일양분포

- 일양분포의 사용법
- R-언어의 분폿값, 확률계산, 난수발생 함수 이름의 규칙
- 일양분포의 분폿값, 확률, 난수의 계산

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

일양분포의 사용법

일양분포(uniform distribution)

- 임의의 구간 (min, max) 사이에서 동일한 확률분포
- 확률변수 X가 일양분포를 따르면 $X \sim U$(min, max)로 표현함

일양분포에서 난수, 분위수 등을 계산하는 R 함수

```
dunif(x, min = 0, max = 1, ...)
punif(q, min = 0, max = 1, lower.tail = TRUE, ...)
qunif(p, min = 0, max = 1, lower.tail = TRUE, ...)
runif(n, min = 0, max = 1,...)
```

- 매개변수
  - x : x에서 확률밀도함수를 구함
  - q : q까지의 누적확률을 구함
  - p : 제 p번째 분위수를 구함
  - n : n개의 난수를 발생함
  - min, max : 일양분포의 구간(기본값: 0과 1로 0과 1사이에 균일한 분포)
  - lower.tail : 누적확률(punif)와 분위수(qunif)를 계산할 때 작은 쪽이 아닌 반대쪽을 구함

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

R-언어의 분폿값, 확률계산, 난수발생 함수 이름의 규칙

분포이름이 name인 경우

- rname
  - 분포의 이름이 name인 분포에서 난수 발생하는 함수로 r이 접두어로 사용
- dname
  - 분포의 이름이 name인 분포에서 확률밀도함수값을 계산하는 함수로 d를 접두어로 사용
- pname
  - 분포의 이름이 name인 분포에서 누적확률함수 $F(x) = Pr[X \leq x]$를 계산
- qname
  - 분포의 이름이 name인 분포에서 분위수를 계산

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

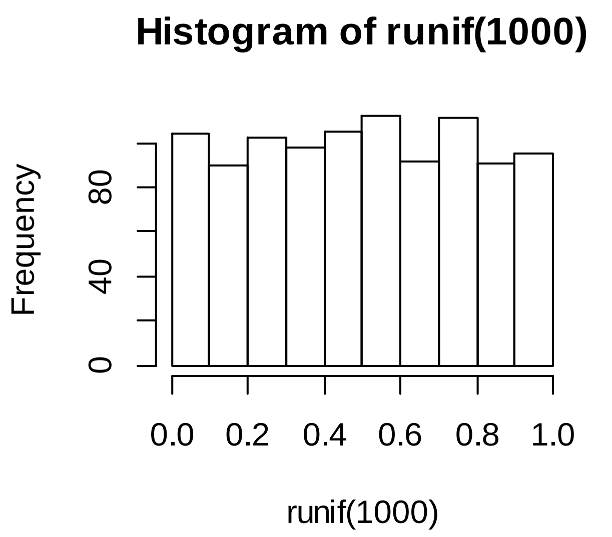
일양분포의 분폿값, 확률, 난수의 계산

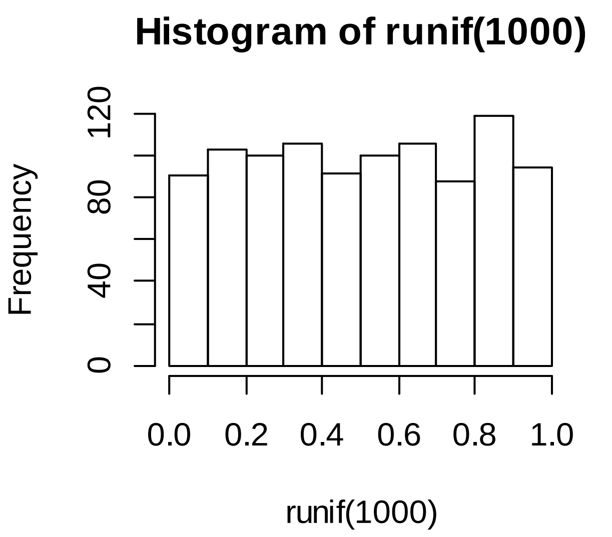
예시 1

난수발생과 분포의 모양

0과 1사이에서 일양분포에서 1000개의 난수를 발생하여 히스토그램으로 난수의 분포를 확인해보자.

```
> hist(runif(1000))
> hist(runif(1000))
```





- 매번 난수를 발생할 때마다 모양은 조금씩 달라지나 각 구간에 포함될 확률은 모두 같음

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

일양분포의 분푯값, 확률, 난수의 계산

예시 2

분위수, 분포함수, 확률밀도 함수

확률변수 X가 일양분포 $U(0,1)$에서 x가 0.25, 0.5 및 0.75일 때 확률밀도함수와 사분위수를 확인해보자.

