여러가지 확률분포

5조

01 정규분포

02 표준정규분포
정규성 판정

03 중심극한정리

04 기타 특이분포

여러가지 확률분포

이산형

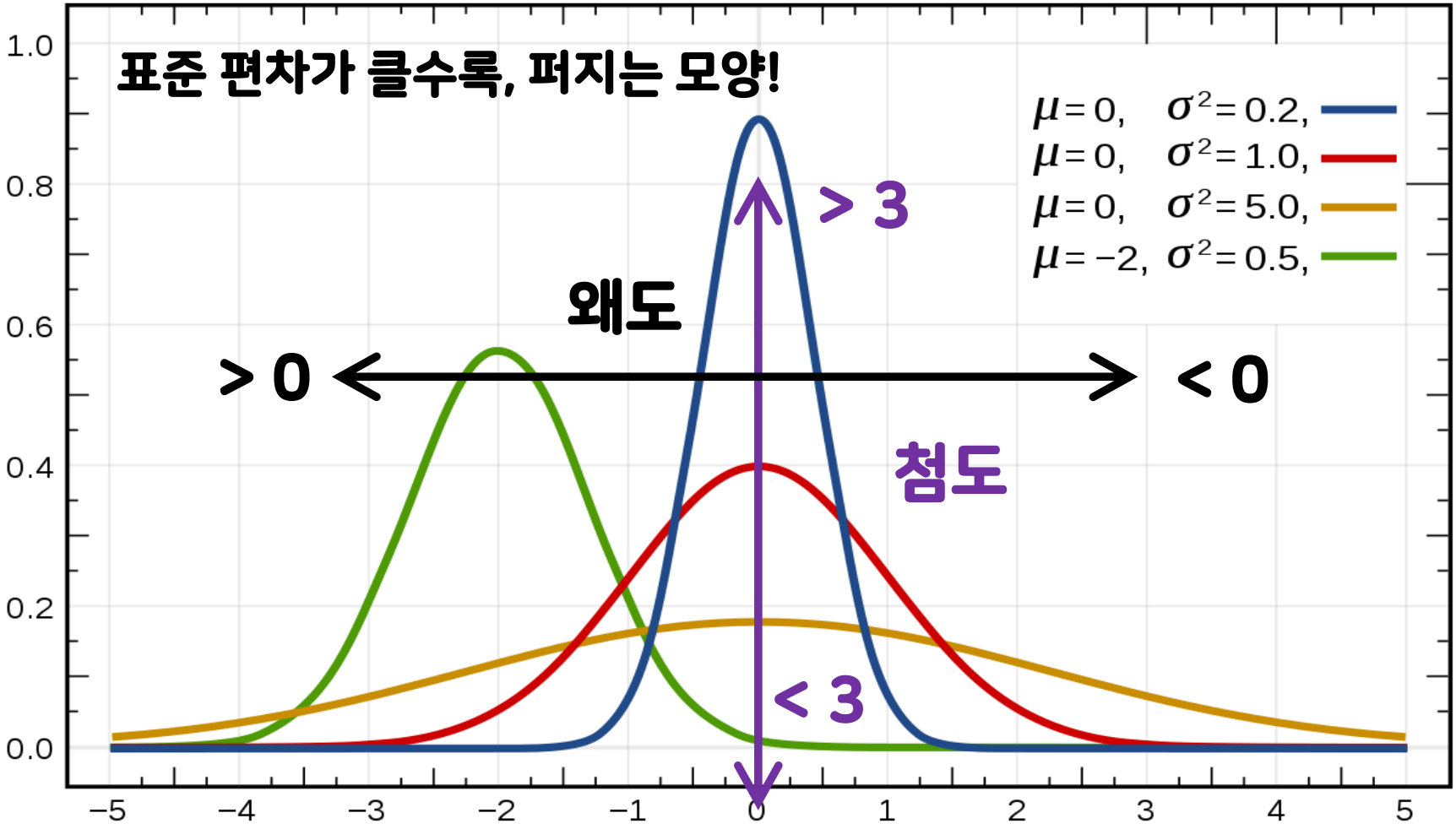
- 베르누이
- 이항분포
- 포아송분포
- 기하분포
- 음이항분포
- 초기하분포

연속형

- 균일분포
- 지수분포
- 정규분포
- 카이제곱분포
- t분포
- F분포

정규분포

세상의 거의 모든 분포

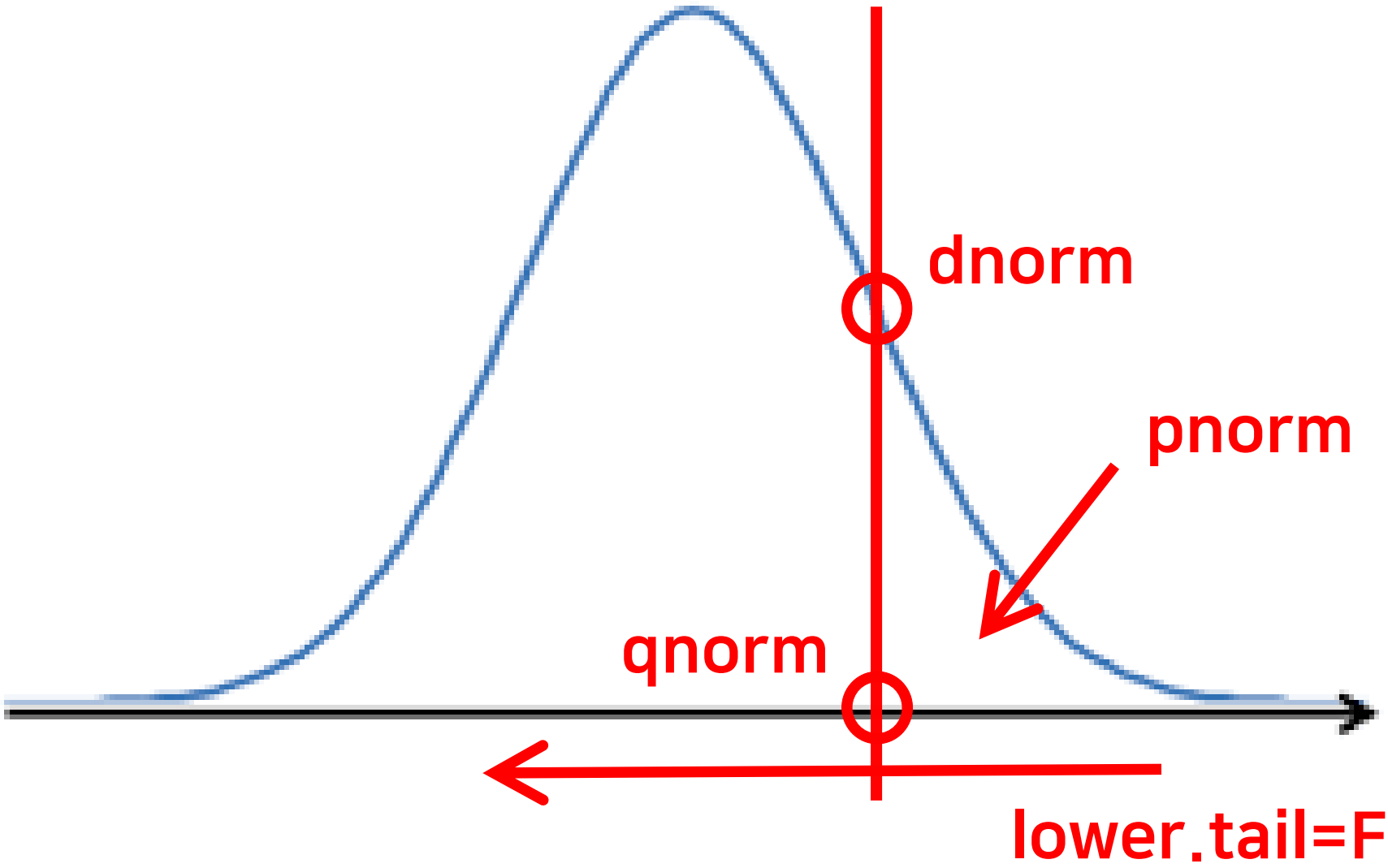


자료의 분포가 평균을 중심으로 대칭인 종모양을 이루는 분포 !

