제5강 : 확률분포

금융 통계 및 시계열 분석

TRADE INFORMATIX

2014년 1월 21일

Outline

통계에서 확률 모형의 의미와 역할

통계 분석의 가정

Population Assumption
우리가 고려하는 자료 집합은 특정한 특성을 가진 모형으로부터 생성된 샘플 집단이다.

통계 분석의 문제

Inference on Population
우리가 고려하는 샘플 집단을 생성한 모형의 특성을 구한다.

통계 분석을 이용한 예측

Prediction
우리가 고려하는 샘플 집단이 생성된 모집단의 특성이 변하지 않는다면 미래에 생성될 샘플 집단의 특징을 예측할 수 있다.

확률의 정의

확률 실험 (random experiment)

예측할 수는 없지만 가능한 결과 (possible outcomes, ω) 가 정의되어 있는 절차

- 예: 동전을 100번 던진 결과는?
- 확률 실험의 결과는 숫자로 나타낼 수 없는 것을 모두 포함

샘플 공간 (sample space) Ω

확률 실험의 모든 가능한 결과의 집합

- 예: 동전을 100번 던졌을 때 나올 수 있는 모든 경우의 집합

확률의 정의 (계속)

확률 사건 (random event) $\{\omega\}$

샘플 공간의 부분집합

- 예: 동전이 모두 앞면이 나온 결과 하나로 이루어진 집합

시그마 대수 (sigma-algebra) $\mathcal{F}$

확률사건의 집합 즉, 샘플공간의 부분집합의 집합 중 다음 조건을 만족하는 것

1. 공집합 포함 ($\emptyset \in \mathcal{F}$)
2. 모든 원소의 여집합을 원소로 포함 ($A \in \mathcal{F} \rightarrow A^C \in \mathcal{F}$)
3. 모든 원소의 합집합을 원소로 포함 ($A, B \in \mathcal{F} \rightarrow A \cup B \in \mathcal{F}$)

확률의 정의 (계속)

확률 측도 (probability measure) $P(A)$

시그마 대수가 존재할 때 그 시그마 대수의 원소인 확률사건에 대해 실수값을 대응시킨 함수 중 다음 조건을 만족하는 것

1. 확률값은 0과 1사이의 값 $(0 \leq P(A) \leq 1)$
2. 샘플 공간의 확률값은 1 $(P(\Omega) = 1)$
3. 서로 소인 두 확률 사건의 합집합의 확률값은 각 확률사건에 대한 확률값의 합 $(A \cap B = \emptyset \ \rightarrow \ P(A \cup B) = P(A) + P(B))$

확률은 개별 결과값이 아닌 그 결과의 집합인 확률 사건에 대해서만 정의

확률변수

보렐집합 (Borel Set)

모든 닫힌 실수 구간 (closed interval) 포함하는 시그마 대수의 원소

확률 변수 (random variable) X

샘플 공간의 원소 즉 **개별적인 확률 실험 결과**에 대해 **실수값 (real value)**를 정의한 함수 중 다음 조건을 만족하는 함수.

- ❏ 모든 보렐집합에 대해 그 보렐집합에 대응하는 확률사건이 시그마 대수의 원소여야 한다.

확률 사건 즉 확률실험 결과의 집합에 대해 정의된 확률값과 달리 확률 변수는 개별 확률 실험 결과에 대해 정의

확률분포

- 확률분포
  - 확률변수를 수학적으로 정의하기 위한 함수
- 이산 확률 분포
  - 확률변수의 값이 이산값 (discrete value)
  - 이산 확률분포함수 (누적분포함수, 확률밀도함수) 로 표현 가능
- 연속 확률 분포
  - 확률변수의 값이 연속값 (continuouse value)
  - 연속 확률분포함수 (누적분포함수, 확률밀도함수) 로 표현 가능

이산 확률분포의 정의

- 누적확률함수 cumulative probability mass function
  - 확률변수 X에 대해 특정한 x값보다 같거나 작은 값이 나올 수 있는 확률

$$c.m.f(x) = P\{X \leq x\} \tag{1}$$

- 확률함수 probability mass function
  - 확률변수 X에 대해 각각의 x값이 나올 수 있는 확률

$$p.m.f(x) = P\{X = x\} \tag{2}$$

연속 확률분포의 정의

- 누적분포함수 cumulative probability density function
  - 확률변수 X에 대해 특정한 x값보다 같거나 작은 값이 나올 수 있는 확률

$$c.d.f(x) = P\{X \leq x\} \tag{3}$$

- 확률밀도함수 probability density function
  - 누적분포함수를 x로 미분한 함수

$$p.d.f(x) = \frac{\partial c.d.f(x)}{\partial x} \tag{4}$$

모멘트 (moment)

- 1차 모멘트 : 평균 (mean)
  - 확률변수의 기대값
  - 확률변수분포의 중앙 위치

$$\mu = E[x] \tag{5}$$

- 2차 모멘트 : 분산 (variance)
  - 확률변수의 평균으로부터의 오차의 제곱의 기대값
  - 확률변수분포가 양쪽으로 퍼진 정도

$$\sigma^2 = E[(x-\mu)^2] \tag{6}$$

- 3차 모멘트 : 왜도 (skewness)
  - 확률변수의 평균으로부터의 오차를 분산으로 나눈 값의 세제곱의 기대값
  - 확률변수분포가 한쪽으로 쏠린 정도

$$\sigma^2 = E[((x-\mu)/\sigma)^3] \tag{7}$$

- 4차 모멘트 : (초과) 첨도 (excessive kurtosis)
  - 확률변수의 평균으로부터의 오차를 분산으로 나눈 값의 네제곱의 기대값
  - 초과첨도는 여기에서 normal 분포의 첨도인 3을 뺀 값
  - 확률변수분포가 normal 분포에 비해 양 끝단으로 퍼진 정도

$$\sigma^2 = E[((x-\mu)/\sigma)^4] \tag{8}$$

랜덤 넘버 제어

- `RNGkind(kind, normal.kind)` 생성 알고리즘 설정
  - `kind` 표준 랜덤 넘버 생성 알고리즘
    - "Wichmann-Hill", "Marsaglia-Multicarry", "Super-Duper", "Mersenne-Twister", "Knuth-TAOCP-2002", "Knuth-TAOCP", "L'Ecuyer-CMRG"
  - `normal.kind` Normal 랜덤 넘버 생성 알고리즘
- `set.seed(seed, kind, normal.kind)` 시드(seed) 넘버 설정

```
> set.seed(1)
> runif(1)
[1] 0.2655087
> runif(1)
[1] 0.3721239
> set.seed(1)
> runif(1)
[1] 0.2655087
> runif(1)
[1] 0.3721239
```

샘플링

- `sample(data, size, replace)` : 지정된 모집단에서 원하는 갯수의 데이터 샘플 채취
  - `data` 모집단
  - `size` 샘플 갯수
  - `replace` : `TRUE`면 샘플 채취 확률이 언제나 같음. 같은 샘플이 여러번 채취 가능
    `FALSE`면 채취된 샘플은 모집단에서 없어짐. 같은 샘플은 채취 불가

```
> set.seed(1)
> x <- 1:10
> sample(x, size=8, replace=TRUE)
[1]  3  4  6 10  3  9 10  7
> sample(x, size=8, replace=FALSE)
[1]  7  1  2  8  5 10  4  6
```

