

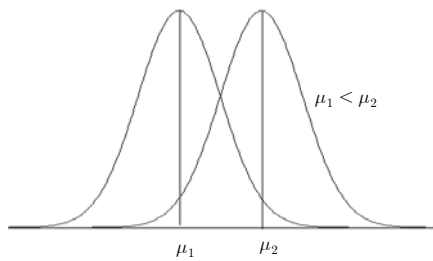
## 제 4장. 모집단과 표본

### 4.1 정규분포

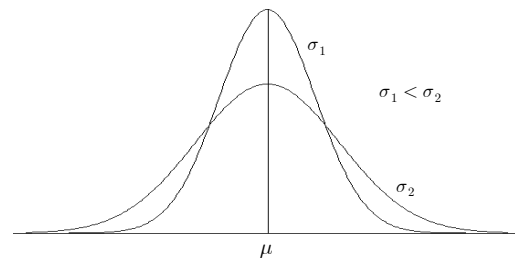
정규분포는 가우스(Gauss, 1777-1855)에 의해 제시된 분포로써 가우스 분포(Gauss distribution)라고 불리며 가장 대표적인 연속형 확률 분포이다. 평균이  $\mu$  이고 표준편차가  $\sigma$  인 정규분포의 밀도 함수는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty$$

확률변수  $X$  가 평균이  $\mu$ , 표준편차가  $\sigma$  인 정규분포를 따를 때, 이를  $X \sim N(\mu, \sigma^2)$ 로 나타낸다. 정규분포는 평균을 중심으로 좌우 대칭의 모양이며 대칭점과 변곡점 사이의 거리가 표준편차이다.

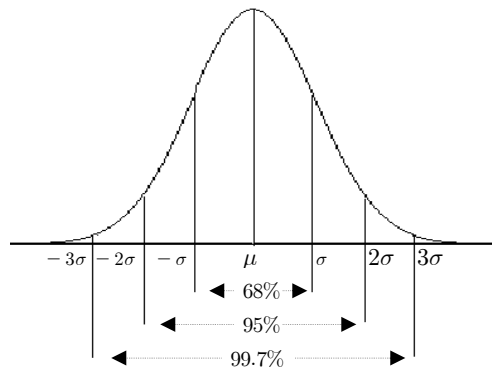


<표준편차는 같고 평균이 다른 두 정규분포>



<평균은 같고 표준편차가 다른 두 정규분포>

$X$ 가 정규분포  $N(\mu, \sigma^2)$ 을 따를 때



표준 정규분포 : 평균이 0이고 표준편차가 1인 정규분포.

$$X \sim N(\mu, \sigma^2) \text{ 일 때, } Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

#### 4.1.1 정규분포의 확률밀도함수 그리기

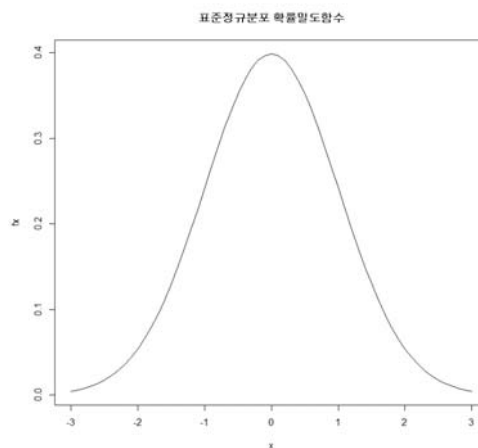
정규분포의 확률밀도는 `dnorm()` 함수를 사용한다.

```
> dnorm(x, mean=mu, sd=sigma)
:  $X \sim N(\mu, \sigma^2)$  일 때  $\Pr(X=x)$  값을 구함
```

다음은 표준 정규분포의 확률밀도함수를 그리는 방법이다.

```
x<-seq(from=-3, to=3, by=0.1)
fx<-dnorm(x, mean=0, sd=1)
plot(x, fx, type="l", main="표준정규분포 확률밀도함수")
```

- ▶ `seq(from=a, to=b, by=c)` : a부터 b의 구간에서 간격이 c인 벡터를 생성
- ▶ `plot(x, y, type="l")` : (x,y)의 산점도를 선으로 연결하여 그림



#### 4.1.2 정규분포의 확률 계산

정규분포의 누적 분포는 `pnorm()`을 사용한다.

```
> pnorm(x, mean=mu, sd=sigma)
:  $X \sim N(\mu, \sigma^2)$  일 때  $\Pr(X \leq x)$  값을 구함
```

예 ) A회사에서 생산되는 전구의 수명은 평균이 2000시간이고 표준편차가 200시간인 정규분포를 따른다고 한다.