? norm

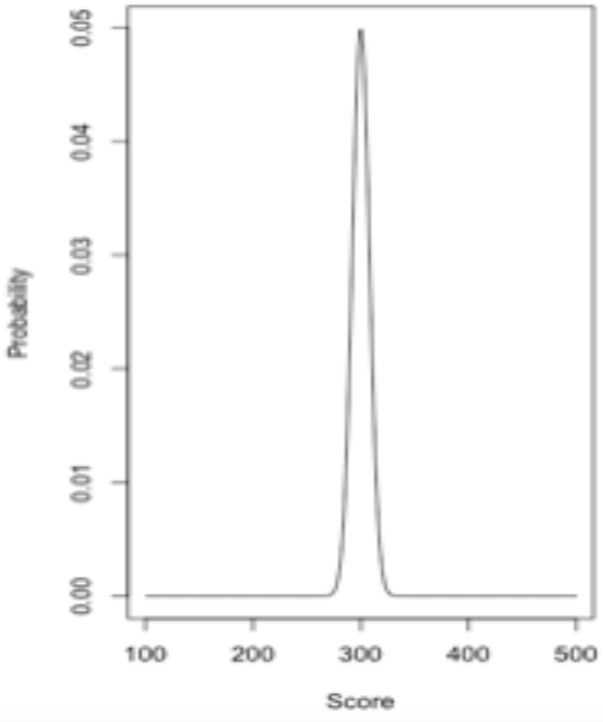
함수 구분		R 함수 및 Parameter
밀도함수 (Density function)	d	dnorm(x, mean, sd)
누적분포 함수 (Cumulative distribution function)	p	pnorm(q, mean, sd, lower.tail=T/F)
분위 수 함수 (Quantile function)	q	qnorm(p, mean, sd, lower.tail=T/F)
난수 발생 (Random number function)	r	rnorm(n, mean, sd)

lower.tail : 구하는 함수 값의 방향 설정
기본 설정 : T (양의 방향)


dnorm
pnorm
qnorm
lower.tail=F

d = probability Density = 확률을 계산(정규분포의 PDF)
p = cumulative Probability = 누적 확률을 계산(정규분포의 CDF)

#PDF(Probability Density Function)
#1
plot(x=100:500, y=dnorm(100:500, mean=300, sd=8), type = 'l',
xlab = 'Score', ylab = 'Probability')
#평균이 300, 표준편차가 8인 정규분포에서 x값이 100~500인 확률 값.



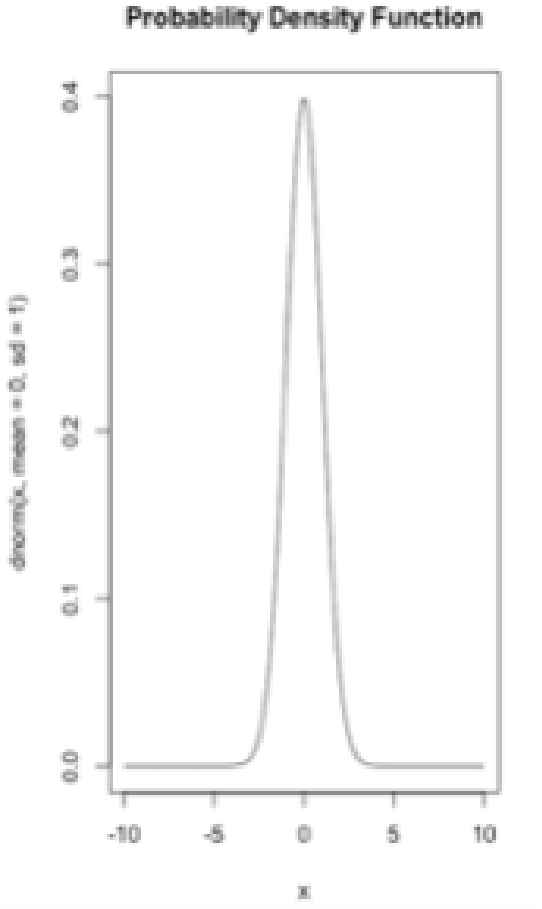
q = probability Quantility = 누적 확률의 inverse
#분위수
qnorm(p=0.025, mean = 0, sd=1, lower.tail = T)
qnorm(p=0.975, mean = 0, sd=1, lower.tail = T)
#따라서 95% 적중구간에 대한 값은 많이 나오는 -1.96 이상 +1.96 이하의 범위로 설명된다.

```
> qnorm(p=0.025, mean = 0, sd=1, lower.tail = T)
[1] -1.959964
> qnorm(p=0.975, mean = 0, sd=1, lower.tail = T)
[1] 1.959964
```

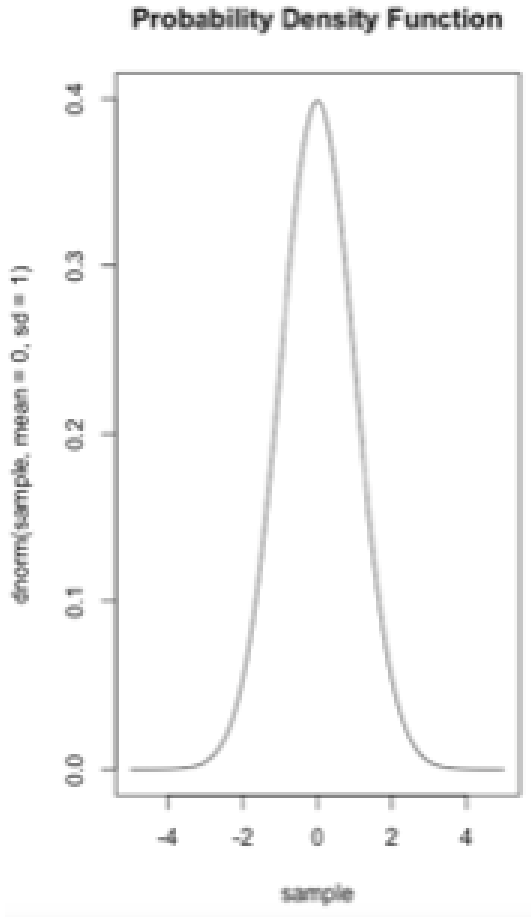
r = probability Random = 주어진 정규분포로 랜덤 값 생성
#난수
rnorm(100, mean = 0, sd = 1)

```
> rnorm(100, mean = 0, sd = 1)
  [1]  0.1425882274  0.3780482646  2.9651462017 -0.4355449371 -0.1693383399  0.6245904061 -0.4193200429
  [8]  1.1845970283  0.2791339820 -0.6543569823  1.1191789295 -0.2888360921  1.4088205185 -0.8858937289
 [15]  0.0848899141 -1.0081253749 -2.3328940705 -0.0007770748 -0.8061031685  0.4296193478  0.5412254021
 [22] -1.4266550068 -1.5088710892  0.3744251415  0.4244792353  0.4408818026 -0.1061013139 -0.2644580501
 [29]  0.9993653269  0.8019542862 -0.3112900683  0.3540460950 -0.2421727736  1.1422321881 -0.1161320862
 [36]  1.6926443354  0.2638518439 -1.8166480028  0.1667330890  0.1439148086 -0.3193549063 -0.8130390967
 [43]  0.1353488552 -0.9536357112 -0.8434297310 -1.4671398101  2.7348345035 -0.1568724990 -0.3557644010
 [50] -1.0169512716 -0.0988519893  0.3112801378 -1.5962753254  0.9349539729  2.4177745246  1.8529011082
 [57] -1.7134403982 -1.0197860035  1.0417327500  0.1004212148 -0.8133341436 -0.1512747681 -0.5695733479
 [64] -1.6030258965  1.2376885062 -1.6430004870  1.5255186753 -0.0145190705 -0.0504484156 -1.2415377465
 [71]  1.1986881832  1.1528452194  1.2008189500  0.4199886526  1.5904246382  0.4633569236  0.4289453825
```

#2
```
x <- seq(-10,10, length = 200)
plot(x, dnorm(x, mean=0, sd=1), type = 'l', main="Probability Density Function")
```



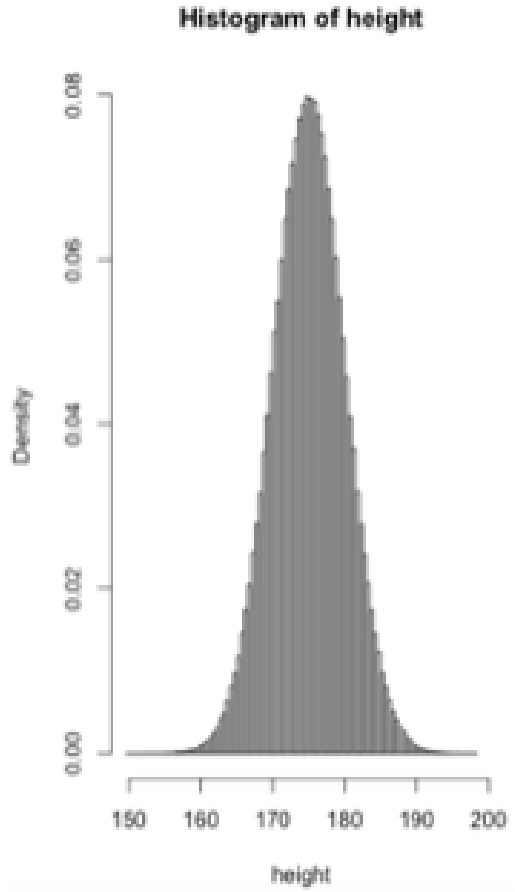
#3
```
sample <- seq(-5,5, length = 101)
plot(sample, dnorm(sample, mean = 0, sd=1),type = 'l', main = "Probability Density Function")
```



#4
#확률밀도함수
height=rnorm(n=1000000, mean=175, sd=5) #데이터생성
hist(height, breaks=100, probability = T)

lines(density(height))

(line을 넣고 안 넣고의 차이.)