R에서 제공하는 확률분포 관련 기능

기능	prefix	사용예
확률밀도함수 density/mass function	`d`	`dnorm`
누적분포함수 cumulative distribution function	`p`	`pnorm`
분위수계산함수 quantile function, inverse of cdf	`q`	`qnorm`
랜덤샘플생성 sample realization	`r`	`rnorm`

R에서 제공하는 확률분포 목록

분포 종류	분포 이름	R 명칭	인수
이산 분포	binomial	`binom`	`size, prob`
	negative-binomial	`nbinom`	`size, prob, mu`
	geometric	`geom`	`prob`
	hypergeometric	`hyper`	`m, n, k, p`
	Poisson	`pois`	`lambda`
연속 분포	uniform	`unif`	`min, max`
	normal	`norm`	`mean, sd`
	log-normal	`lnorm`	`meanlog, sdlog`
	exponential	`exp`	`rate`
	Gamma	`gamma`	`shape, scale`
	Weibull	`weibull`	`shape, scale`
연속 분포 (test)	student-t	`t`	`df`
	F	`f`	`df1, df2`
	Chi-Squared	`chisq`	`df`
	Wilcoxon	`wilcox`	`m, n`
	Cauchy	`cauchy`	`location, scale`
연속 분포 (Bayesian)	Beta	`beta`	`shape1, shape2`
	Dirichlet	`dirichlet`	bayesm 패키지

Bernoilli/Binomial 분포

확률밀도함수	`dbinom`	평균	p
누적분포함수	`pbinom`	분산	pq
분위수계산함수	`qbinom`	왜도	$(q-p)/\sqrt{pq}$
랜덤샘플생성	`ebinom`	첨도	$(1-6pq)/\sqrt{pq}$

- "베르누이 시도 (Bernoilli trial)"은 두가지 결과값만 있는 사건을 말한다. 예를 들어 성공/실패 혹은 동전의 앞면/뒷면 혹은 1/0 값 등이다. 베르누이 시도는 1 값 (성공 혹은 앞면) 이 나올 성공확률 p로 정의된다. 실패확률은 1-p가 된다.

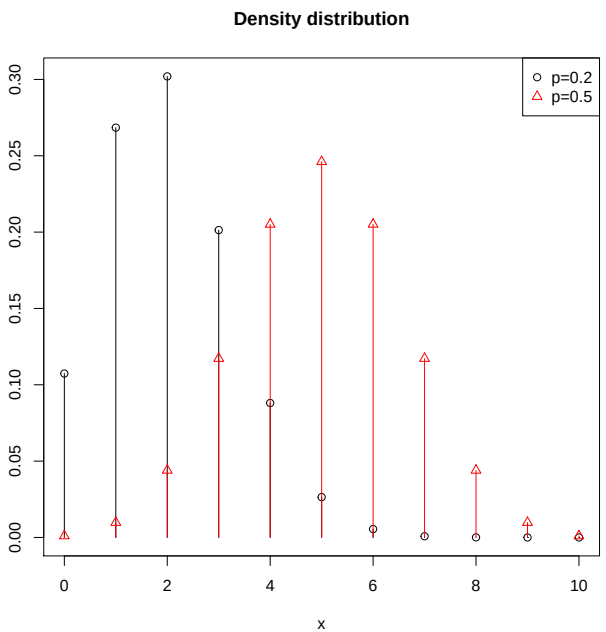
$$p(x) = \begin{cases} p, & \text{if } x = 1. \\ 1-p, & \text{if } x = 0. \end{cases} \tag{9}$$

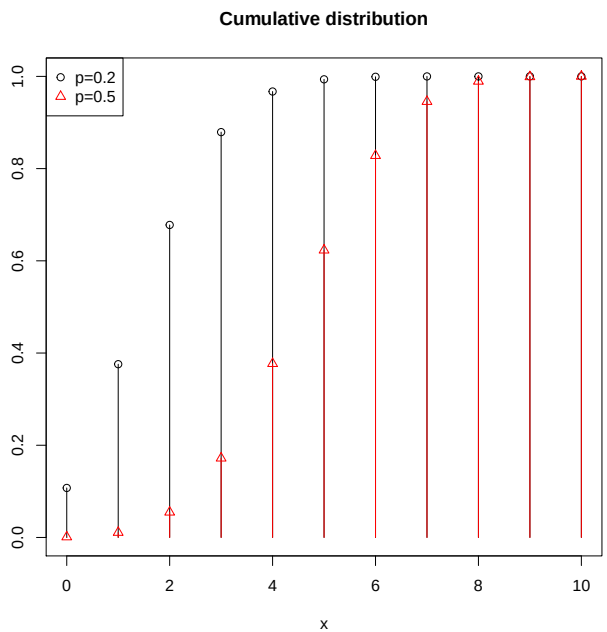
- "Binomial 분포"는 동일한 성공확률 p를 가지는 베르누이 시도가 n번 연속되었을 때 전체 성공횟수에 대한 분포이다. n번 짜리 Binomial 분포의 값은 당연히 0부터 n까지의 정수이다. 베르누이 분포는 n=1인 Binomial 분포이다.
- 만약 일간 주가 이벤트가 상승/하락의 두 가지로만 정의되고 상승확률이 p, 하락확률이 1-p이며 매일의 주가이벤트가 같은 확률분포이고 독립적이면 (iid) n일간의 상승일수는 n-Binomial 분포가 된다.

Binomial 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,10,1)
> d1 <- dbinom(x, 10, prob=0.2)
> d2 <- dbinom(x, 10, prob=0.5)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("p=0.2", "p=0.5"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,10,1)
> p1 <- pbinom(x, 10, prob=0.2)
> p2 <- pbinom(x, 10, prob=0.5)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("p=0.2", "p=0.5"),
+        pch=1:2, col=1:2)
```





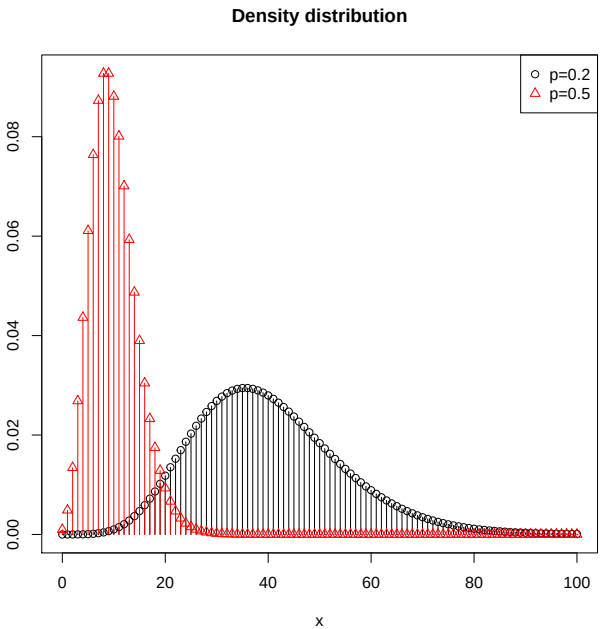
Negative Binomial 분포

확률밀도함수	`dnbinom`	평균	$pr/(1-p)$
누적분포함수	`pnbinom`	분산	$pr/(1-p)^2$
분위수계산함수	`qnbinom`	왜도	$(1+p)/\sqrt{pr}$
랜덤샘플생성	`enbinom`	첨도	$6/r + (1-p)^2/\sqrt{pr}$

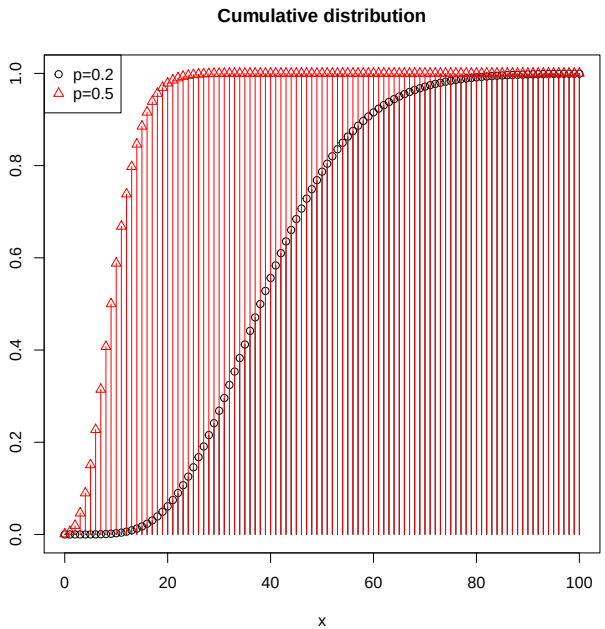
- 동일한 성공확률 p를 가지는 베르누이 시도의 결과값이 r이 되기 위해 필요한 전체 베르누이 시도의 횟수에 대한 분포 Binomial 분포와 달리 0부터 무한대의 값을 가진다.

Negative Binomial 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,100,1)
> d1 <- dnbinom(x, 10, prob=0.2)
> d2 <- dnbinom(x, 10, prob=0.5)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("p=0.2", "p=0.5"),
+        pch=1:2, col=1:2)
```




```
> x <- seq(0,100,1)
> p1 <- pnbinom(x, 10, prob=0.2)
> p2 <- pnbinom(x, 10, prob=0.5)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("p=0.2", "p=0.5"),
+        pch=1:2, col=1:2)
```



