소프트웨어시스템 실습

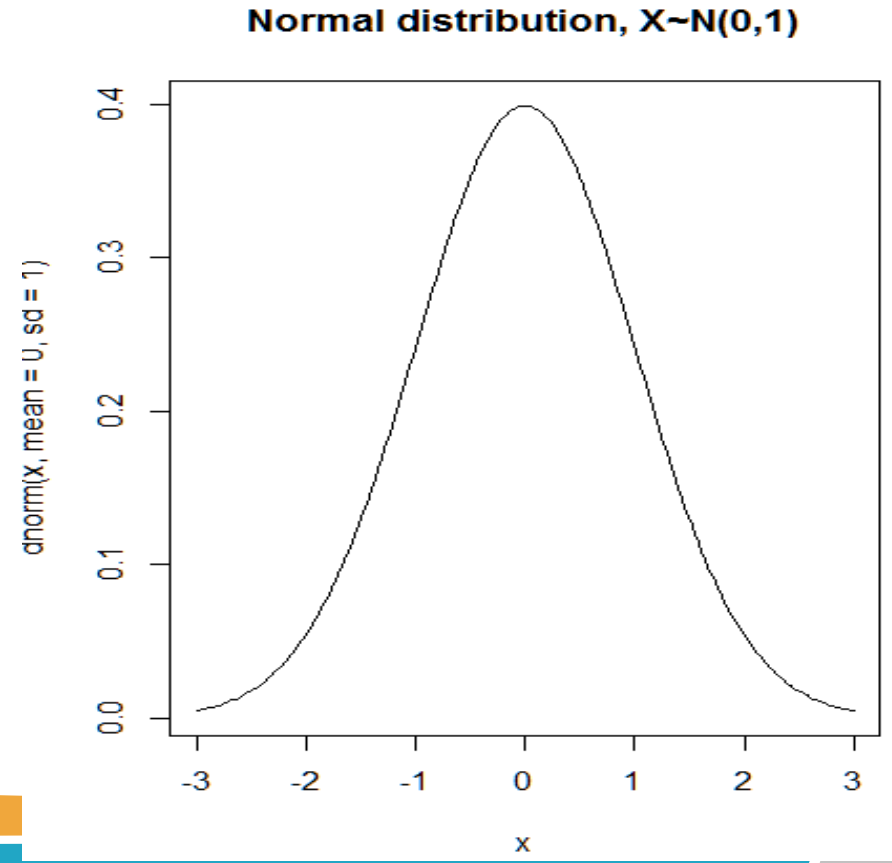
통계분석

난수생성 및 분포함수

확률 분포	난수	확률 밀도	누적 분포	분위수
이항분포(Binomial)	rbinom	dnorm	pnorm	qnorm
F 분포(F)	rf	df	pf	qf
기하분포(Geometric)	rgeom	dgeom	pgeom	qgeom
초기하분포(Hypergeometric)	rhyper	dhyper	phyper	qhyper
음이항분포(Negative Binomial)	rnbinom	dnbinom	pnbinom	qnbinom
정규 분포(Normal)	rnorm	dnorm	pnorm	qnorm
포아송 분포(Poisson	rpois	dpois	ppois	qpois
t 분포(Student t)	rt	dt	pt	qt
연속 균등 분포(Uniform)	runif	dunif	punif	qunif