표준정규분포

정규분포의 표준화!

평균이 0이고 표준편차가 1인 정규분포

$$X \qquad Z=\frac{X-m}{\sigma} \qquad Z$$

$$N(m,\sigma^2) \Longrightarrow N(0,1^2)$$

정규분포 — 표준화 → 표준정규분포

$$P(a \leqq X \leqq b) \Rightarrow P\left(\frac{a-m}{\sigma} \leqq Z \leqq \frac{b-m}{\sigma}\right)$$

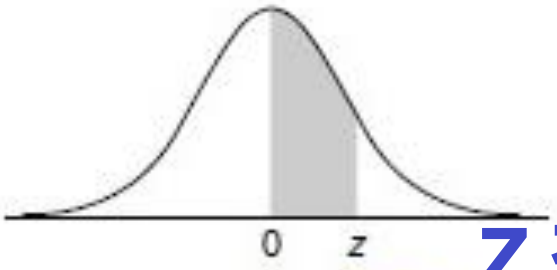
표준정규분포표

Example

0과 1.55사이의 확률!

Area between 0 and z

$P(0<Z<1.55)$



Z 값

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406

확률 값


무한도전
MBC
늦었다고 생각할 때가
진짜 너무 늦었다

정규성 판정

정규성을 어떻게 판단하는가?

함수 구분	R 함수
왜도	skewness()
첨도	kurtosis()
왜도 검정	anscombe.test()
첨도 검정	agostino.test()
정규성 검정	shapiro.test(), qqplot(), qqnorm(), qqline()

왜도의 절댓값이 3이상 , 첨도가 10 이상이면 정규성을 만족하지 않는다!

정규성 판정

정규성을 어떻게 판단하는가?

```
> set.seed(7777)
> data=rnorm(100,mean=0,sd=1)
> library(moments)
> skewness(data)
[1] 0.3609106
> kurtosis(data)
[1] 3.383699
```

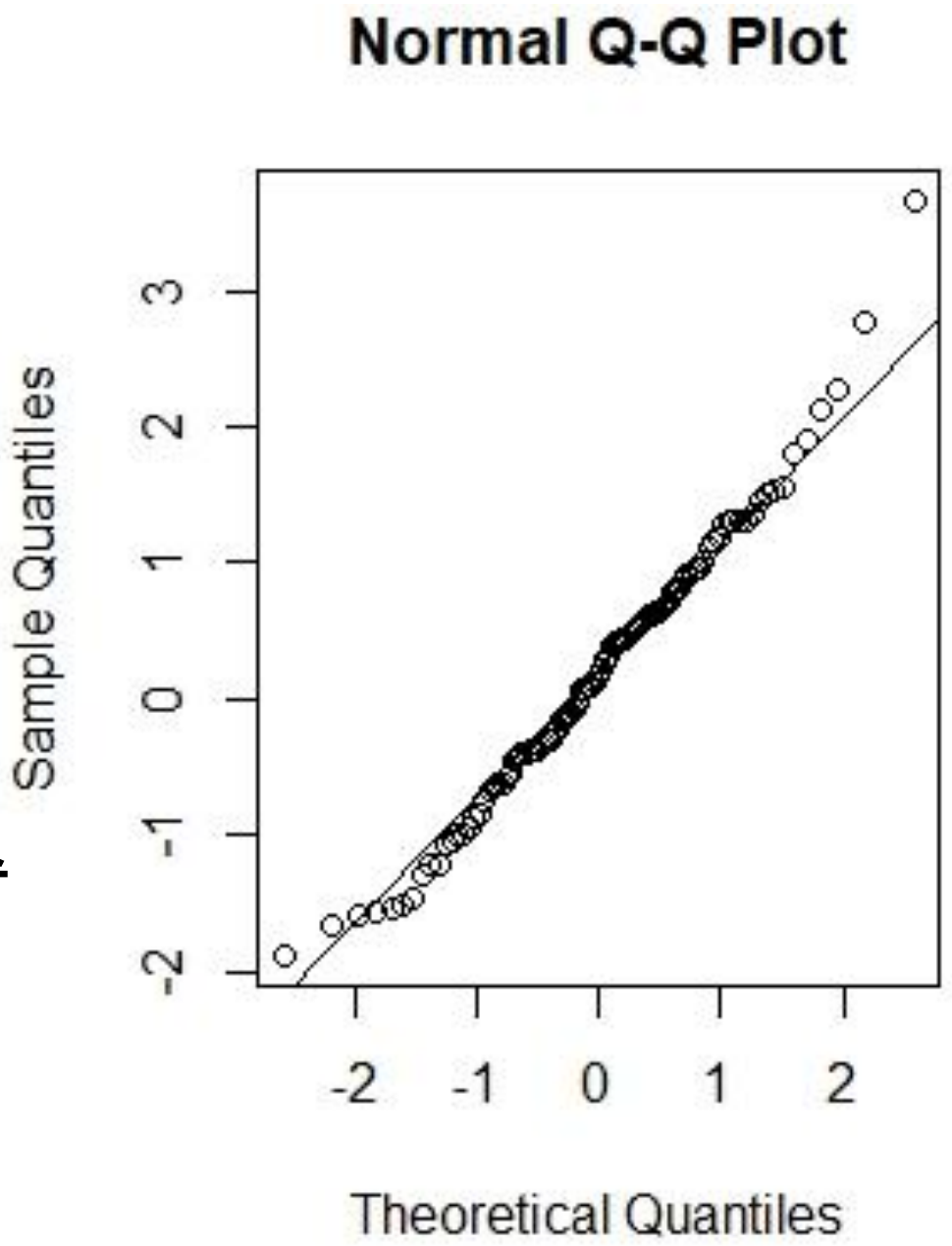
```
> shapiro.test(data)
Shapiro-Wilk normality test
data: data W = 0.98521,
p-value = 0.3288
```

```
> agostino.test(data)
D'Agostino skewness test
data: data skew = 0.36091,
z = 1.52958, p-value = 0.1261
alternative hypothesis:
data have a skewness
```

```
> anscombe.test(data)
Anscombe-Glynn kurtosis test
data: data kurt = 3.3837,
z = 1.0547, p-value = 0.2916
alternative hypothesis:
kurtosis is not equal to 3
```

왜도의 절댓값이 3이상 , 첨도가 10 이상이면 정규성을 만족하지 않는다!

```
> qqnorm(data)
> qqline(data)
```



x축 : 표준정규분포의 분위수
y축 : 표본데이터의 분위수

qqnorm은 정규성 검정
qqplot은 두개의 데이터 비교!

중심극한정리

N이 커지면 정규분포로 수렴한다!

크기가 n인 확률표본 $X_1, X_2, \cdots, X_n$이

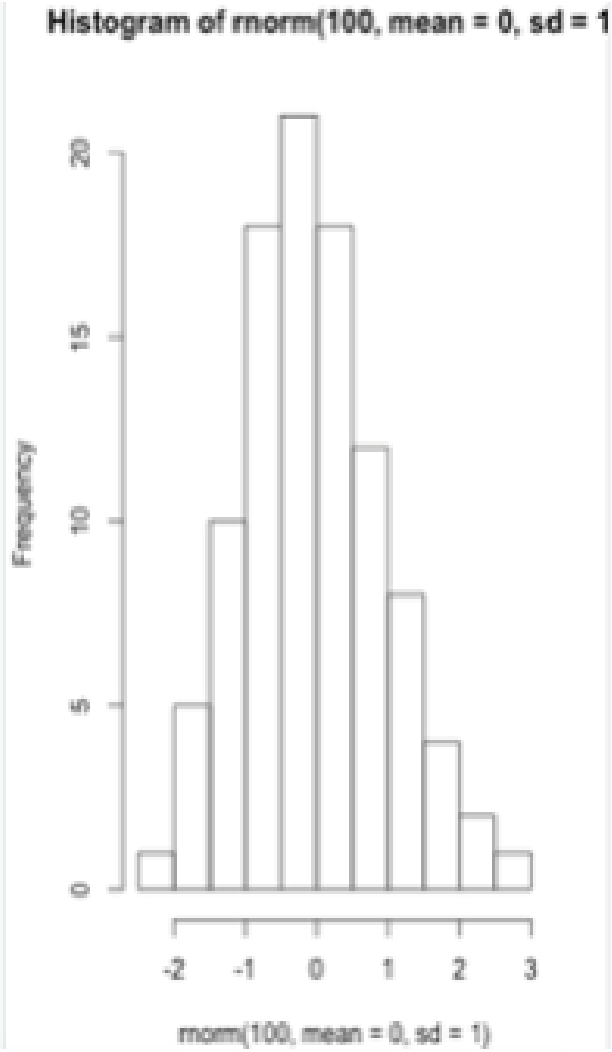
$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \cdots, n$$

이면, 표본평균 $\overline{X}$의 확률분포는 다음과 같은 정규분포를 따르게 된다.