1) 이 회사 제품인 전구의 수명이 2500 시간 이하일 확률을 구하여라

: 전구의 수명을  $X$ 라고 한다면  $X \sim N(2000, 200^2)$ 의 분포를 따른다. 따라서 구하는 확률은  $\Pr(X < 2500)$  이므로 결과를 확인해 보면 다음과 같다.

```
> pnorm(2500, mean=2000, sd=200)
[1] 0.9937903
>
```

2) 이 회사 제품인 전구의 수명이 1800 시간 이상일 확률을 구하여라  
: 구하는 확률은  $\Pr(X > 1800)$  이므로 다음과 같이 구할 수 있다.

```
> 1-pnorm(1800, mean=2000, sd=200)
[1] 0.8413447
>
```

또는 pnorm()의 lower.tail 옵션을 이용하면 다음과 같다.

```
> pnorm(1800, mean=2000, sd=200, lower.tail=F)
[1] 0.8413447
>
```

#### 4.1.3 정규 분위수의 계산

정규분포의 분위수를 구하기 위해서는 qnorm()을 사용한다.

```
> qnorm(p, mean=mu, sd=sigma)
:  $X \sim N(\mu, \sigma^2)$  일 때,  $\Pr(X \leq x) = p$  를 만족하는  $x$ 를 구함
```

예 ) IQ 테스트에서 상위 2% 이내의 점수를 받아야 멘사 회원의 자격이 주어진다고 한다. 만약 IQ 테스트의 점수가 평균이 100이고 표준편차가 15인 정규분포를 따를 때, 멘사 회원이 되기 위해서는 약 몇 점 이상의 점수를 받아야 하는가?

: IQ 테스트의 점수를  $X$  라고 한다면  $X \sim N(100, 15^2)$  이고 구하는 값은  $\Pr(X \leq x) = 0.98$  을 만족하는 점수  $x$  이다. 결과는 다음과 같다.

```
> qnorm(0.98, mean=100, sd=15)
[1] 130.8062
>
> qnorm(0.02, mean=100, sd=15, lower.tail=F)
[1] 130.8062
>
```

qnorm() 함수도 pnorm() 함수와 마찬가지로 lower.tail 옵션을 사용할 수 있다.

## 4.2 이항 분포

### 베르누이 시행

: 시행의 결과가 (성공, 실패)와 같이 두 가지 결과 중의 하나로 나타나는 시행

### 이항 분포

: 성공 확률이  $p$ 인 베르누이 시행을  $n$ 회 반복할 때, ' $X$  = 성공의 횟수'의 분포

시행 횟수가  $n$  이고 성공의 확률이  $p$ 인 이항분포의 확률밀도함수는 다음과 같다.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

#### 이항분포의 평균, 분산

:  $X$ 가 이항분포  $B(n, p)$ 를 따를 때

$$E(X) = np, \quad \text{Var}(X) = np(1-p)$$

이항분포의 정규근사 :  $X$ 가 이항분포  $B(n, p)$ 를 따를 때

$$(1) X \sim N(np, np(1-p))$$

$$(2) Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

참고. 이항분포의 정규근사는  $n$ 이 충분히 커서 ' $np \geq 5$  이고  $n(1-p) \geq 5$ '일 때 사용하는 것이 안전하다.

#### 연속성 수정(continuity correction)

: 두 정수  $a, b$ 에 대하여

$$P(a \leq X \leq b) = P(a - 0.5 \leq X \leq b + 0.5)$$

$$\doteq P\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

#### 4.2.1 이항 분포의 밀도함수 그리기

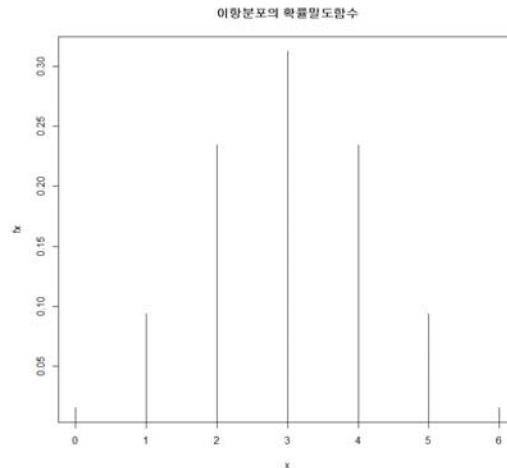
이항분포의 확률 밀도는 `dbinom()`을 사용한다.