Geometric 분포

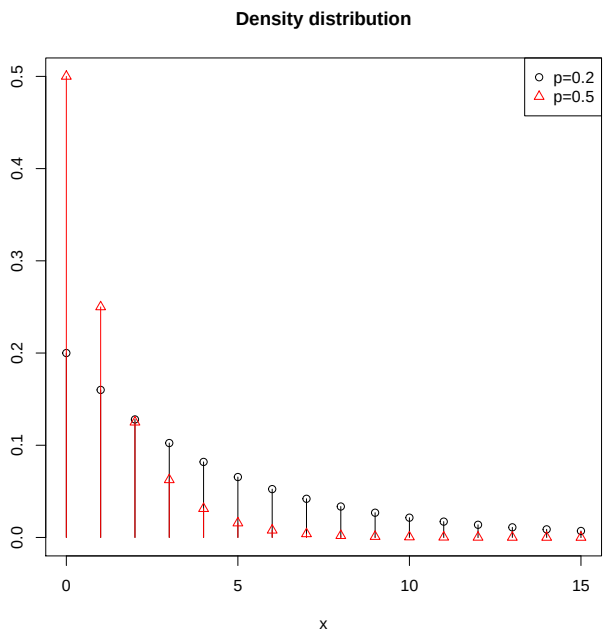
확률밀도함수	`dgeom`	평균	$1-p$
누적분포함수	`pgeom`	분산	$(1-p)/p^2$
분위수계산함수	`qgeom`	왜도	$(2-p)/\sqrt{1-p}$
랜덤샘플생성	`rgeom`	첨도	$6+p^2/\sqrt{1-p}$

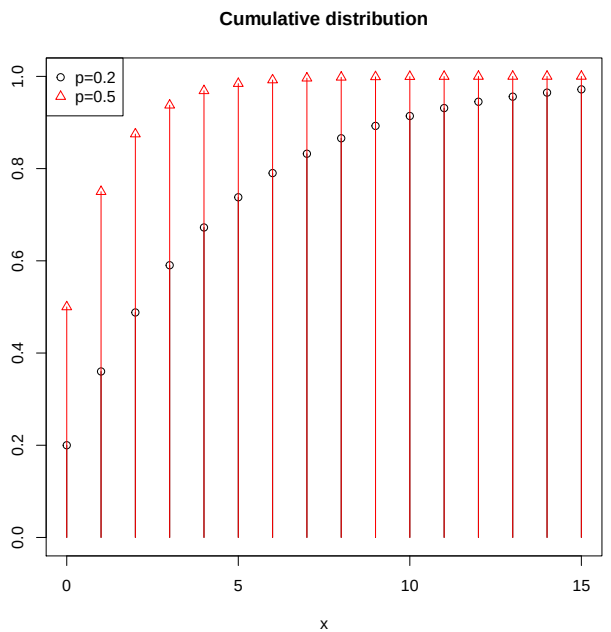
- 동일한 성공확률 p를 가지는 베르누이 시도의 결과값이 최초로 1(성공)이 되는데 필요한 전체 베르누이 시도의 횟수에 대한 분포

Geometric 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,15,1)
> d1 <- dgeom(x, prob=0.2)
> d2 <- dgeom(x, prob=0.5)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("p=0.2", "p=0.5"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,15,1)
> p1 <- pgeom(x, prob=0.2)
> p2 <- pgeom(x, prob=0.5)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("p=0.2", "p=0.5"),
+        pch=1:2, col=1:2)
```





Hypergeometric 분포

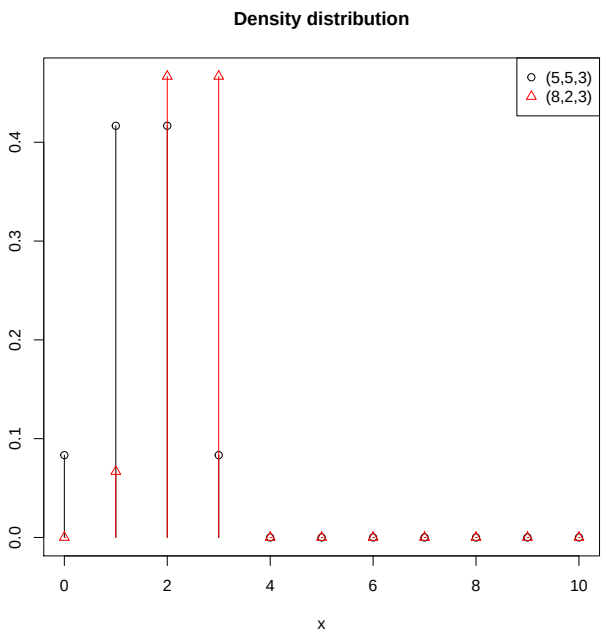
확률밀도함수	`dhyper`	평균	$km/(m+n)$
누적분포함수	`phyper`	분산	$kmn(mn-k)/((m+n)^2(mn-1))$
분위수계산함수	`qhyper`	왜도	
랜덤샘플생성	`rhyper`	첨도	

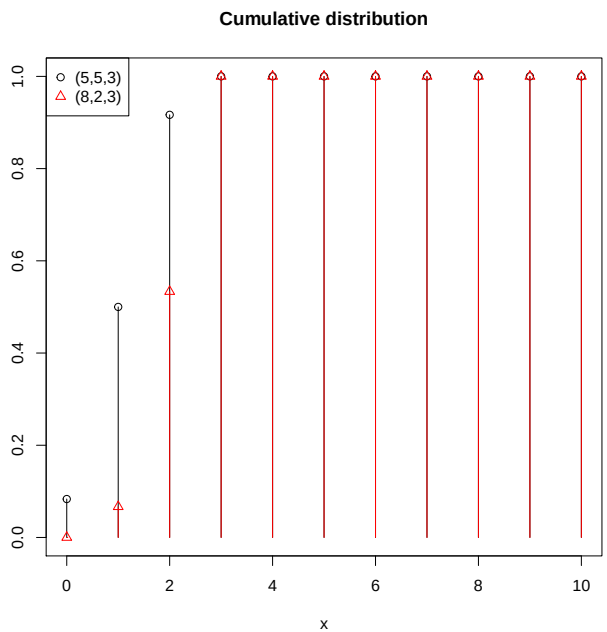
- 베르누이 시도와 같이 결과값이 0 혹은 1의 두가지 경우만을 가지는 시도의 성공횟수. 그러나 binomial 분포와 달리 각 시도의 확률이 같지 않으며 전체 샘플의 수가 N이고 성공 샘플의 수가 m, 실패 샘플의 수가 n(=N-m)인 모집합에서 k개의 샘플을 구하는 경우처럼 대체(replacement)가 없다.

Hypergeometric 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,10,1)
> d1 <- dhyper(x, 5, 5, 3)
> d2 <- dhyper(x, 8, 2, 3)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("(5,5,3)", "(8,2,3)"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,10,1)
> p1 <- phyper(x, 5, 5, 3)
> p2 <- phyper(x, 8, 2, 3)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("(5,5,3)", "(8,2,3)"),
+        pch=1:2, col=1:2)
```