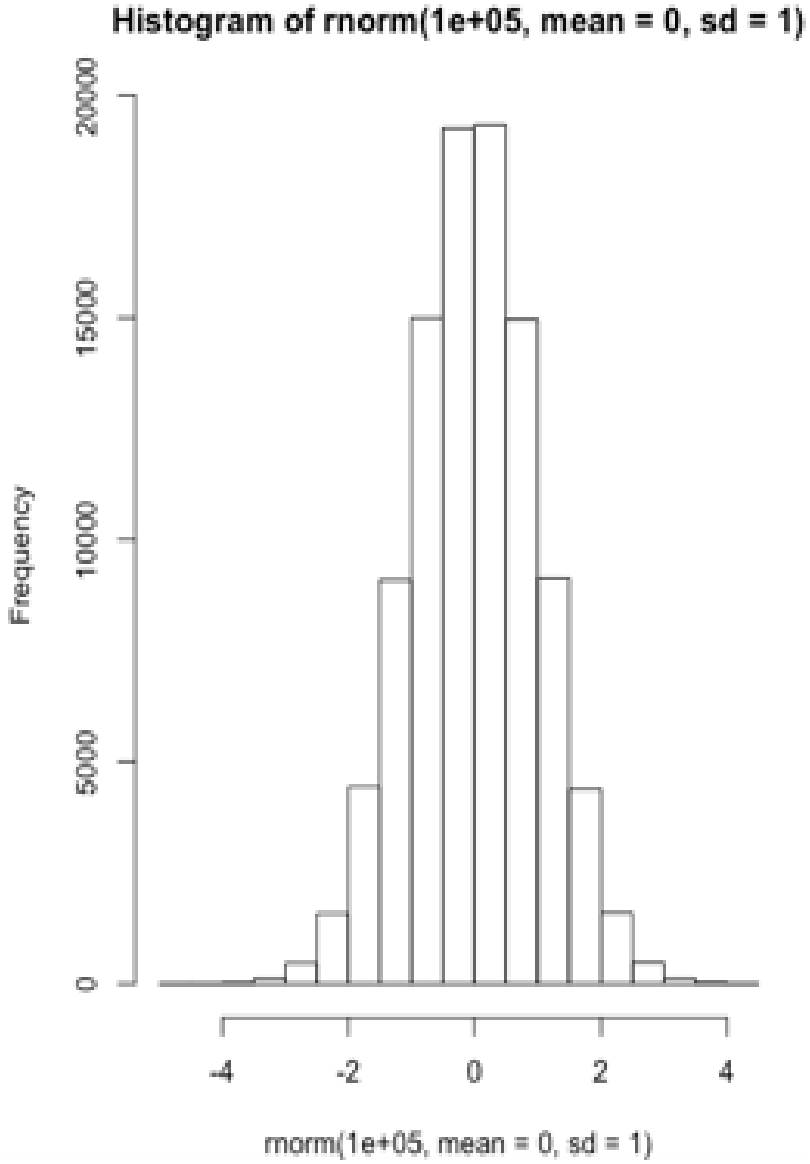
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Example 1

```
hist(rnorm(100, mean = 0, sd = 1))
```



hist(rnorm(100000, mean = 0, sd = 1))
#중심극한 정리에 의해 랜덤 개수 n이 증가할 수록 점차 표준 정규분포에 가까운 hist가 그려지는것을 확인할 수 있다.



Example 2

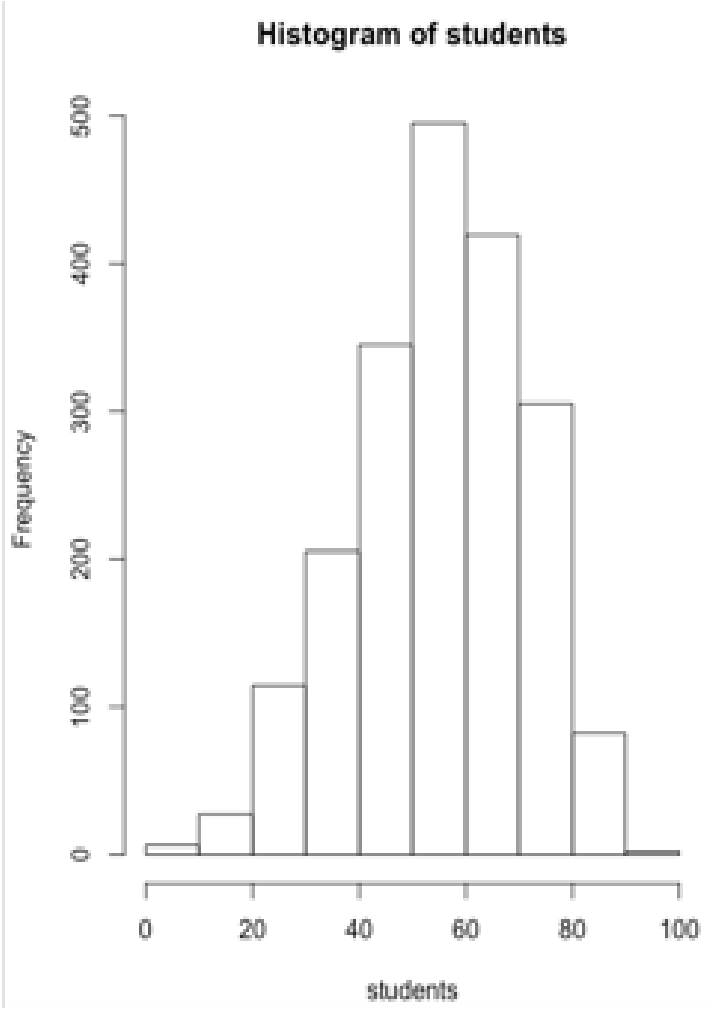
#예를들어 한 반의 학생이 40명인데 시험을 보았다. 시험은 100점 만점이다.
#3명에게 무작위로 점수를 물어보고 평균을 낸 다음, 이와 같은 행위를 2000번 정도 하면 표본의 평균이 2000개가 생기게 된다.
#이 2000번에 대한 분포를 히스토그램으로 보면 평균에 해당하는 히스토그램 막대가 가장 높이 나타난다.
#1
set.seed(0529) #난수 시드를 설정
score<-sample(1:100, 40, replace=T) #1~100까지 40명에 대해 난수의 값을 할당해준다.
score
mean(score)#표본에 대한 평균과 분포를 확인한다.

```
> #1
> set.seed(0529) #난수 시드를 설정
> score<-sample(1:100, 40, replace=T) #1~100까지 40명에 대해 난수의 값을 할당해준다.
> score
 [1] 78 45 29 73 83 87 49  9 62  3 80 91  9 58 87 56 87 27  3 44 79 81 29 47 41 23 71 15 96 72 67
[32]  4 80 65 43 70 76 58 92 70
> mean(score)#표본에 대한 평균과 분포를 확인한다.
[1] 55.975
```

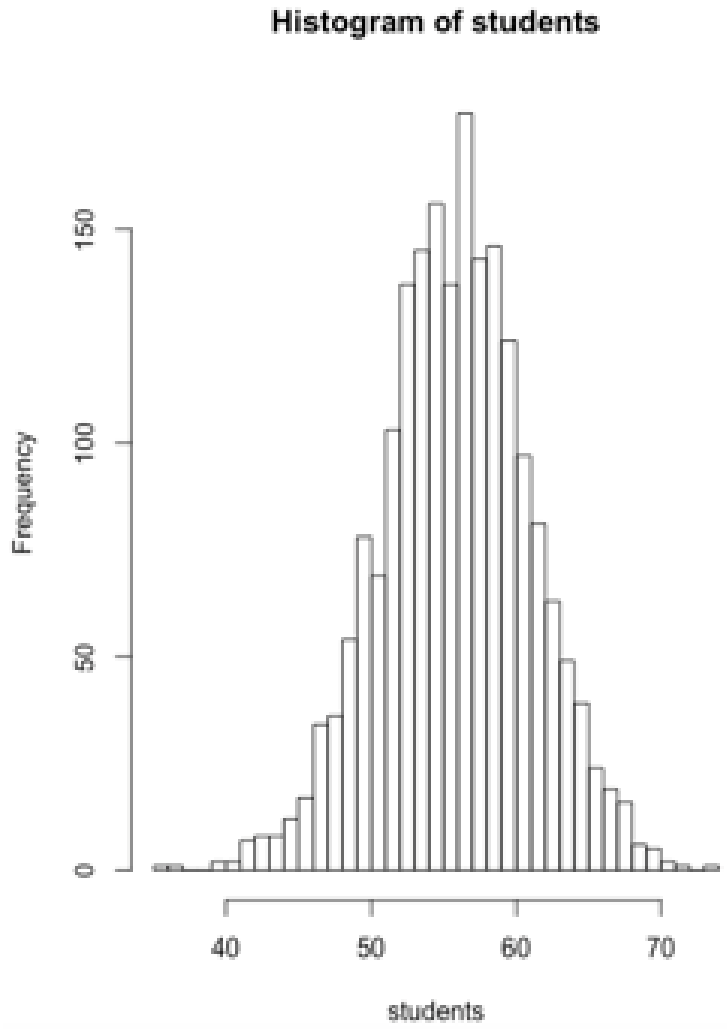
```
students <- NULL
students <- sapply(1:2000,
function(i)mean(sample(score,3, replace=T)))
hist(students)  #3명을 복원추출로 뽑아서 평균을 계산하
고, 이를 2000번 반복한다.

mean(students)
```

```
> mean(students)
[1] 55.42283
```



```
#이번에는 30명을 복원추출로 뽑아서 평균을 계산하고, 이를 2000번 반복한다.
students <- NULL
students <-  sapply(1:2000,
function(i)mean(sample(score, 30, replace=T)))
hist(students, breaks = 30)
mean(students)
```