dunif 함수를 사용하여 확률밀도감수 구하기

```
> p <- c(.25, .5, .75)
> dunif(p)
[1] 1 1 1
```

- 일양분포 $U(0,1)$의 경우 모든 x에 대해 확률밀도함수는 1임

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

일양분포의 분폿값, 확률, 난수의 계산

예시 2

qunif 함수를 사용하여 사분위수 구하기

```
> qunif(p)
[1] 0.25 0.50 0.75
```

- (0,1)에서 균일한 분포이므로 0.25보다 작을 확률은 25%, 0.5보다 작을 확률은 50%, 0.75보다 작을 확률은 75%임

punif 함수를 사용하여 사분위수 구하기

```
> punif(0.8)   # 0.8보다 같거나 작을 확률
[1] 0.8
```


본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포

- 정규분포의 사용법
- 분포, 확률, 난수의 계산

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포의 확률, 난수, 분위수 계산

정규분포

- 기댓값(평균)을 기준으로 좌우대칭이며 전체 분포의 모양은 종모양인 분포

기댓값 μ, 분산 σ^2인 정규분포에서 사용하는 함수

- 난수 : rnorm
- 확률밀도함수 : dnorm
- 누적분포함수 : pnorm
- 분위수 : qnorm

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포의 확률, 난수, 분위수 계산

사용함수

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

* mean에는 기댓값 μ 지정
* sd에는 표준편차 σ 지정

- 기댓값 μ와 표준편차 σ는 생략될 수 있으며 생략된 경우 mean과 sd는 각각 0과 1인 표준정규분포를 사용함

매개변수

- x, q : 확률밀도함수와 누적분포함수의 값을 얻고자 하는 벡터를 설정함
- p : 분위수를 얻고자 하는 확률값의 벡터를 설정함
- n : 생성할 난수의 개수를 지정함
- lower.tail 논리값을 가지며 TRUE이면(기본값), 확률이 $\Pr[X \le x]$로 계산되며 FALSE이면 $\Pr[X > x]$로 계산됨

통계청 통계교육원
본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포의 확률, 난수, 분위수 계산

예시 1

분포함수 값의 계산

어떤 과목의 점수는 평균이 70점, 표준편차가 5점인 정규분포이다.
이 과목의 점수가 80점이면 상위 몇 %에 해당하는지 확인해보자.

```
> 1-pnorm(80, 70, 5)    # 80 이하일 확률
[1] 0.02275013
```

* 상위 약 2.3%에 해당되는 점수임

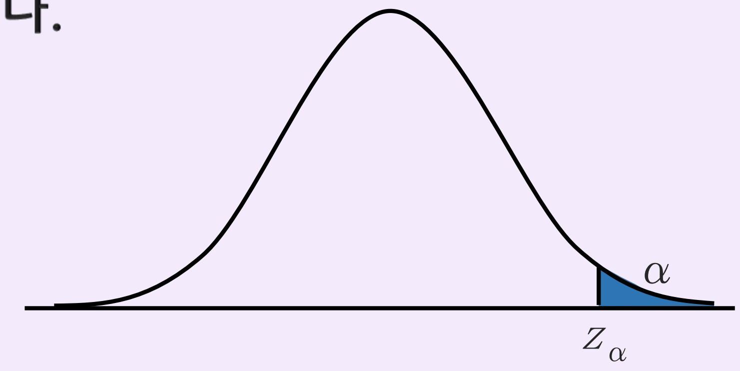
본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포의 확률, 난수, 분위수 계산

예시 2

자료

통계학에서 z_α를 표준정규분포의 제 100(1 - α)%백분위수로 표시한다.
예를 들어 $z_{0.05}$는 95% 백분위수이다.



z_α의 위치 및 해당 영역

표준정규분포에서 2.5% 백분위수 $z_{0.975}$

```
> qnorm(0.025)
 [1] -1.959964
```

* 잘 알려진 값 -1.96임을 알 수 있음(반올림 적용)

통계청 통계교육원

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포의 확률, 난수, 분위수 계산

예시 2

dnorm 함수를 사용하여 확률밀도함수 구하기

```
> dnorm(c(-1, 0, 1))
 [1] 0.2419707 0.3989423 0.2419707
```

- 표준정규분포의 확률밀도함수 $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ 의 값을 각각 -1, 0, 1에서의 값을 계산한 것임

pnorm 함수를 사용하여 누적분포함수 구하기

```
> pnorm(c(-2.54,-1.96,0, 1.96, 2.54))
[1] 0.005542623 0.024997895 0.500000000 0.975002105 0.994457377
```

- Z가 표준정규분포일 때 $\Pr[Z \leq -2.54]$, $\Pr[Z \leq -1.96]$, $\Pr[Z \leq 0]$, $\Pr[Z \leq 1.96]$, $\Pr[Z \leq 2.54]$의 확률을 계산한 것임

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포의 확률, 난수, 분위수 계산

예시 3

자료

20대 한국남성의 키의 분포는 평균이 174, 표준편차가 2.5인 정규분포라고 가정하자.
키 기준으로 상위 5%이내에 들어가려면 키가 최소 얼마라야 하는지 확인해보자.