> `dbinom(x, size=n, prob=p)`  
 :  $X \sim B(n, p)$  일 때,  $\Pr(X=x)$ 의 값을 구함

이항분포의 확률밀도함수는 다음과 같이 그릴 수 있다.

```
x<-0:6
fx<-dbinom(x, 6, 0.5)
plot(x, fx, main="이항분포의 확률밀도함수", type="h")
```

▶ `plot(x, fx, type="h")` : 막대 형태의 히스토그램으로 그래프를 그리는 옵션



#### 4.2.2 이항 분포의 확률 계산

이항 분포의 누적확률은 pbinom()을 이용한다.

```
> pbinom(x, size=n, prob=p)
:  $X \sim B(n, p)$  일 때,  $\Pr(X \leq x)$ 의 값을 구함
```

예 1) 어느 생산 공정의 불량률은 10%라고 한다. 이 공정에서 임의로 5개의 표본을 추출 할 때, 불량품이 3개일 확률을 구하여라.

: 불량품의 개수를  $X$  라고 하면  $X \sim B(5, 0.1)$ 을 따른다. 따라서 구하는 확률  $\Pr(X=3)$ 은 다음과 같다.

```
> dbinom(3,size=5, prob=0.1)
[1] 0.0081
>
```

예 2) 그렇다면 임의로 추출한 20개의 표본 중에서 불량품이 5개 이하일 확률은 얼마인가?

: 불량품의 개수  $X$  는  $X \sim B(20, 0.1)$ 의 분포를 따르고, 구하는 확률은  $\Pr(X \leq 5)$ 이다.

```
> pbinom(5,size=20, prob=0.1)
[1] 0.9887469
>
```

#### 4.2.3 이항 분포의 정규 근사

표본의 크기에 따라 이항분포의 정규근사가 어떻게 달라지는 지 확인해보자. 다음의 코드를 스크립트 창에 입력하고 실행한 후, 결과를 확인해보면 다음과 같다.

```

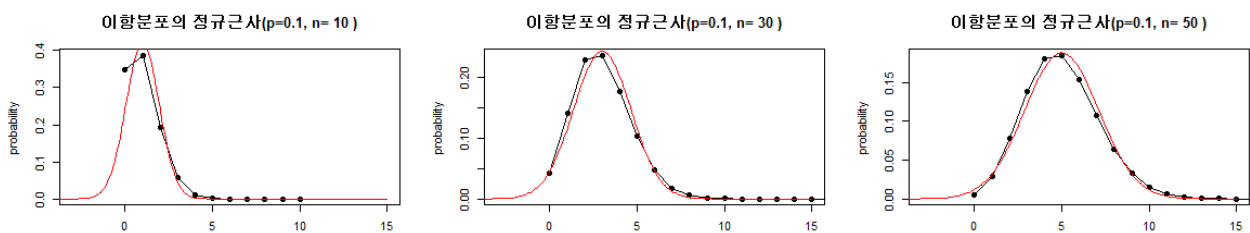
p<-0.1
n<-c(10,30,50)

par(mfrow=c(1,3))
for (i in 1:3){
  x<-seq(from=0, to=n[i], by=1)          #이항 분포의 그래프를 그리는 부분
  fx<-dbinom(x, n[i], p)
  plot(x, fx, pch=16, xlim=c(-3,15), ylab="probability",xlab="",
       main=paste("이항분포의 정규근사(p=",p,", n=",n[i],")"))
  lines(x, fx)

  y<-seq(from=-5, to=15, by=0.1)        #근사된 정규분포의 그래프를 그리는 부분
  mu<-n[i]*p
  sd<-sqrt(n[i]*p*(1-p))
  fy<-dnorm(y, mu, sd)
  lines(y, fy, col="red")
}

```

- ▶ `par(mfrow=c(nc, nr))` : `nr`개의 행과 `nc`개의 열을 갖는 다중 그래프 창을 생성함. 여러개의 그래프를 한 화면에 그리고자 할 때 사용한다
- ▶ `plot(x,y, pch=16)` : 산점도를 그릴 때 점의 모양을 바꾸는 명령어. 각 번호에 해당되는 점의 모양이 사전에 지정되어 있다.
- ▶ `lines(x, y)` : `(x, y)` 의 그래프를 선으로 이어서 그리는 명령어. 기존에 이미 생성되어 있는 그래프 위에 선 그림을 더해서 그린다.