기타 다른 분포 :
포아송 분포 및 정규근사 X ~ Poisson(lamda)

단위 시간 (공간) 당
사건 발생 횟수(이산적인 횟수)

- 확률질량함수: P(μ)

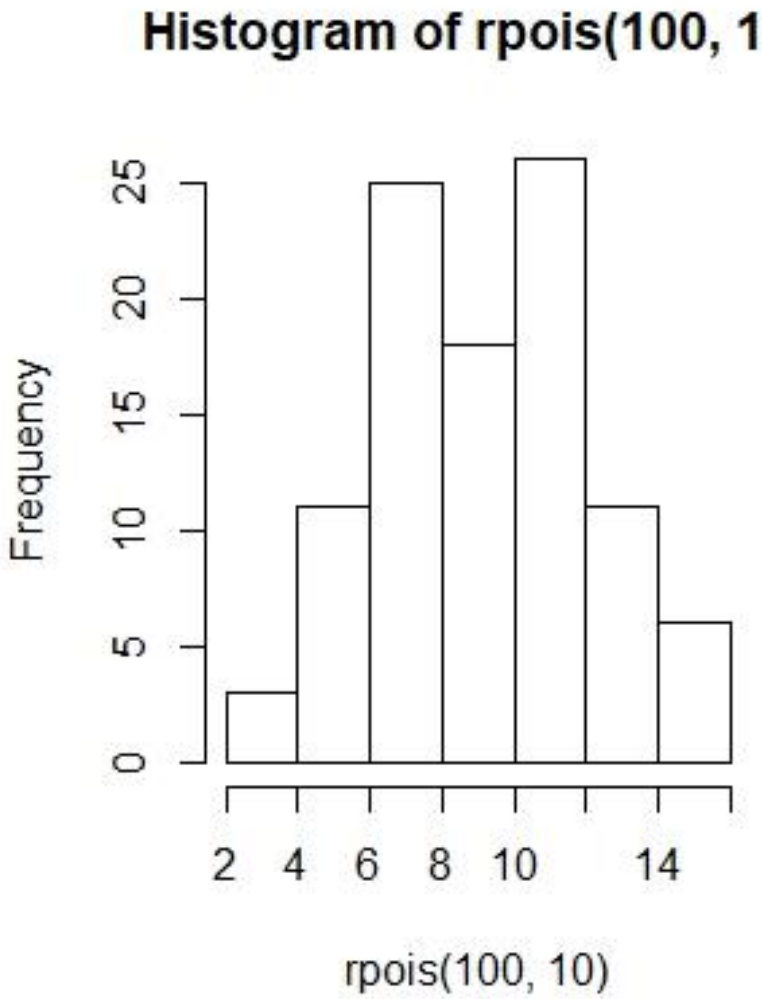
$$p(X=x)=f(x)=\frac{e^{-\mu}\mu^{x}}{x!},\ x=0,1,2\ldots$$

