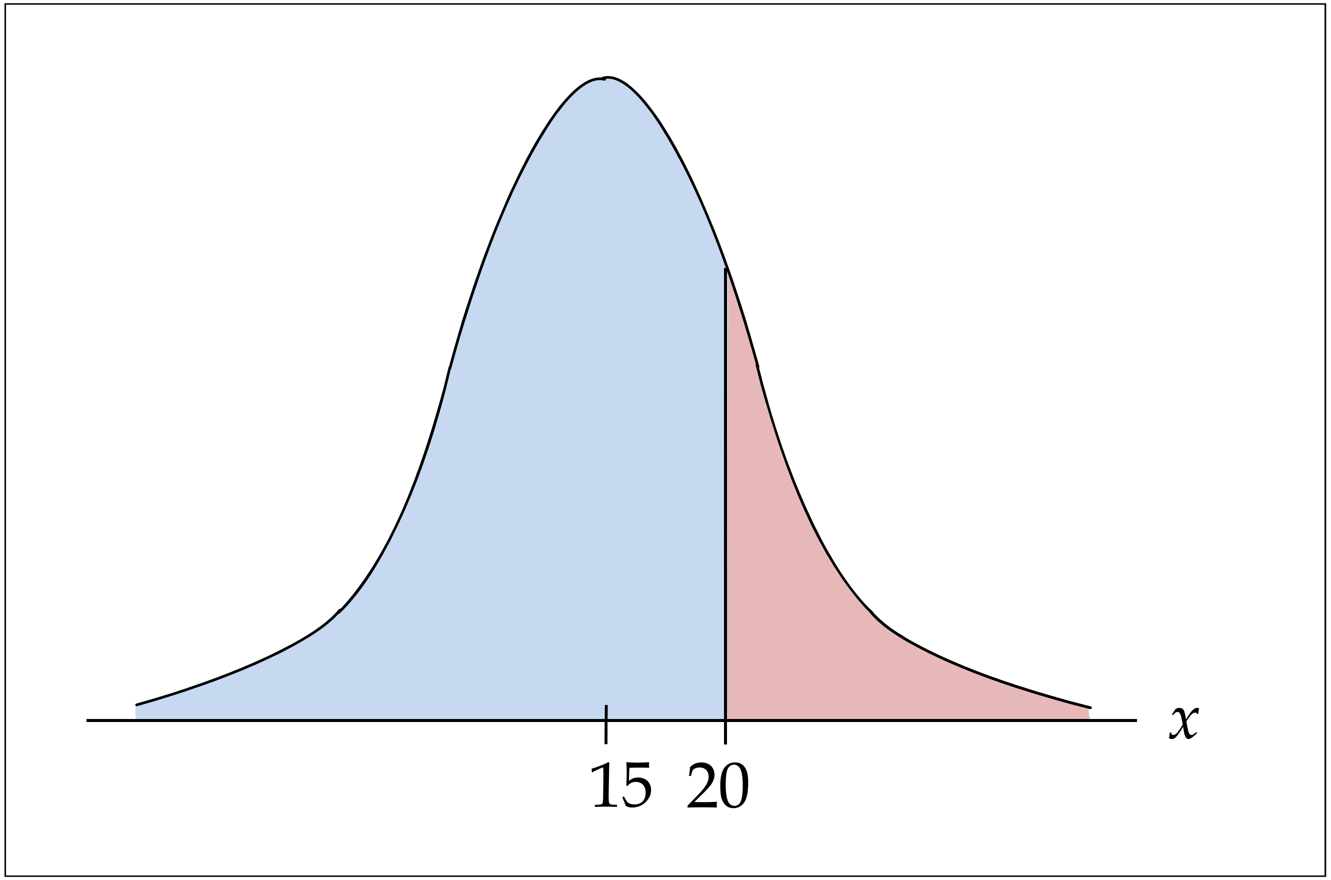
판매에 필요한 양의 분포는

평균 15 gallons 과 표준편차 6 gallons

이다.

관리자는 재고가 바닥날 확률, 즉 $P(x > 20)$에 대하여 알고 싶어 한다.

Graphical Representation of P(x>20)




```
> pnorm(20,15,6)
[1] 0.7976716
> pnorm(20,15,6,lower.tail=FALSE)
[1] 0.2023284
```

예: Pep Zone - Finding Percentiles

- Reorder Point(주문시 재고량)

만약 Pep Zone의 관리자가 재고가 바닥날 확률을 .05정도로 하고 싶어 하면, 주문 시점의 재고량 (Reorder Point)을 얼마로 하면 되겠는가?