Possison 분포

확률밀도함수	`dpois`	평균	λ
누적분포함수	`ppois`	분산	λ
분위수계산함수	`qpois`	왜도	$\lambda^{-1/2}$
랜덤샘플생성	`rpois`	첨도	λ^{-1}

- 0부터 t시간까지 특정 이벤트가 발생한 횟수
- 이벤트 사이의 대기시간은 파라미터가 λ인 지수(exponentional) 분포

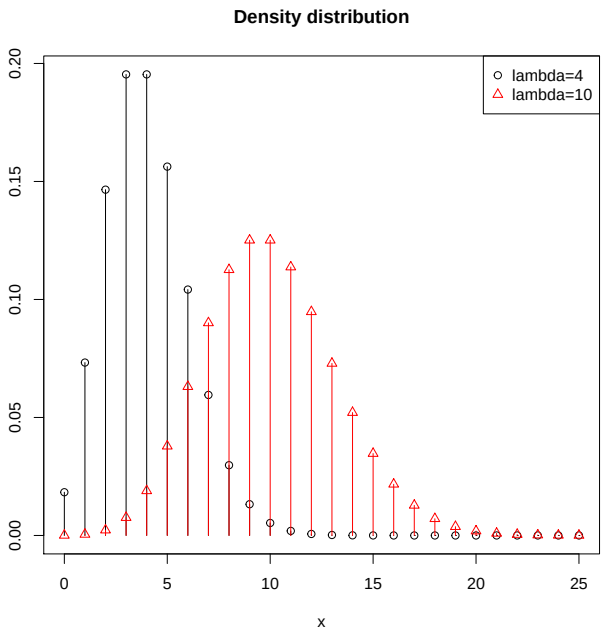
$$p(x = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{10}$$

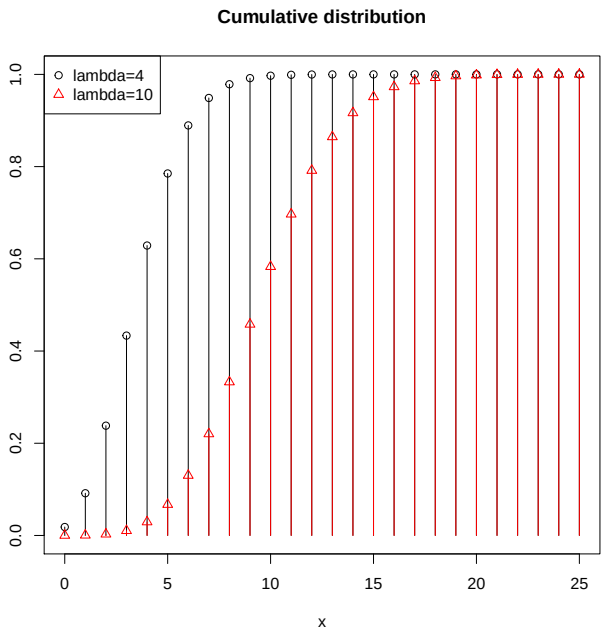
- 신용모형의 파산확률 분석
- high-frequency tick data 분석에서 duration 분석
- 시스템 트레이딩에서 특정 시그널의 발생 빈도 분석

Possison 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,25,1)
> d1 <- dpois(x, lambda=4)
> d2 <- dpois(x, lambda=10)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("lambda=4", "lambda=10"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,25,1)
> p1 <- ppois(x, lambda=4)
> p2 <- ppois(x, lambda=10)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("lambda=4", "lambda=10"),
+        pch=1:2, col=1:2)
```





지수 분포

확률밀도함수	`dexp`	평균	$1/\lambda$
누적분포함수	`pexp`	분산	$1/\lambda^2$
분위수계산함수	`qexp`	왜도	2
랜덤샘플생성	`rexp`	첨도	6

❏ 이벤트 사이의 대기시간에 대한 일반적 분포

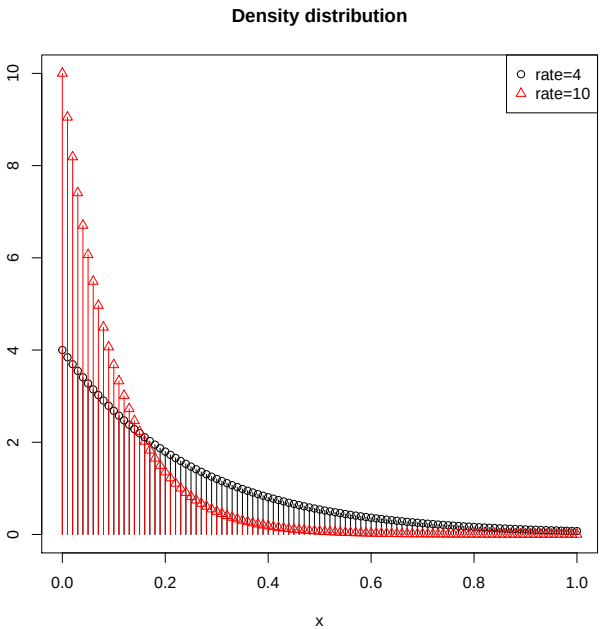
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0. \\ 0, & \text{if } x < 0. \end{cases} \tag{11}$$

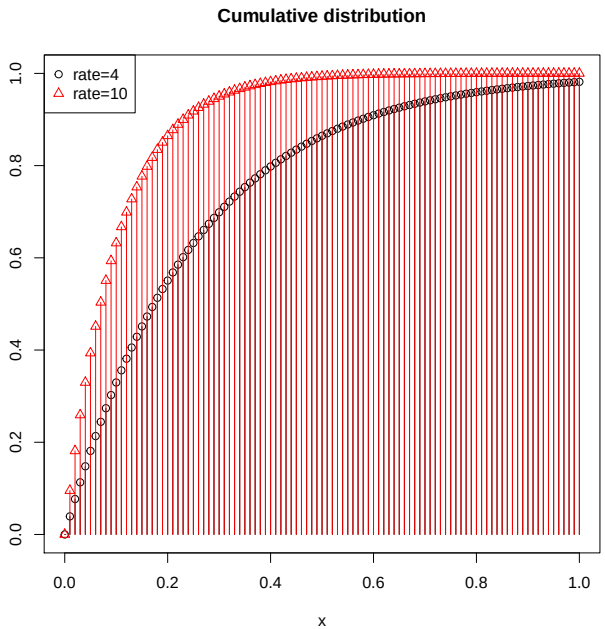
$$F(x) = \begin{cases} 1 - \lambda e^{-\lambda x}, & \text{if } x \geq 0. \\ 0, & \text{if } x < 0. \end{cases} \tag{12}$$

지수 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,1,0.01)
> d1 <- dexp(x, rate=4)
> d2 <- dexp(x, rate=10)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("rate=4", "rate=10"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,1,0.01)
> p1 <- pexp(x, rate=4)
> p2 <- pexp(x, rate=10)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("rate=4", "rate=10"),
+        pch=1:2, col=1:2)
```