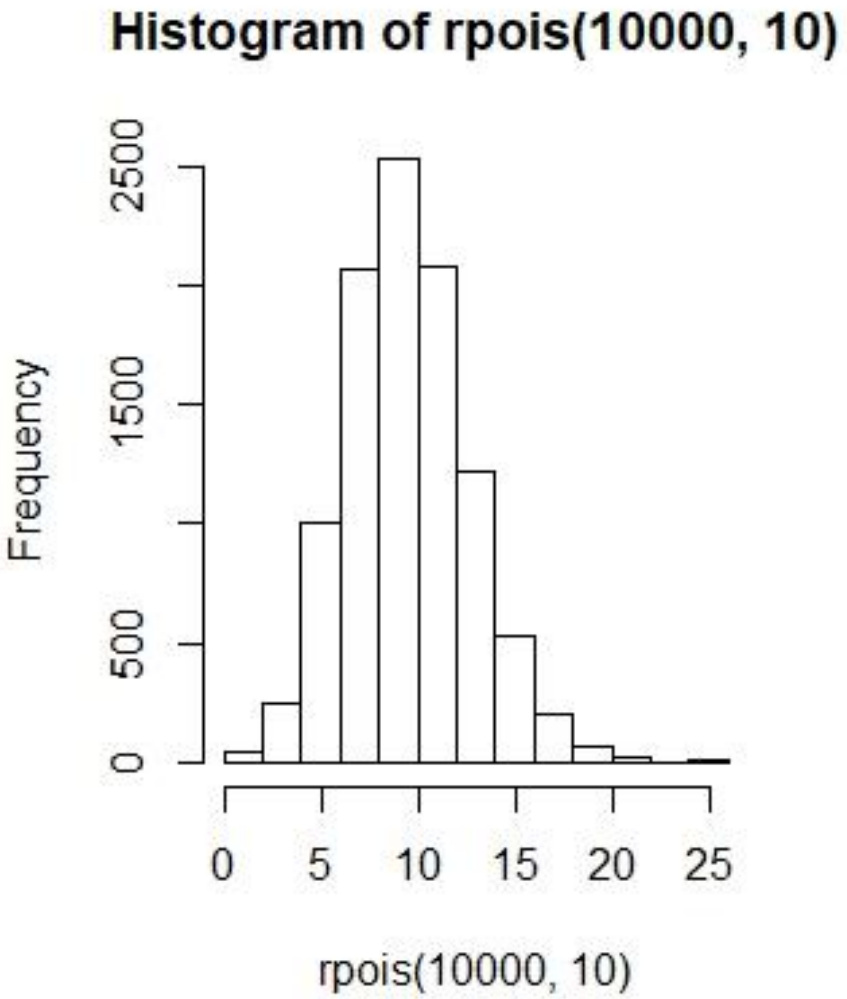
- 평균과 분산

$$\mu=E(X)=\mu,\quad \sigma^{2}=Var(X)=\mu$$

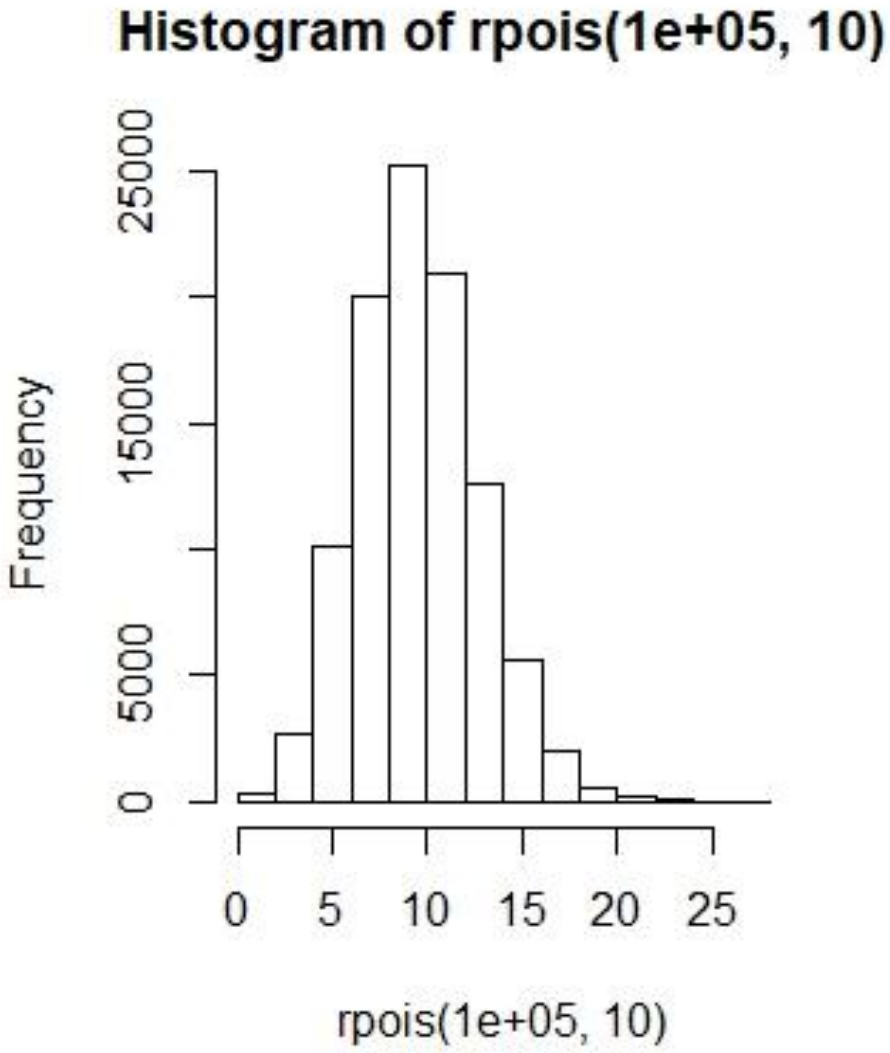
표준화! lamda>5

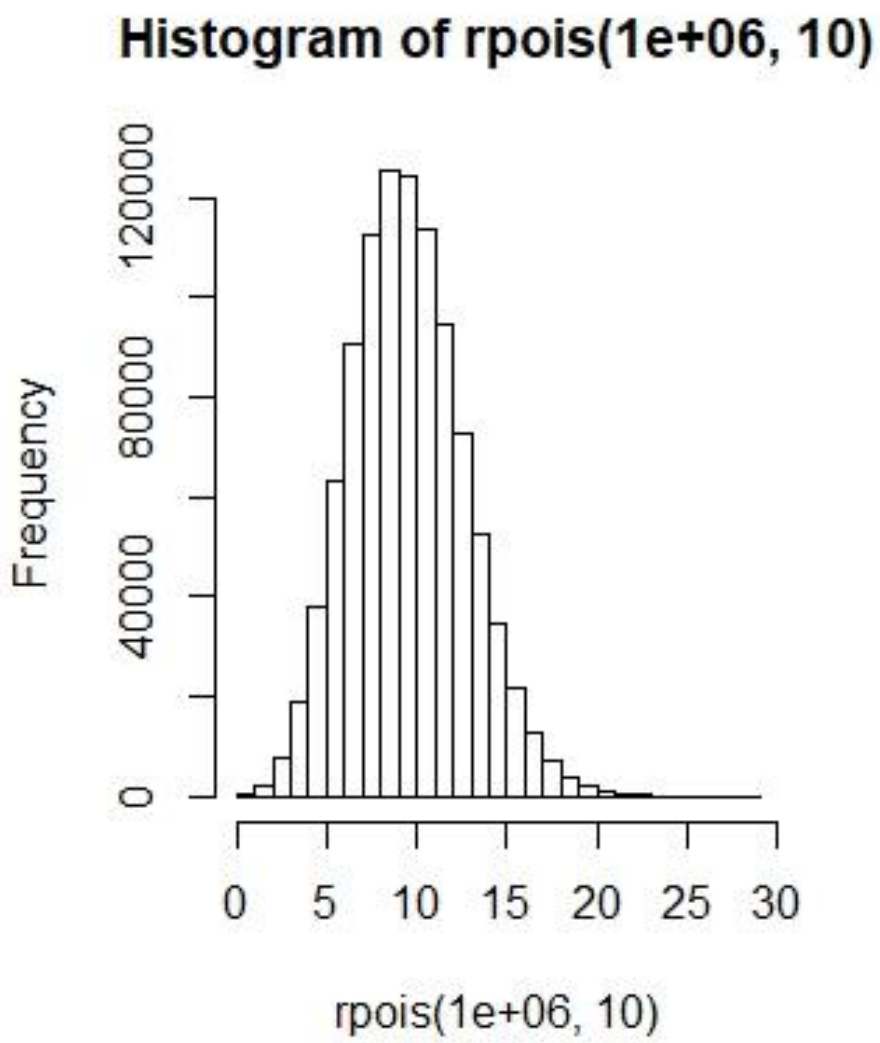
```
> hist(rpois(100,10))
> hist(rpois(10000,10))
```





```
> hist(rpois(100000,10))
> hist(rpois(1000000,10))
```