평균과 표준편차가 각각 174, 2.5인 분포에서 95% 백분위 찾기

```
> qnorm(0.95, 174, 2.5)   # 키 기준으로 상위 5%
[1] 178.1121
```

* 178.1121보다 크면 됨

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

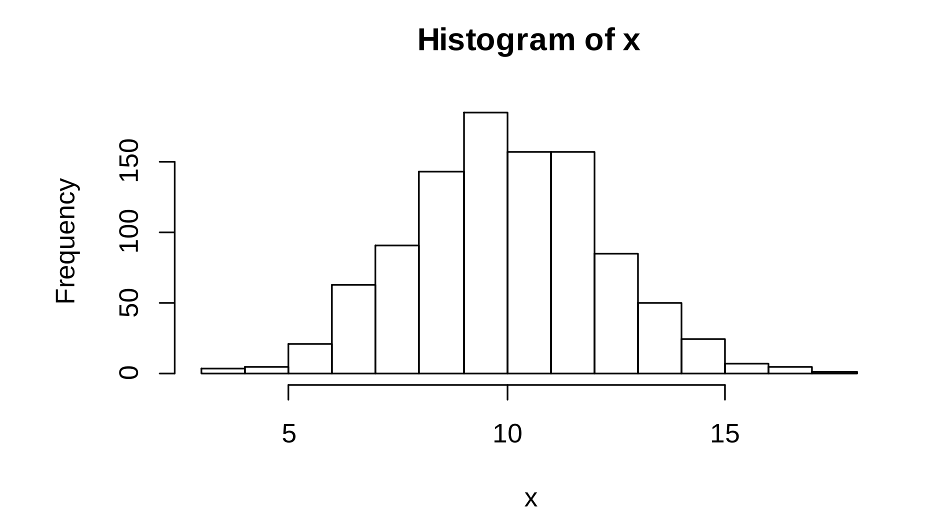
정규분포의 확률, 난수, 분위수 계산

예시 4

난수로 분포 확인하기

평균이 10, 분산이 5인 정규분포에서 1,000개의 난수를 발생하여 이 1000개의 평균과 분산이 각각 10과 5에 가까운지 확인해 보고 1,000개의 분포가 종모양, 좌우대칭인지 히스토그램으로 확인해보자.

```
> x <- rnorm(1000, 10, sqrt(5))
> mean(x)
[1] 9.974489
> var(x)
[1] 4.913132
> hist(x)
```



참고 : 발생한 x는 난수이므로 위의 프로그램은 실행할 때마다 약간씩 다른 결과를 얻을 수 있음에 유의

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

정규분포 난수 만들기

- 일양분포로 정규분포 난수
- Box-Muller 변환

통계청 통계교육원

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

Box-Muller 변환

u_1과 u_2가 독립이며 일양분포 $U(0,1)$을 따를 경우

$$z_1 = \sqrt{-2\log(u_1)}\ cos(2\pi u_2)$$

$$z_2 = \sqrt{-2\log(u_1)}\ sin(2\pi u_2)$$

- 위와 같이 z_1과 z_2를 얻으면 z_1과 z_2의 분포는 독립인 표준정규분포를 따름
- 이를 Box-Muller 변환이라고 함

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

Box-Muller 변환

예시 1

난수의 응용

Box-Muller 변환을 사용하여 n개의 표준정규분포를 따르는 난수를 생성하는 함수를 작성하여 난수를 발생하여 얻은 결과를 사용하여 히스토그램을 그려서 분포가 표준정규분포에 가까운지 확인해보자.

```
boxmuller <- function(size) {
  if (size %% 2 == 0) {
     nn <- size/2
  } else {
     nn <- (size + 1)/2
  } # end if-else

  u1 <- runif(nn)
  u2 <- runif(nn)

  z1 <- sqrt(-2*log(u1))*cos(2*pi*u2)
  z2 <- sqrt(-2*log(u1))*sin(2*pi*u2)

  z <- c(z1, z2)

  if (size %% 2 != 0) {
     z <- z[-1]
  } # end if
  list (z = z)
} # end function
```