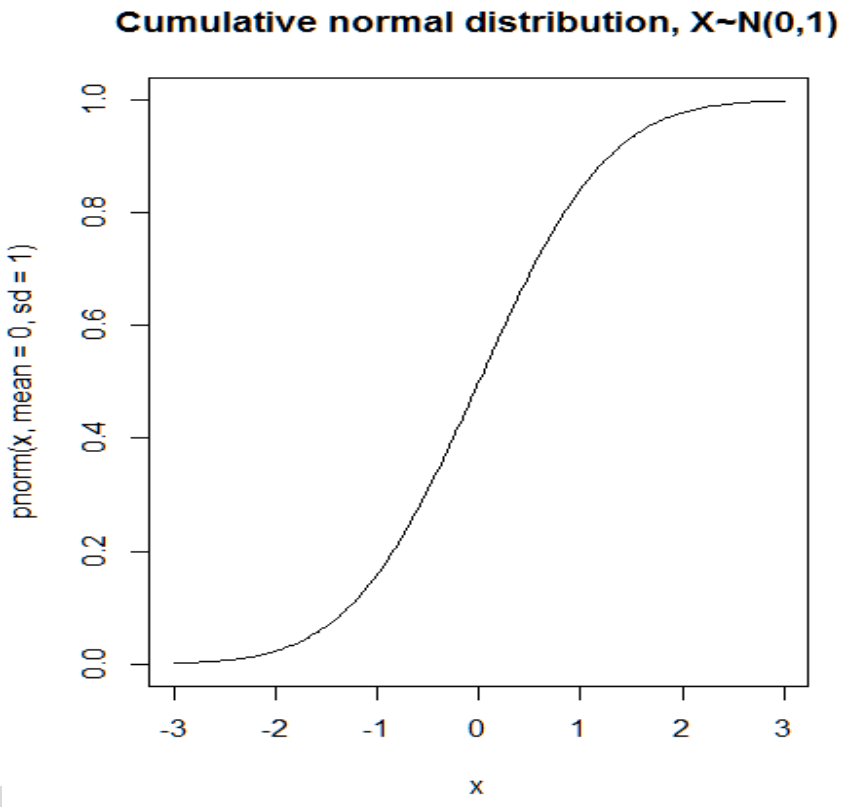
난수생성 및 분포함수

```
x <- seq(-3, 3, length=200)
plot(x, dnorm(x, mean=0, sd=1), type='l',
                    main="Normal distribution, X~N(0,1)")
```



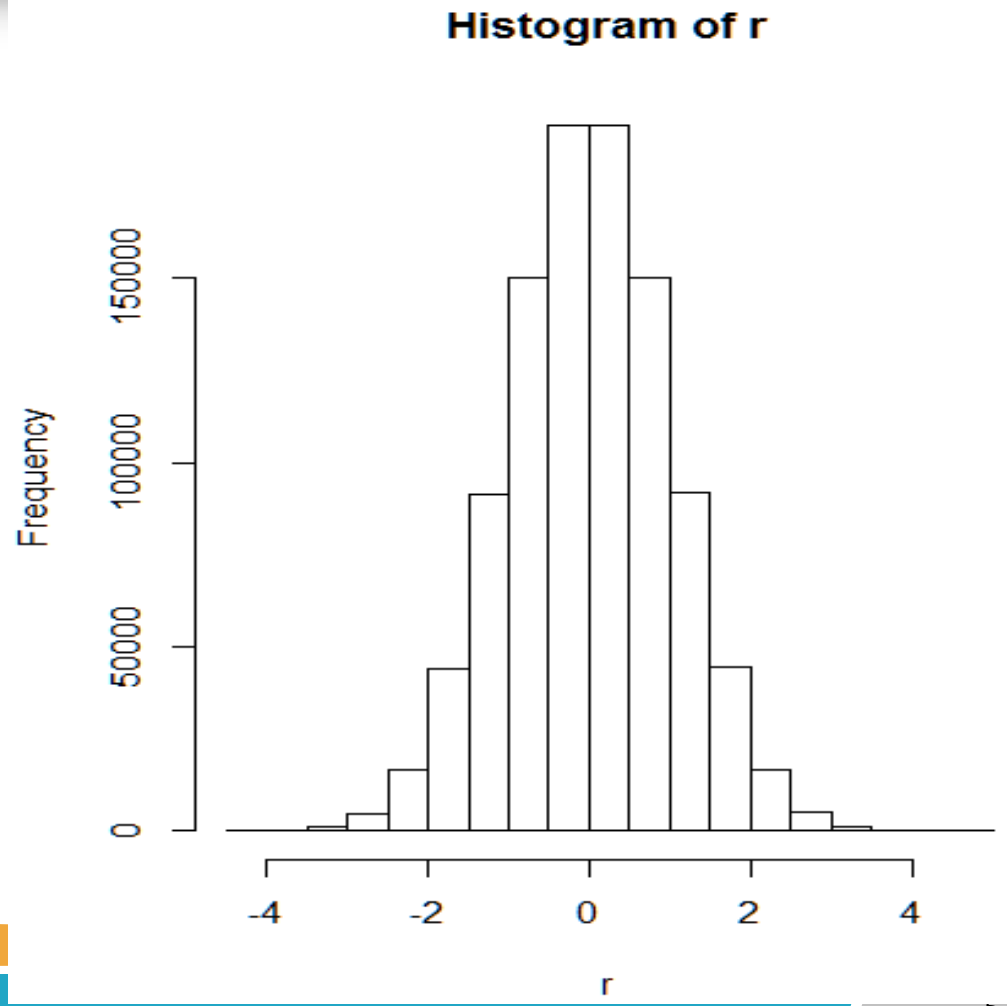
난수생성 및 분포함수

```
x <- seq(-3, 3, length=200)
plot(x, pnorm(x, mean=0, sd=1), type='l',
        main="Cumulative normal distribution, X~N(0,1)")
```