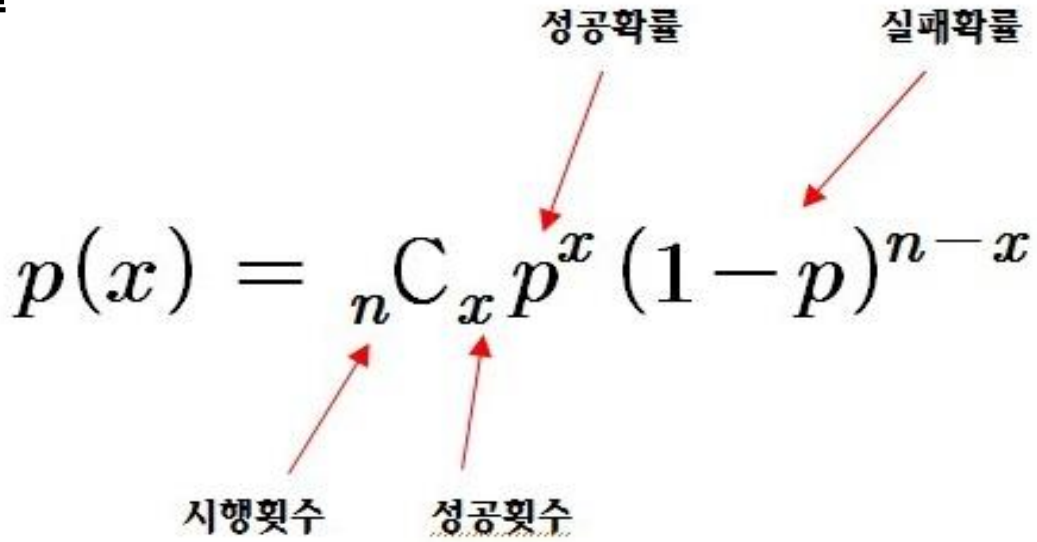




기타 다른 분포 : 이항분포 및 정규근사

$$X \sim Bin(n, p)$$

p의 성공확률을 가진 사건을
n번 독립시행시 성공횟수



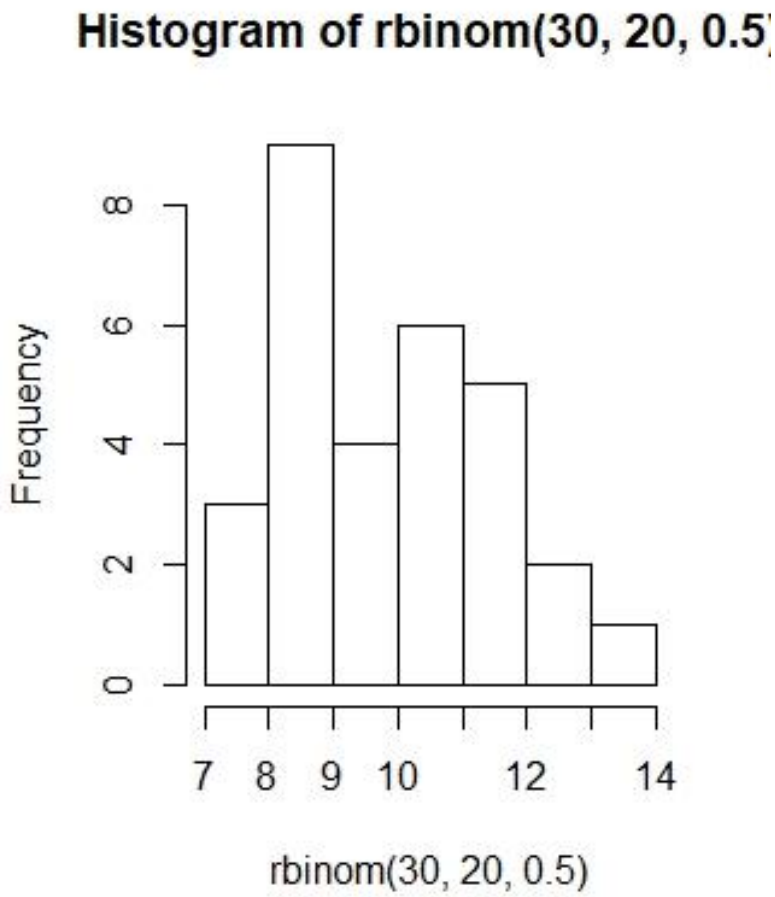
$$E(x) = np$$

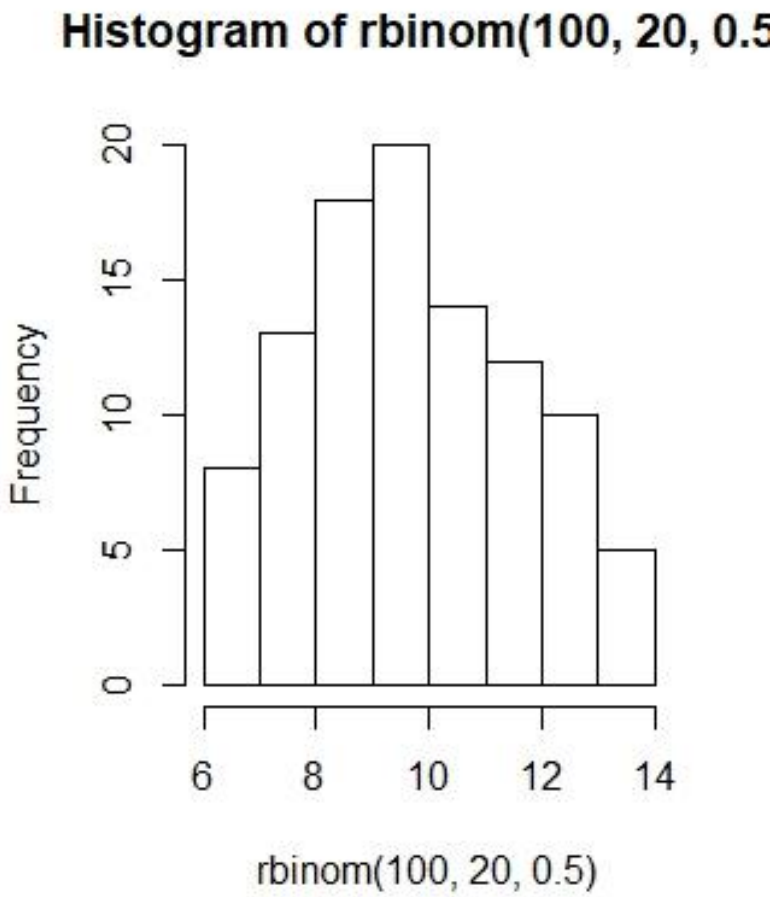
$$Var(x) = np(1-p)$$

표준화!

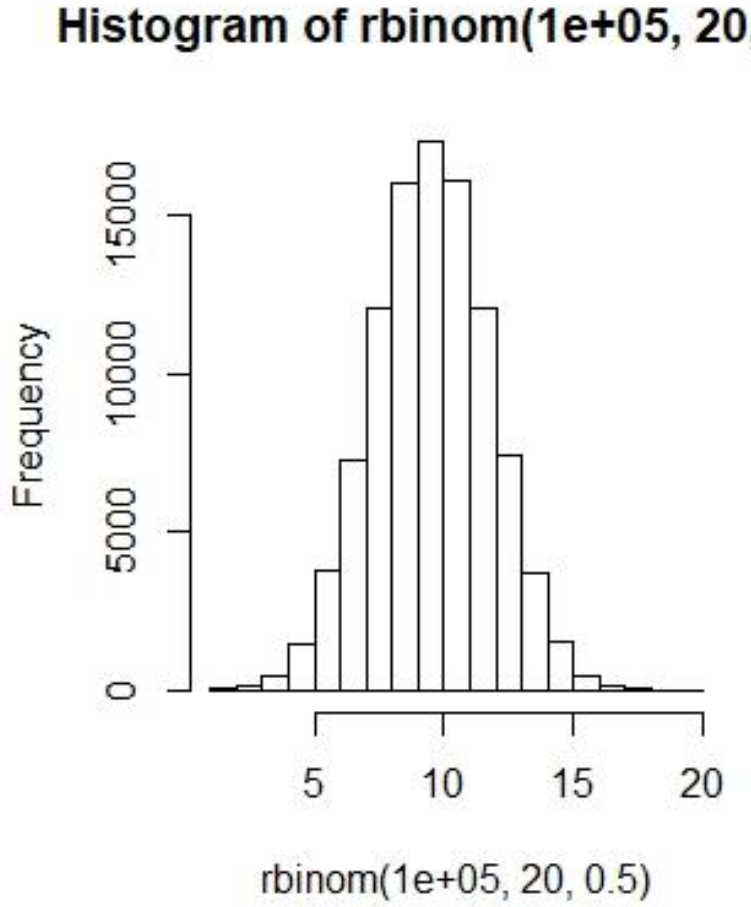
np > 5
n(1-p) > 5

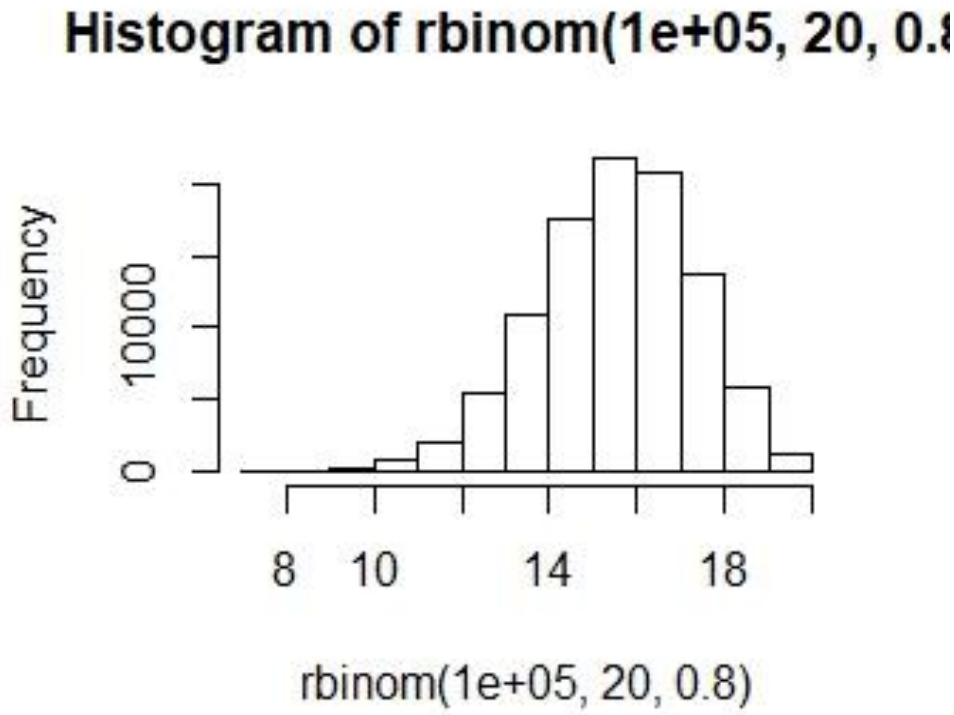
```
> hist(rbinom(30,20,0.5))
> hist(rbinom(100,20,0.5))
```