예 )  $p = 0.5$  인 경우에 이항분포의 정규근사를 실행해 보고 결과를 확인해보자

### 4.3 표본 평균의 분포

#### 1) 모집단의 분포가 정규분포인 경우

모집단의 분포가 정규분포  $N(\mu, \sigma^2)$  일 때, 표본평균  $\bar{X}$ 는 정규분포  $N\left(\mu, \frac{\sigma^2}{n}\right)$  를 따른다.

#### 2) 모집단의 분포가 정규분포가 아닌 경우

평균이  $\mu$  이고 분산이  $\sigma^2$ 인 임의의 무한 모집단에서 표본의 크기  $n$ 이 충분히 크면, 랜덤 표본의 표본평균  $\bar{X}$ 는 근사적으로 정규분포  $N\left(\mu, \frac{\sigma^2}{n}\right)$ 를 따른다.

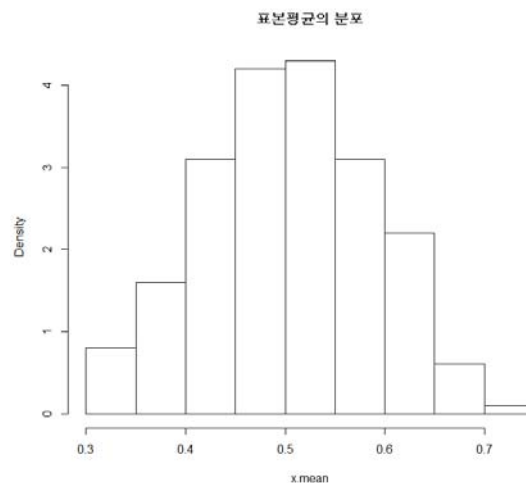
: 중심극한정리 (Central Limit Theorem)

예 1) 균등분포  $U(0,1)$  에서 10개의 표본을 추출하여 표본평균  $\bar{X}$ 를 구하는 실험을 200회 반복하고, 이들 200개의 표본평균들을 히스토그램으로 나타내보자.

```
n<-10                                # 1회 시행에서 추출할 표본의 개수
x.mean<-c()                          # 각 시행에서 추출된 표본의 평균을 저장할 벡터

for (j in 1:200){                    # 총 200번의 반복시행을 위한 반복문
  x<-runif(n, min=0, max=1)
  x.mean[j]<-mean(x) }
hist(x.mean, xlim=c(0,1), probability=T, main=paste("표본평균의 분포 n=",n))
```