Normal 분포

확률밀도함수	`dnorm`	평균	μ
누적분포함수	`pnorm`	분산	σ^2
분위수계산함수	`qnorm`	왜도	0
랜덤샘플생성	`rnorm`	첨도	0

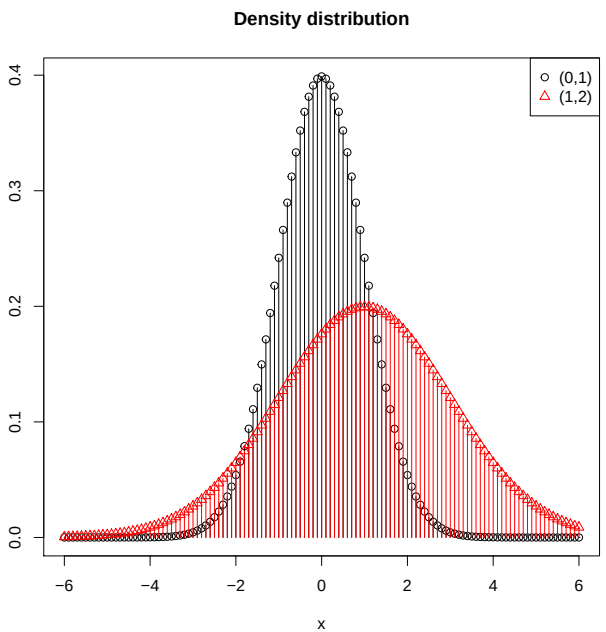
❏ 평균μ, 분산σ의 두 파라미터로 정의되는 가장 일반적 확률분포

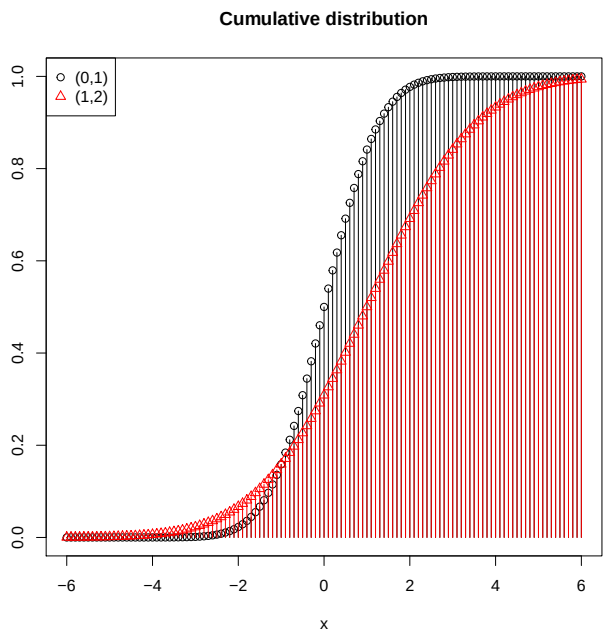
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{13}$$

Normal 분포의 확률밀도함수/누적분포함수

```
> x <- seq(-6,6,0.1)
> d1 <- dnorm(x, 0, 1)
> d2 <- dnorm(x, 1, 2)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("(0,1)", "(1,2)"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(-6,6,0.1)
> p1 <- pnorm(x, 0, 1)
> p2 <- pnorm(x, 1, 2)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("(0,1)", "(1,2)"),
+        pch=1:2, col=1:2)
```





Gamma 분포

확률밀도함수	`dexp`	평균	$1/\lambda$
누적분포함수	`pexp`	분산	$1/\lambda^2$
분위수계산함수	`qexp`	왜도	2
랜덤샘플생성	`rexp`	첨도	6

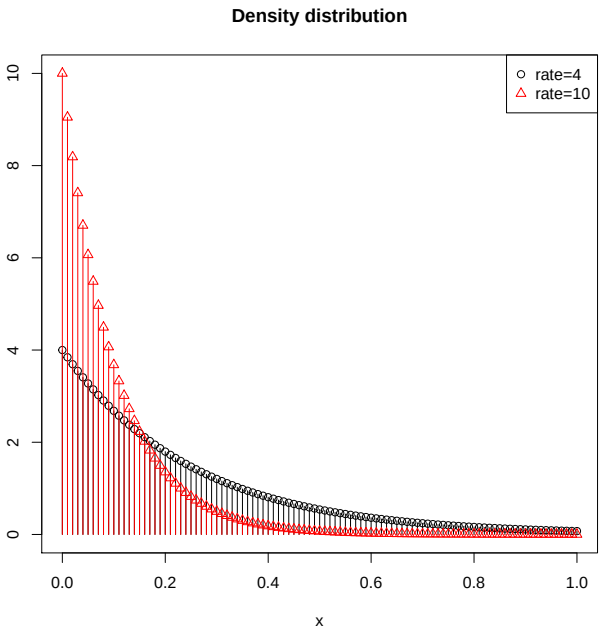
❏ 지수분포의 일반화 버전

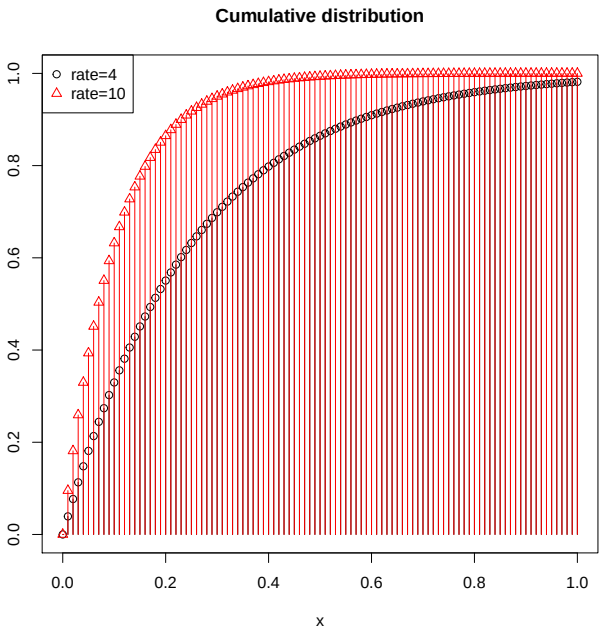
$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \quad \text{for } x > 0 \text{ and } k, \theta > 0. \tag{14}$$