❑ 지금까지와 반대의 과정: finding percentiles

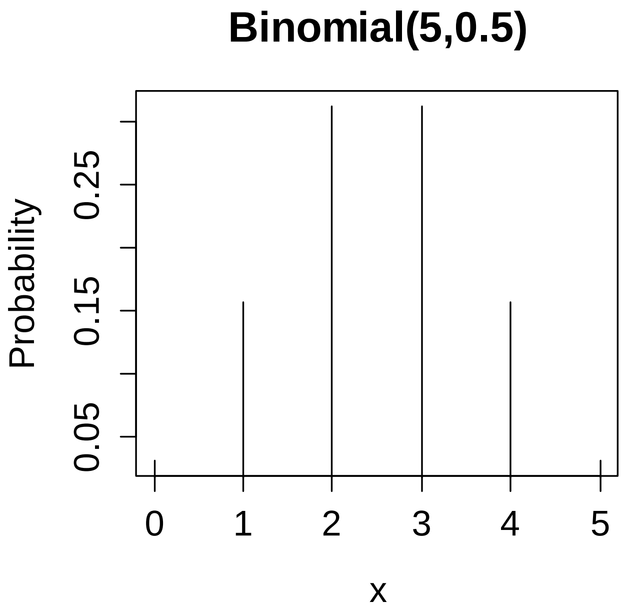
확률 α가 주어진 상황에서 **P(X ≤ x) = α를 만족하는 x값을 구하고 싶다.**

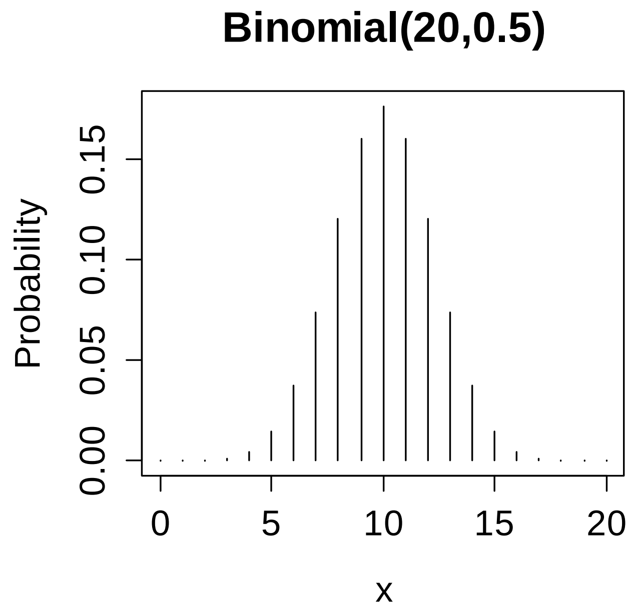
```
> qnorm(0.95,15,6)
[1] 24.86912
> qnorm(0.05,15,6,lower.tail=FALSE)
[1] 24.86912
```

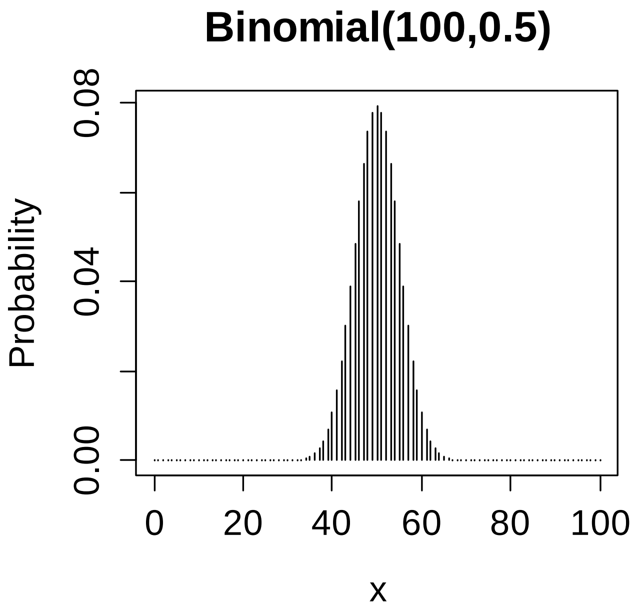
이항확률의 정규근사 (normal approximation of binomial probabilities)

시행횟수 n 이 커지면, 이항확률함수를 손이나 계산기로 계산하는 것이 어렵게 된다.

정규분포는 $n > 20$, $np > 5$, 그리고 $n(1 - p) > 5$ 일 경우의 이항확률에 대한 쉬운 근사치를 제공한다.


Binomial(5,0.5)
Probability
0.05
0.15
0.25
0
1
2
3
4
5
x


Binomial(20,0.5)
Probability
0.00
0.05
0.10
0.15
0
5
10
15
20
x


Binomial(100,0.5)
Probability
0.00
0.04
0.08
0
20
40
60
80
100
x

이항확률의 정규근사

설정

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

0.5 (연속성수정계수, continuity correction factor)를 더하고 빼라. 이는 이산분포를 근사하기 위해 연속분포가 사용되기 때문이다.