▶ `runif(n, min=a, max=b)` : (a,b)를 모수로 갖는 균등분포에서  $n$ 개의 랜덤 표본을 생성한다

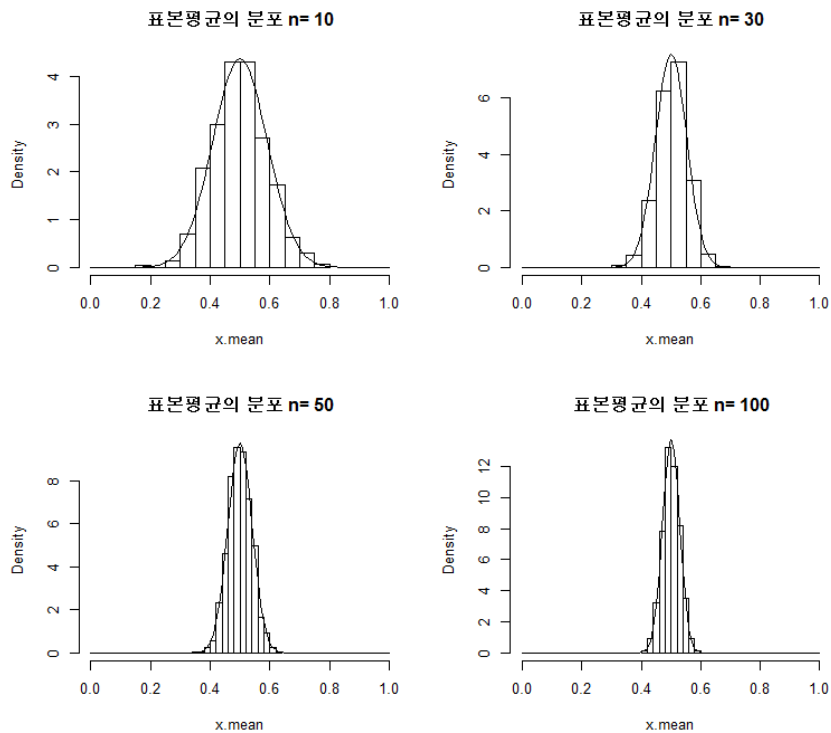


예 2) 이번에는 표본의 개수를 20개, 30개, 50개로 증가시키며 실험을 반복해보자. 표본 평균의 분포는 어떻게 달라지는가? 주어진 스크립트를 실행해보고 그 결과를 확인해보자.

```
n<-c(10,30,50,100)
par(mfrow=c(1,4))                                # 다중 그래프창 생성

for (i in 1:4){                                    # 총 4번의 서로 다른 표본의 개수로 실험
  x.mean<-c()
  for (j in 1:1000){                                # 각 표본의 크기별로 1000번씩 반복
    x<-runif(n[i], 0, 1)
    x.mean[j]<-mean(x) }
  hist(x.mean, xlim=c(0,1), probability=T, main=paste("표본평균의 분포 n=",n[i]))

  y<-seq(0, 1, 0.01)
  mu<-0.5                                           # (0,1) 균등분포의 모평균
  sigma<-sqrt(1/12)                                # (0,1) 균등분포의 모표준편차
  fy<-dnorm(y, mu, sigma/sqrt(n[i]))
  lines(y, fy) }                                    #표본평균이 근사적으로 따르는 정규 분포
```





#### 4.4 정규분포 분위수 대조도

: 정규모집단의 가정을 검토하는 방법으로 정규분포의 분위수와 이에 대응하는 자료 분포의 분위수를 좌표평면에 나타낸 그림을 말한다. 점들이 직선 주위에 밀집하여 나타날수록 모집단의 분포가 정규분포라는 가정이 타당하다고 할 수 있다.

예 ) (bodydims.csv) 주어진 자료는 247명의 남성과 260명의 여성에 관한 신체 측정 자료이다. 원 자료는 신체의 각 부위를 측정한 총 25개의 변수가 있으며 본 예제에서는 그 중 8개 변수만 간추린 자료를 이용하기로 한다. 아래는 8개의 변수에 대한 설명이다. (원자료에 관한 자세한 설명은 다음을 참조하도록 한다. :

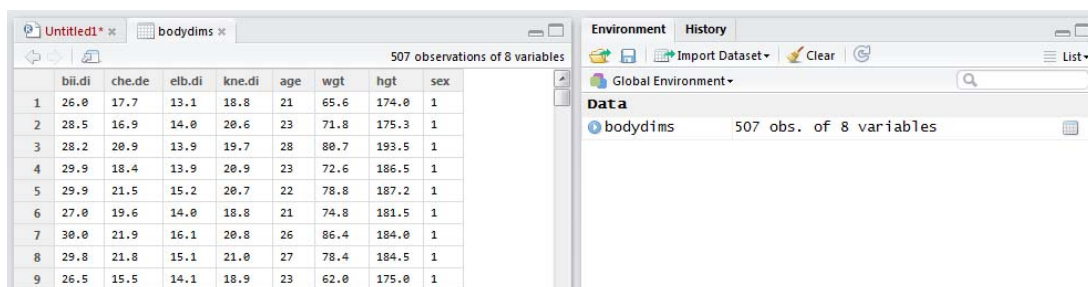
<http://www.openintro.org/stat/data/bdims.php>)

- bii.di : 숫자형 변수, 응답자의 골반의 넓이 (cm)
- che.de : 숫자형 변수, 응답자의 가슴 깊이 (cm)
- elb.di : 숫자형 변수, 응답자의 양쪽 팔꿈치 지름의 합. (cm)
- kne.di : 숫자형 변수, 응답자의 양쪽 무릎의 지름의 합. (cm)
- age : 숫자형 변수, 응답자의 나이 (years)
- wgt : 숫자형 변수, 응답자의 몸무게 (kg)
- hgt : 숫자형 변수, 응답자의 신장 (cm)
- sex : 범주형 변수, 응답자의 성별 (1=남성, 0=여성)

csv 형태의 자료를 읽기 위해서는 read.csv()를 사용한다.

```
> read.csv("파일 저장 경로", header=T)
```

다음은 bodydims.csv 파일을 불러와서 bodydims 라는 변수명으로 저장한 결과이다. 자료를 확인해 보면 총 507개의 관찰치와 8개의 변수로 구성된 것을 볼 수 있다.