Gamma 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,1,0.01)
> d1 <- dexp(x, rate=4)
> d2 <- dexp(x, rate=10)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("rate=4", "rate=10"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,1,0.01)
> p1 <- pexp(x, rate=4)
> p2 <- pexp(x, rate=10)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("rate=4", "rate=10"),
+        pch=1:2, col=1:2)
```





Weibull 분포

확률밀도함수	`dweibull`	평균	$\lambda\Gamma(1+1/k)$
누적분포함수	`pweibull`	분산	$\lambda^2\Gamma(1+2/k)-\mu^2$
분위수계산함수	`qweibull`	왜도	
랜덤샘플생성	`rweibull`	첨도	

- k, λ의 두 파라미터로 정의되는 연속 확률분포

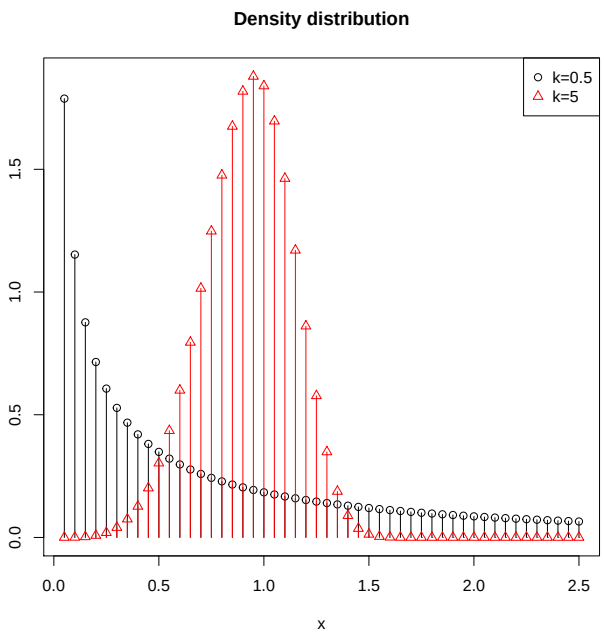
$$f(x;\lambda,k)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k} & x\geq 0,\\ 0 & x<0,\end{cases} \tag{15}$$

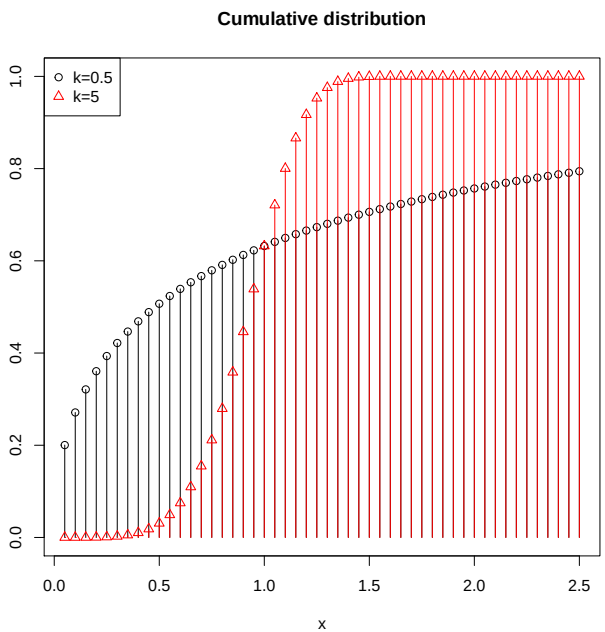
- 지수함수, 정상 함수 등 다양한 연속분포 근사화 가능
- duration 분석, 파산 분석, 수명 분석 등에 사용

Weibull 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0.05,2.5,0.05)
> d1 <- dweibull(x, 0.5)
> d2 <- dweibull(x, 5)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("k=0.5", "k=5"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0.05,2.5,0.05)
> p1 <- pweibull(x, 0.5)
> p2 <- pweibull(x, 5)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("k=0.5", "k=5"),
+        pch=1:2, col=1:2)
```





Chi-Squared 분포

확률밀도함수	`dchisq`	평균	k
누적분포함수	`pchisq`	분산	$2k$
분위수계산함수	`qchisq`	왜도	$\sqrt{8/k}$
랜덤샘플생성	`rchisq`	첨도	$12/k$

- normal 분포를 따르는 k 개의 확률변수의 제곱의 합

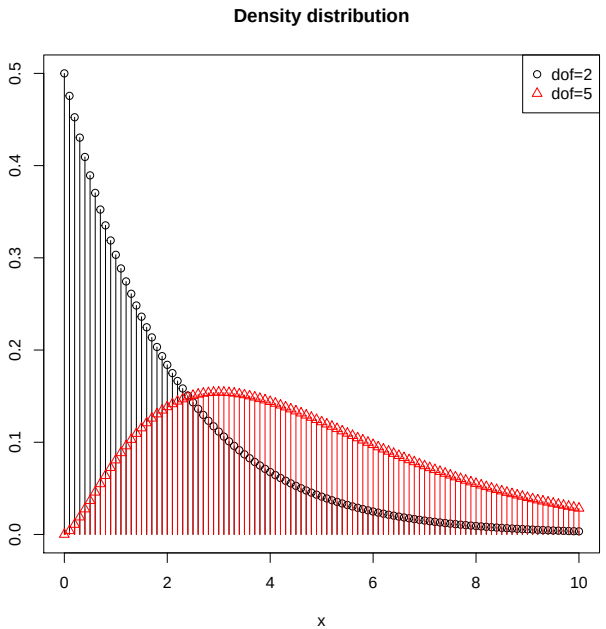
$$f(x;k) = \sum_{i=1}^{k} x_i^2 \tag{16}$$