난수생성 및 분포함수

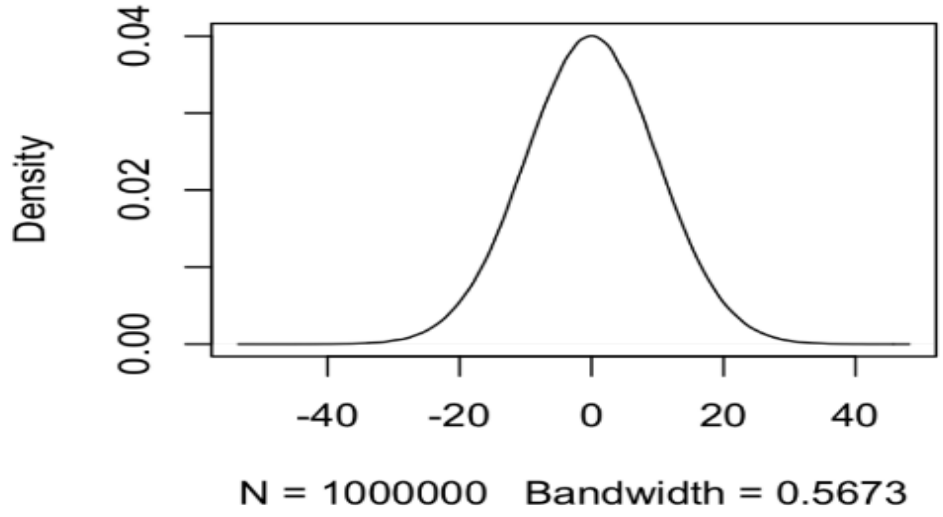
```
> r <- rnorm(1000000, mean=0, sd=1)
> hist(r)
```



난수생성 및 분포함수

```
> rnorm(100, 0, 10)
  [1]    6.35522264 -15.91675609    0.11219825    2.81311412    8.94825134
  ...
 [96]   -4.70195484   12.33659335 -15.98517300 -13.41173703 -12.91536521
```

```
> plot(density(rnorm(1000000, 0, 10)))
```



기초통계량

평균, 분산

```
> mean(1:5)
[1] 3
> var(1:5)
[1] 2.5
> sum((1:5-mean(1:5))^2)/(5-1)
[1] 2.5
```

N-1

다섯 수치 요약

```
> fivenum(1:11)
[1]  1.0  3.5  6.0  8.5 11.0
> summary(1:11)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     3.5     6.0     6.0     8.5    11.0
```