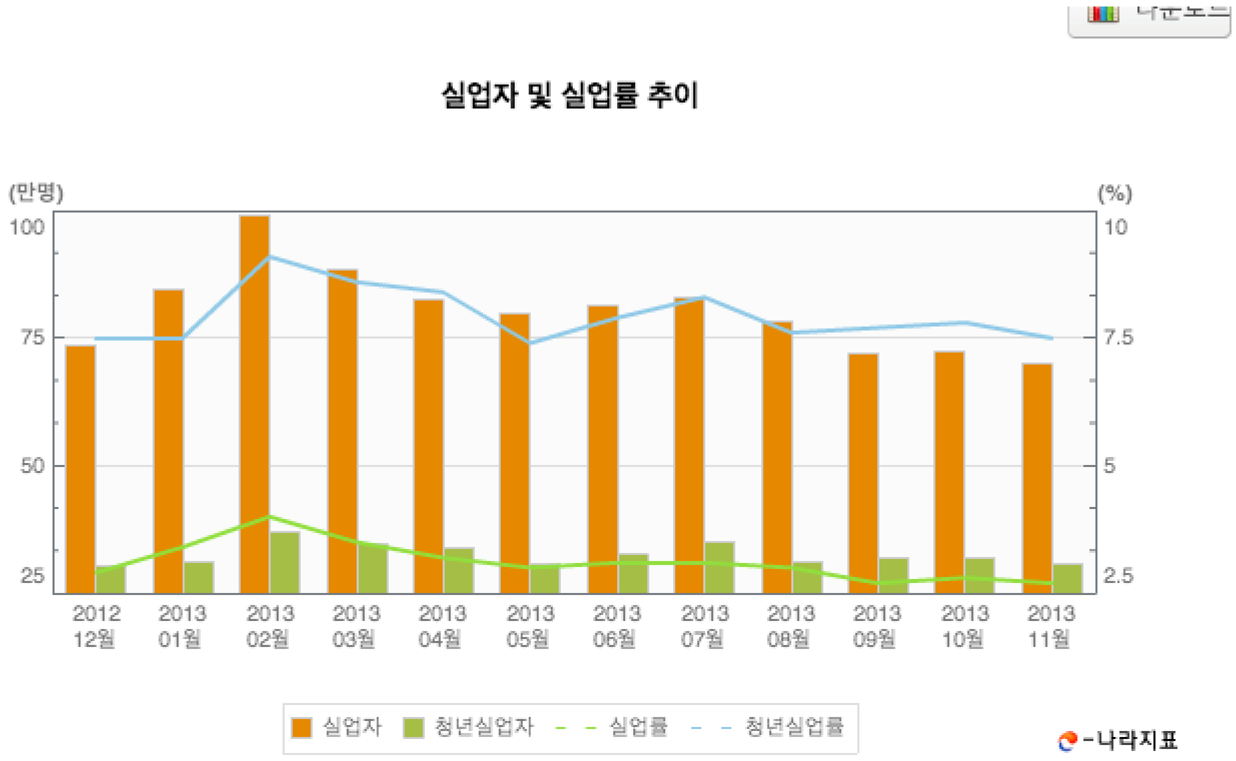
예를 들어, 이산이항분포에서 $P(x = 10)$는 연속정규분포에서 $P(9.5 \leq x \leq 10.5)$ 로 근사된다.

실업률

[단위 : 만명, %]

	2008	2009	2010	2011	2012	2013 06월	2013 07월	2013 08월	2013 09월	2013 10월	2013 11월
취업자 증감	14.5	-7.2	32.3	41.5	43.7	36.0	36.7	43.2	46.3	47.6	58.8
- 농림어업	-3.7	-3.8	-8.2	-2.5	-1.4	-1.9	0.3	2.8	2.6	1.5	-3.1
- 제조업	-5.2	-12.6	19.1	6.3	1.4	9.6	5.3	0.5	2.1	3.1	3.5
- 건설업	-3.7	-9.1	3.3	-0.2	2.2	0.3	1.1	-0.7	0.7	-1.4	-1.8
- 서비스업	26.0	17.9	20.0	38.6	41.6	26.2	28.2	38.6	39.7	43.2	59.8
실업자	76.9	88.9	92.0	85.5	82.0	81.3	82.8	78.3	72.0	72.4	70.0
실업률(%)	3.2	3.6	3.7	3.4	3.2	3.1	3.1	3.0	2.7	2.8	2.7
청년실업자	31.5	34.7	34.0	32.0	31.3	32.7	35.2	31.3	32.0	31.8	30.9
청년실업률(%)	7.2	8.1	8.0	7.6	7.5	7.9	8.3	7.6	7.7	7.8	7.5

실업률


실업자 및 실업률 추이
(만명)
(%)
100
75
50
25
10
7.5
5
2.5
2012 12월
2013 01월
2013 02월
2013 03월
2013 04월
2013 05월
2013 06월
2013 07월
2013 08월
2013 09월
2013 10월
2013 11월
실업자
청년실업자
실업률
청년실업률
-나라지표

실업률

- 2013년 11월 실업률이 2.7%이다. 취업 가능한 사람 100명이 임의적으로 선택되었다.
  - 실업자로 예상되는 사람은 몇 명인가?
  - 실업자의 수에 대한 분산과 표준편차는 얼마인가?
  - 6명이 실업자일 확률은 얼마인가?

```
> n=100
> p=0.027
> dbinom(6,n,p)
[1] 0.03524398
> pnorm(6.5,n*p,sqrt(n*p))- pnorm(5.5,n*p,sqrt(n*p))
[1] 0.03381599
```