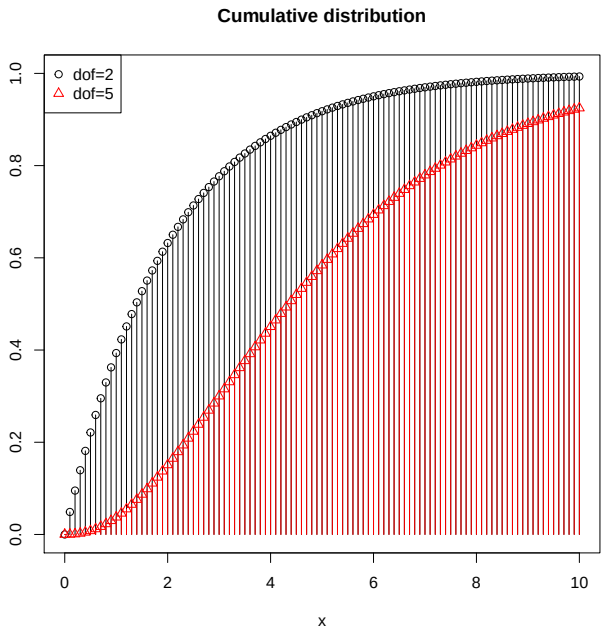
- normal 분포의 분산 추정에 사용

Chi-Squared 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0,10,0.1)
> d1 <- dchisq(x, 2)
> d2 <- dchisq(x, 5)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("dof=2", "dof=5"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0,10,0.1)
> p1 <- pchisq(x, 2)
> p2 <- pchisq(x, 5)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("dof=2", "dof=5"),
+        pch=1:2, col=1:2)
```





Student-t 분포

확률밀도함수	`dt`	평균	0
누적분포함수	`pt`	분산	$k/(k-2)$
분위수계산함수	`qt`	왜도	0
랜덤샘플생성	`rt`	첨도	$6/(k-4)$

- normal분포에서 뽑은 n 개의 샘플에 대해 다음 수식의 결과가 가지는 분포. 여기서 $\bar{x}$ 와 s 는 각각 샘플평균과 샘플표준편차

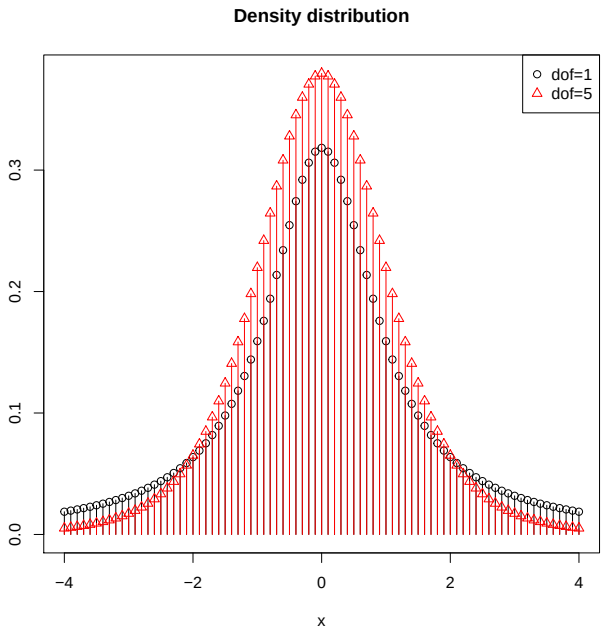
$$t = \frac{\mu - \bar{x}}{s/\sqrt{n}} \tag{17}$$

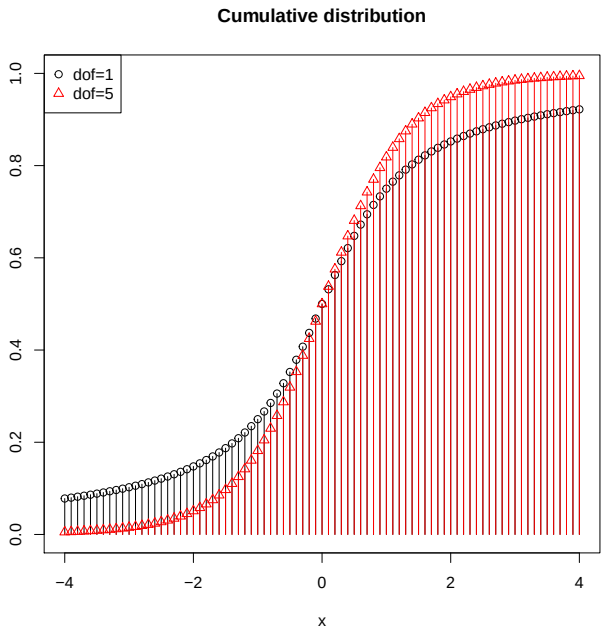
- 정규분포를 chi-quared로 나눈 형태
- 정규분포의 표준 추정에 사용

Student-t 분포의 확률밀도함수/누적분포함수

```
> x <- seq(-4,4,0.1)
> d1 <- dt(x, 1)
> d2 <- dt(x, 5)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("dof=1", "dof=5"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(-4,4,0.1)
> p1 <- pt(x, 1)
> p2 <- pt(x, 5)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("dof=1", "dof=5"),
+        pch=1:2, col=1:2)
```





F 분포

확률밀도함수	`df`	평균	$d_2/(d_2-2)$
누적분포함수	`pf`	분산	$\frac{2\,d_2^2\,(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$
분위수계산함수	`qf`	왜도	
랜덤샘플생성	`rf`	첨도	

- chi-squred 분포를 두 샘플의 경우로 확장한 분포
- 두 normal 분포 샘플의 샘플표준편차의 비율이 가지는 분포

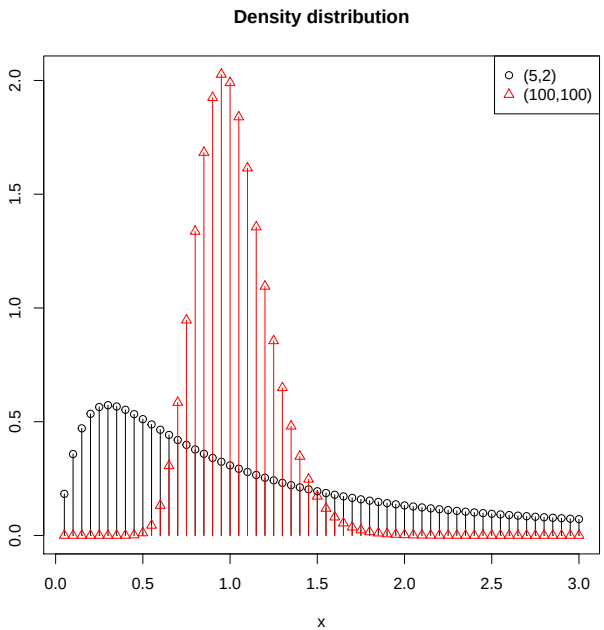
$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{18}$$

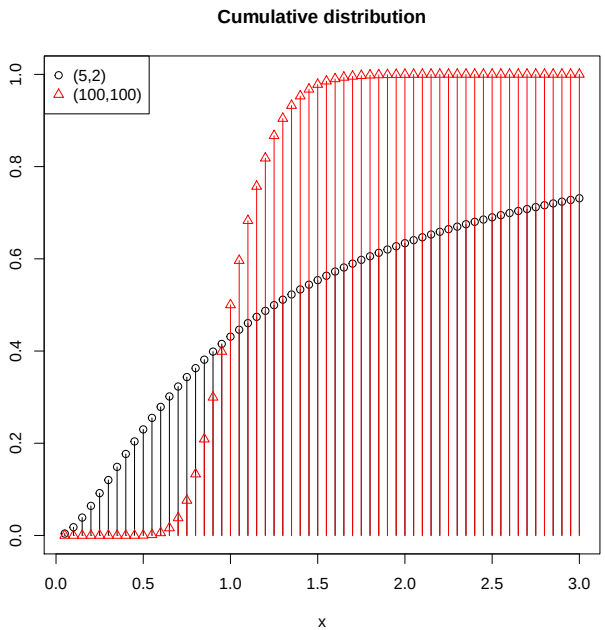
- 두 샘플의 분산 비교에 사용

F 분포의 확률밀도함수/누적분포함수

```
> x <- seq(0.05,3,0.05)
> d1 <- df(x, 5, 2)
> d2 <- df(x, 100, 100)
> plot(x, d1, type='p', pch=1, col=1,
+      ylim=c(0, max(c(d1,d2))), ylab="",
+      main="Density distribution")
> lines(x, d1, type='h', col=1)
> points(x, d2, pch=2, col=2)
> lines(x, d2, type='h', col=2)
> legend("topright", c("(5,2)", "(100,100)"),
+        pch=1:2, col=1:2)
```