최빈값

- 가장 자주 나타나는 값

```
> x <- factor(c("a", "b", "c", "c", "c", "d", "d"))
> x
[1] a b c c c d d
Levels: a b c d
> table(x)
x
a b c d
1 1 3 2
> which.max(table(x))
c
3
> names(table(x))[3]
[1] "c"
```

분할표(table) 작성
(각 값에 대한 빈도수)

숫자벡터에서 최대값

분할표

- table()

```
> table(c("a", "b", "b", "b", "c", "c", "d"))
a b c d
1 3 2 1
```

```
> table(c('a','b','a', 'b', 'b', 'b'), c(1,2,1,2,1,2))

   1 2
  a 2 0
  b 1 3
>
```

표본추출

```
> sample(1:10, 5)
[1]  4  5  6 10  9
```

```
> sample(1:10, replace=TRUE)
 [1] 6 7 8 7 4 7 4 3 5 1
```

중복 허용

표본 추출

```
> install.packages("sampling")
> library(sampling)
> x <- strata(c("Species"), size=c(3, 3, 3), method="srswor",
+               data=iris)
> x
        Species ID_unit Prob Stratum
10       setosa      10 0.06       1
20       setosa      20 0.06       1
31       setosa      31 0.06       1
66   versicolor      66 0.06       2
75   versicolor      75 0.06       2
76   versicolor      76 0.06       2
123   virginica     123 0.06       3
125   virginica     125 0.06       3
138   virginica     138 0.06       3

> getdata(iris, x)
    Sepal.Length Sepal.Width Petal.Length Petal.Width     Species
10           4.9         3.1          1.5         0.1      setosa
20           5.1         3.8          1.5         0.3      setosa
```

표본 추출

```
> strata(c("Species"), size=c(3, 1, 1), method="srswr", data=iris)
        Species ID_unit Prob Stratum
5        setosa       5 0.06       1
38       setosa      38 0.06       1
46       setosa      46 0.06       1
89   versicolor      89 0.02       2
116   virginica     116 0.02       3
```

최빈값 및 분할표

```
library(arules)
data("AdultUCI")
str(AdultUCI)
attach(AdultUCI)
table(sex)
table(occupation)
which.max(table(occupation))
names(table(occupation))
table(race)
table(sex, race)
```

분할표

- **xtabs(formula, data)**
  - x, y라는 두가지 속성이 있고 (x, y)에 대한 도수가 num에 저장되어 있을 때 formula는 num ~ x+ y

```
> d <- data.frame(x=c("1", "2", "2", "1"),
+                 y=c("A", "B", "A", "B"),
+                 num=c(3, 5, 8, 7))
> d
  x y num
1 1 A   3
2 2 B   5
3 2 A   8
4 1 B   7
> xt <- xtabs(num ~ x + y, data=d)
> xt
   y
x    A B
  1 3 7
  2 8 5
```

분할표

- 도수를 나타내는 컬럼이 따로 없고, 각 관찰 결과가 서로 다른 행으로 표현되어 있다면 formula는 '~ 변수 + 변수 ...'

```
> d2 <- data.frame(x=c("A", "A", "A", "B", "B"),
+                  result=c(3, 2, 4, 7, 6))
> xtabs(~ x, d2)
x
A B
3 2
```

상관분석(correlation analysis)

- **두 확률변수간의 관련성 검사**
  - 상관계수(-1과 1사이의 값)가 크면 두 변수간에 관련성이 깊음을 의미함
  - -면 음의 상관관계, +면 양의 상관관계
  - 단, 상관관계는 인과관계가 아님
    - 예)
      - "방글라데시의 버터 생산량 <-> 뉴욕 증시의 주가 변동
      - "이산화탄소 농도 변화" <-> 지구 기온
      - "경찰관수" <-> 범죄빈도
  - 확률변수간 독립이면 상관계수는 0

$$P(X,\ Y) = P(X)P(Y)$$

상관분석(correlation analysis)