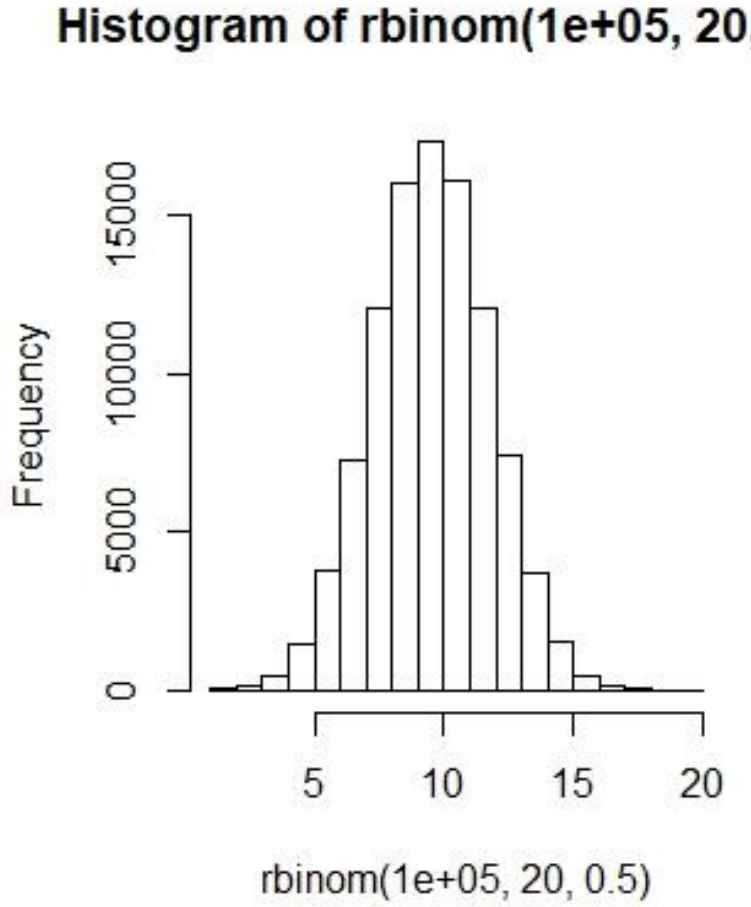


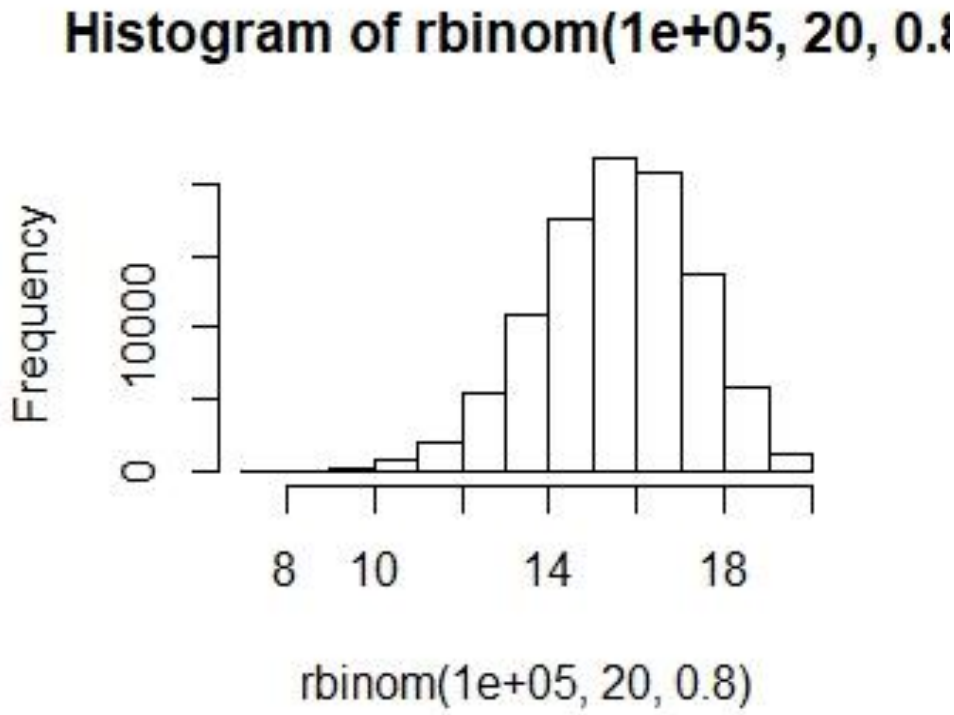
```
> hist(rbinom(100000,20,0.5))
> hist(rbinom(100000,20,0.9))
```





```
> hist(rbinom(100000,20,0.5))
> hist(rbinom(100000,20,0.9))
```





감마분포

a번 사건이 발생할 때 까지 대기시간

$$X \sim \Gamma(\alpha, \beta)$$

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \qquad x > 0,\ \alpha > 0,\ \beta > 0$$

$$0 \qquad \textit{elsewhere}$$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{x}\, dx \qquad \alpha > 0$$

$$E(X) = \alpha\beta \qquad V(X) = \alpha\beta^2$$

Γ(α)의 성질

1) $\Gamma(1) = 1$

2) $\Gamma(n) = (n-1)!$ $\quad n$: 자연수

3) $\Gamma(\alpha) = \alpha\,\Gamma(\alpha-1)$ $\quad \alpha > 1$

카이제곱분포

모분산 검정 시 사용

$$X \sim \chi^2(k)$$

k : 자유도

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}} \qquad x > 0$$

$$\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2 \sim \chi^2(k)$$

$$E(X) = \alpha\beta = 2 \cdot \frac{r}{2} = r$$

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1^2)$$

$$V(X) = \alpha\beta^2 = 2^2 \cdot \left(\frac{r}{2}\right) = 2r$$

$$X_i \sim N(\mu, \sigma^2)$$

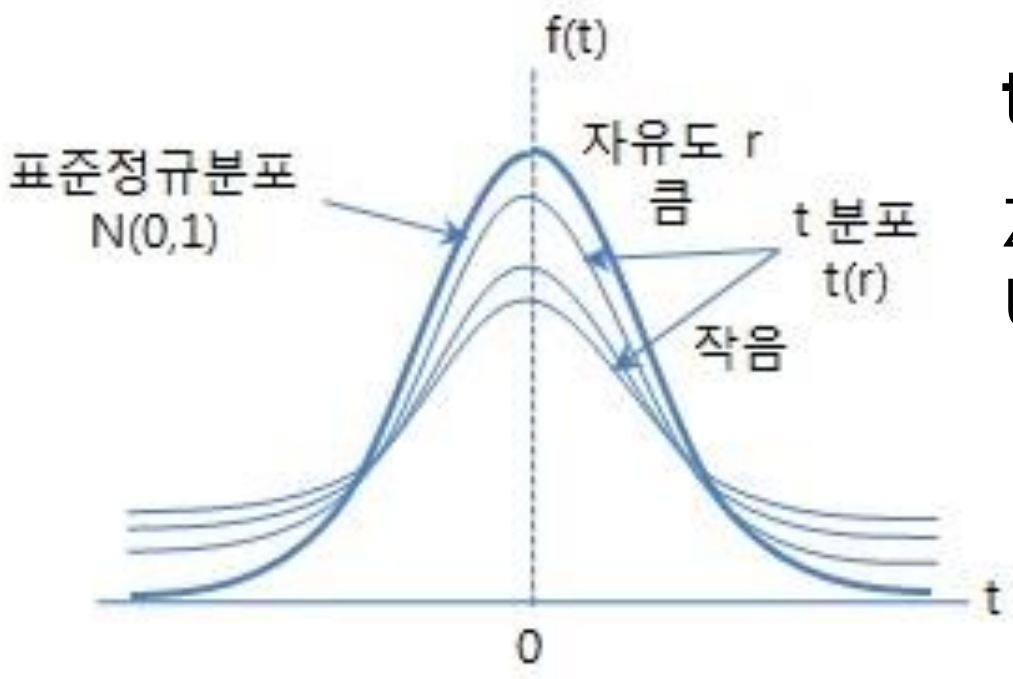
제곱합으로 정의되는 점이
표본분산과 연관성 있음!

T-분포

$t \sim (n-1)$

소표본에서 정규분포 따르는 집단의 평균에 대한 가설검정

두 집단의 평균 차이 검정



$t = Z / \sqrt{U/k}$

$Z \sim N(0,1)$

$U \sim \chi^2(k)$

자유도가 커질 수록
표준정규분포에 수렴!

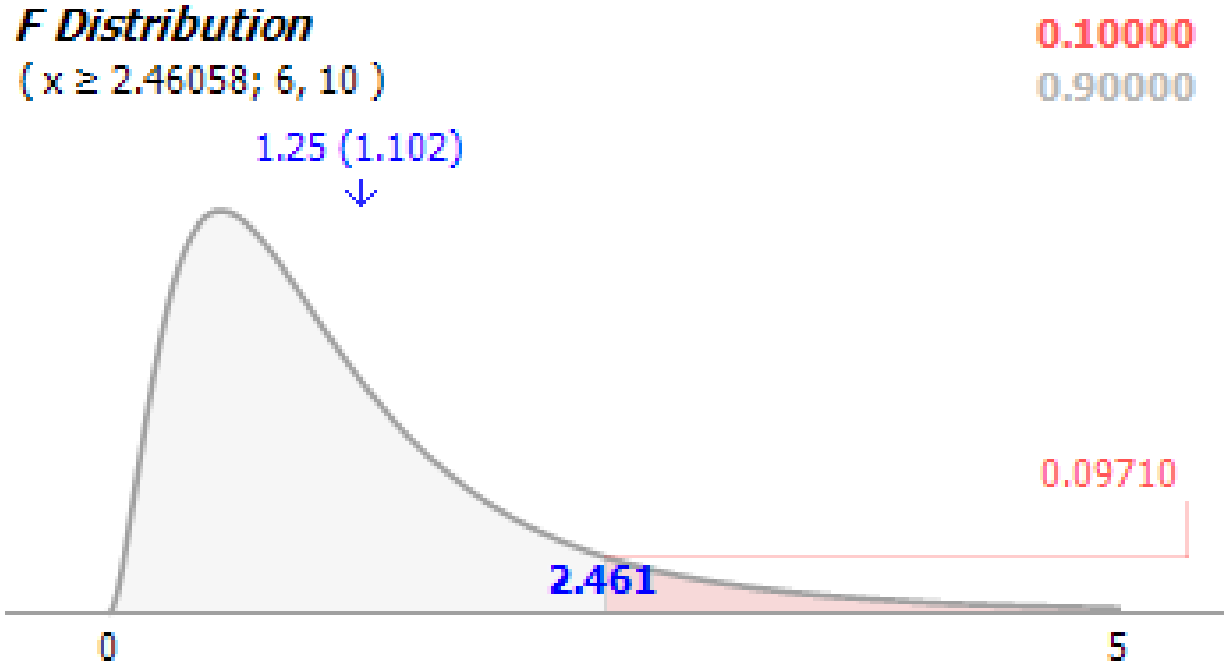
F 분포

X~ F(m,n)

집단 간 분산 비 검정 (ANOVA)

F = (U/m) / (V/n)

U,V는 자유도 m,n을 따르는 카이제곱분포



Thank you