```
> x <- seq(0.05,3,0.05)
> p1 <- pf(x, 5, 2)
> p2 <- pf(x, 100, 100)
> plot(x, p1, type='p', pch=1, col=1,
+      ylim = c(0, 1), ylab="",
+      main="Cumulative distribution")
> lines(x, p1, type='h', col=1)
> points(x, p2, pch=2, col=2)
> lines(x, p2, type='h', col=2)
> legend("topleft", c("(5,2)", "(100,100)"),
+        pch=1:2, col=1:2)
```





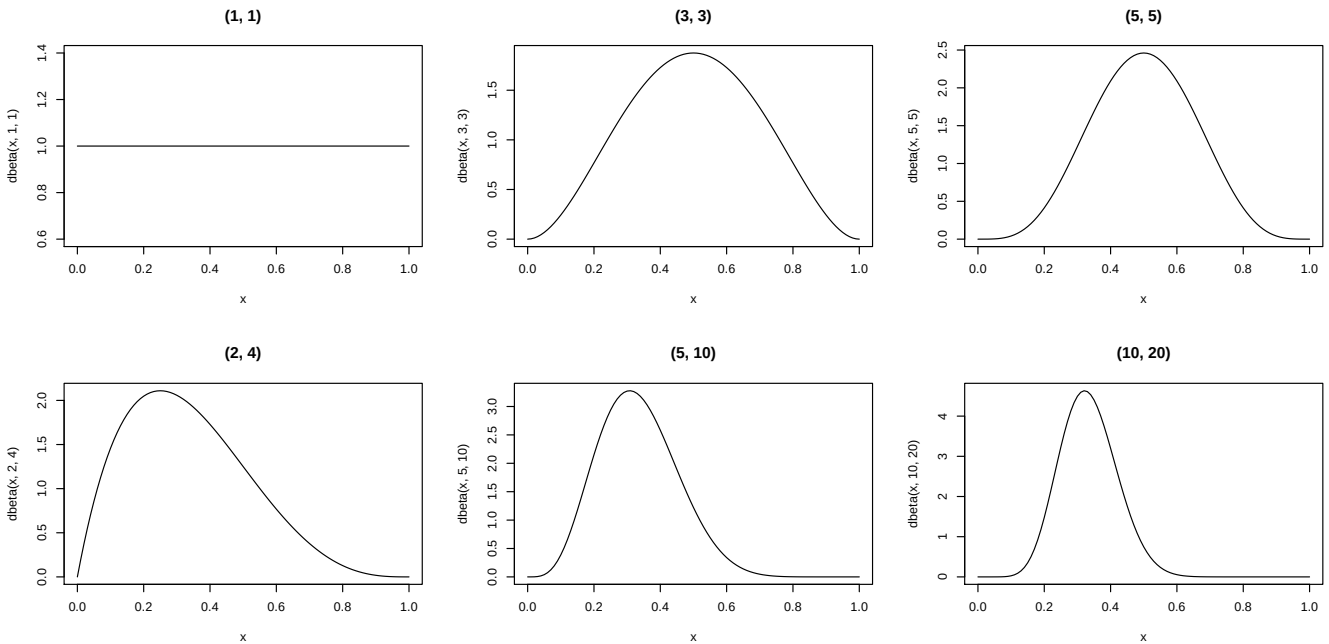
Beta 분포

- uniform 분포부터 시작하여 다양한 uni-modal 형태를 가지므로 파라미터 값의 Inference에 대한 분포 지정에 편리

$$
\begin{aligned}
f(x;\alpha,\beta) &= \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\text{Beta}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$

Beta 분포

```
> x <- seq(0, 1, by=0.01)
> par(mfrow=c(2,3))
> plot(x, dbeta(x, 1, 1), type='l', main="(1, 1)")
> plot(x, dbeta(x, 3, 3), type='l', main="(3, 3)")
> plot(x, dbeta(x, 5, 5), type='l', main="(5, 5)")
> plot(x, dbeta(x, 2, 4), type='l', main="(2, 4)")
> plot(x, dbeta(x, 5, 10), type='l', main="(5, 10)")
> plot(x, dbeta(x, 10, 20), type='l', main="(10, 20)")
> par(mfrow=c(1,1))
```



분포 추정

- 분포 추정의 단계
  1. 분포 결정
     - 관심을 가진 확률변수가 어떤 형태를 가지는가
     - histogram, 커널 밀도 (kernel density)
     - moment 비교
  2. 분포 테스트
     - 관심을 가진 확률변수가 특정한 분포를 따르는가?
     - QQ Plot
  3. 파라미터 추정
     - 확률변수의 분포의 파라미터 (평균, 분산) 값은?
  4. 파라미터 테스트
     - 파라미터 추정치의 신뢰성/정확도는?

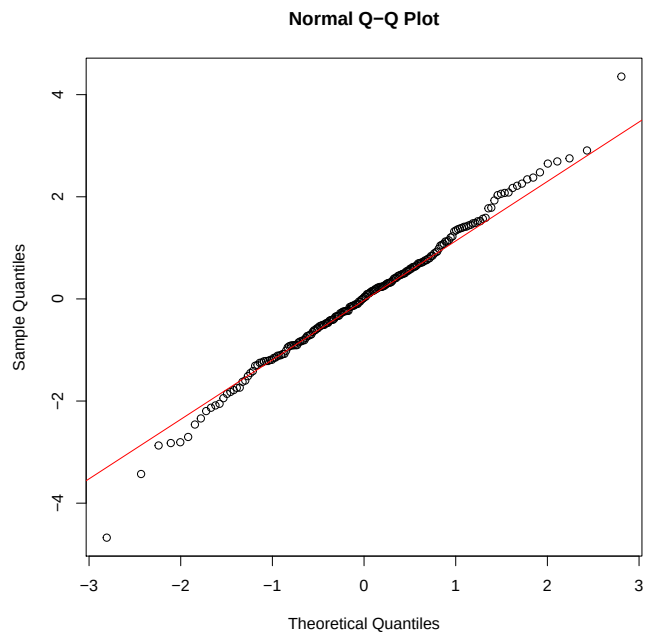
분포 결정 방법 및 관련 테스트

- `hist` : 히스토그램 작성
- `density` : 커널 밀도 (kernel density) 작성
- `qqplot` : QQ Plot 작성
- `ks.test` : Kolmogorov-Smirnov test
- `shapiro.test` : Shapiro-Wilk test

QQ Plot

- 0부터 1까지의 확률값에 대해 normal 분포와 샘플의 quantile 값을 각각 계산하여 이를 x,y 좌표로 점을 찍는다.
- 만일 샘플이 normal 분포라면 직선의 형태
- 샘플이 normal 분포보다 long tail이라면 같은 확률값에 대한 quantile값이 더 커지거나(확률1근처) 작아진다(확률0근처)
- `qqnorm(y)` : normal 분포와 샘플 x의 qq plot
- `qqline(y)` : 샘플 y가 normal 분포일때의 이론적인 qq plot line
- `qqplot(x, y)` : 샘플 x와 샘플 y의 qq plot

```
> y <- rt(200, df = 5)
> qqnorm(y)
> qqline(y, col = 2)
```



Kolmogorov-Smirnov test

- 샘플의 모집단이 특정한 알려진 분포와 일치하는지 비교
- `ks.test(x, y`
  - `x` : 샘플
  - `y` : 비교하고자 하는 분포의 샘플 혹은 그 분포에 대한 R cdf 명령어 문자열

```
> x <- rnorm(50)
> x2 <- rnorm(50)
> y <- runif(30)
> ks.test(x, x2)

Two-sample Kolmogorov-Smirnov test

data:  x and x2
D = 0.14, p-value = 0.7166
alternative hypothesis: two-sided

> ks.test(x, y)

Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.58, p-value = 2.381e-06
alternative hypothesis: two-sided
```

Shapiro-Wilk test

- 샘플의 모집단이 정상 분포인지 테스트
- `shapiro.test(x)`

```
> shapiro.test(rnorm(100))

Shapiro-Wilk normality test

data:  rnorm(100)
W = 0.991, p-value = 0.7443

> shapiro.test(rpois(100, lambda=1))

Shapiro-Wilk normality test

data:  rpois(100, lambda = 1)
W = 0.7903, p-value = 1.263e-10
```