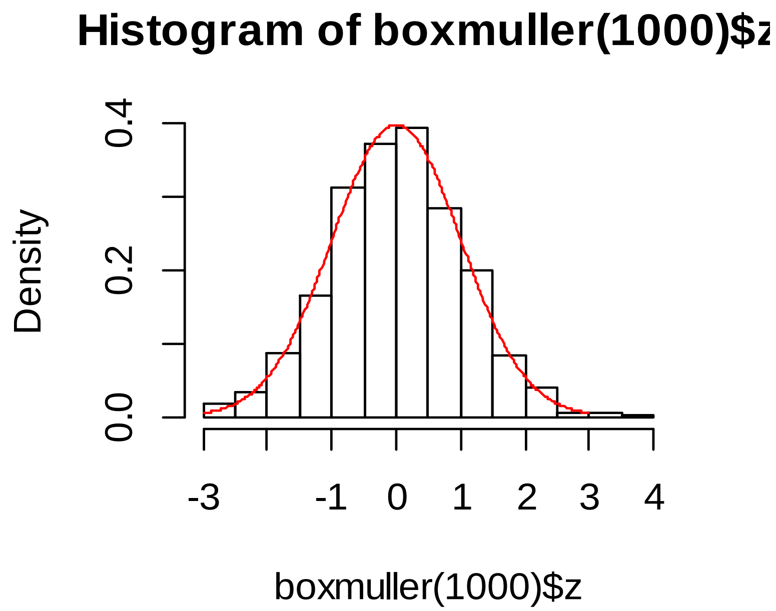

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

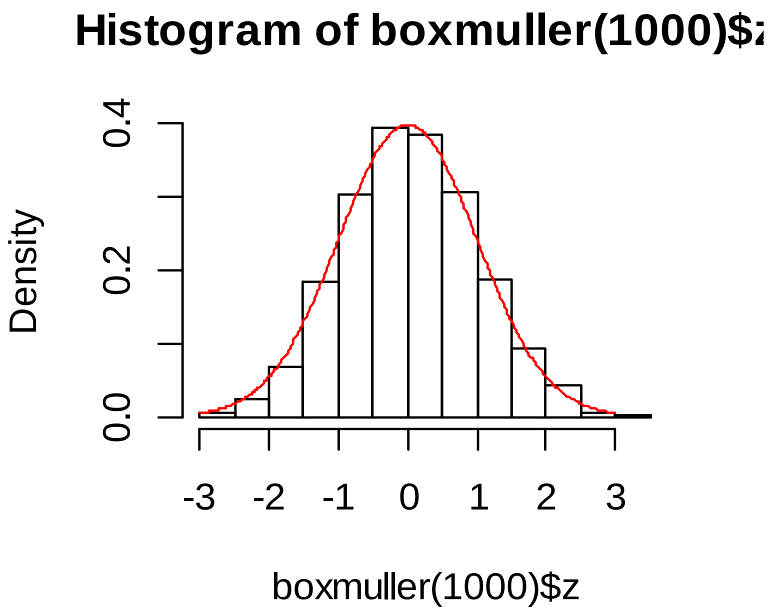
Box-Muller 변환

예시 1

함수 사용 결과

```
> hist(boxmuller(1000)$z, prob=T)
> lines(x <- seq(-3,3, by=0.02), dnorm(x), col= "red")
```





본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

Box-Muller 변환

예시 2

Box-Muller 변환과 rnorm의 비교

Box-Muller 변환으로 얻은 10,000개의 난수에서 얻은 10%, 20%, ... , 90% 백분위수와 표준정규분포의 이론적 백분위수를 비교해보자.

```
> x <- boxmuller(10000)$z
> p <- seq(0.1, 0.9, by=0.1)
> quantile(x, probs=p)
         10%          20%          30%          40%          50%
-1.308919302  -0.844256093  -0.528685954  -0.247099466  0.006313709
         60%          70%          80%          90%
 0.244724218   0.512521851   0.839710845   1.278103702
> qnorm(p)
[1]  -1.2815516  -0.8416212  -0.5244005  -0.2533471   0.0000000   0.2533471
[7]   0.5244005   0.8416212   1.2815516
```

* 가까운 값임을 알 수 있음

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.

Box-Muller 변환

이론적인 값과 난수에서 얻은 값의 차이

quantile와 qnorm의 결과의 차이 계산

```
> quantile(x, probs=p) - qnorm(p) # 차이 계산
         10%          20%          30%          40%          50%
-0.027367736 -0.002634859 -0.004285441  0.006247637  0.006313709
         60%          70%          80%          90%
-0.008622885 -0.011878662 -0.001910388 -0.003447864
```

* 차이는 0에 가까움

통계청 통계교육원

본 문서는 저작권자의 동의 없이 무단 복제 및 배포할 수 없습니다.