- 피어슨(Pearson) 상관계수

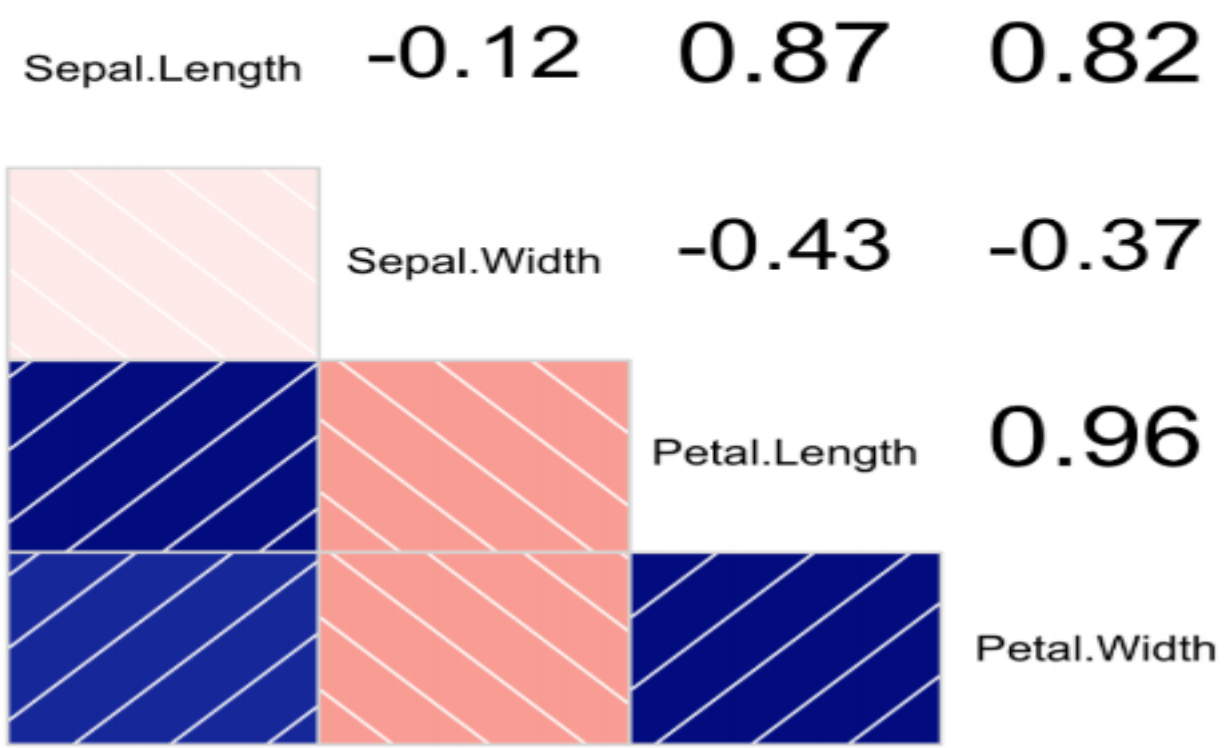
```
> cor(iris$Sepal.Width, iris$Sepal.Length)
[1] -0.1175698
```

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

상관분석(correlation analysis)

```
> install.packages("corrgram")
> library(corrgram)
> corrgram(cor(iris[,1:4]), type="corr", upper.panel=panel.conf)
```



상관계수 검정

- **통계적 유의성 검증**

```
> cor.test(c(1, 2, 3, 4, 5), c(1, 0, 3, 4, 5), method="pearson")

  Pearson's product-moment correlation

data:  c(1, 2, 3, 4, 5) and c(1, 0, 3, 4, 5)
t = 3.9279, df = 3, p-value = 0.02937
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1697938 0.9944622
sample estimates:
      cor
0.9149914
```