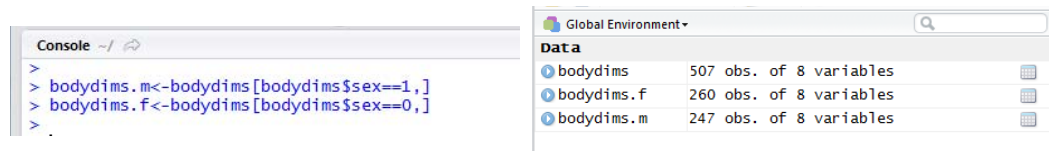


|   | bii.di | che.de | elb.di | kne.di | age | wgt  | hgt   | sex |
|---|--------|--------|--------|--------|-----|------|-------|-----|
| 1 | 26.0   | 17.7   | 13.1   | 18.8   | 21  | 65.6 | 174.0 | 1   |
| 2 | 28.5   | 16.9   | 14.0   | 20.6   | 23  | 71.8 | 175.3 | 1   |
| 3 | 28.2   | 20.9   | 13.9   | 19.7   | 28  | 80.7 | 193.5 | 1   |
| 4 | 29.9   | 18.4   | 13.9   | 20.9   | 23  | 72.6 | 186.5 | 1   |
| 5 | 29.9   | 21.5   | 15.2   | 20.7   | 22  | 78.8 | 187.2 | 1   |
| 6 | 27.0   | 19.6   | 14.0   | 18.8   | 21  | 74.8 | 181.5 | 1   |
| 7 | 30.0   | 21.9   | 16.1   | 20.8   | 26  | 86.4 | 184.0 | 1   |
| 8 | 29.8   | 21.0   | 15.1   | 21.0   | 27  | 78.4 | 184.5 | 1   |
| 9 | 26.5   | 15.5   | 14.1   | 18.9   | 23  | 62.0 | 175.0 | 1   |

예 1) 주어진 자료를 성별에 따라 두 개의 데이터셋으로 나누어 보자. 이 중 여성의 데이터셋을 이용하여 bii.di 변수에 대해 히스토그램과 정규분포 분위수 대조도를 그려보자. 자료는 정규분포를 따른다고 말 할 수 있는가?

먼저 주어진 자료를 성별에 따라 나누어서 각각 bodydims.m 과 bodydims.f 로 저장해보자. 주어진 데이터 프레임의 성별 변수값 중 남성은 1, 여성은 0으로 되어있다. 특정 조건을 만족

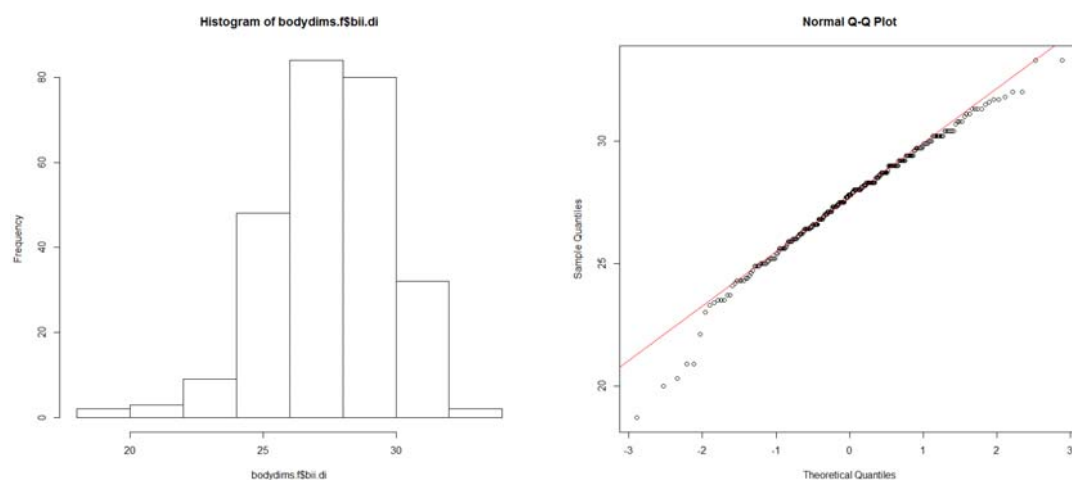
하는 부분집합을 선택하기 위해서는 논리 연산자를 다음과 같이 이용할 수 있다. 주어진 스크립트를 실행하면 두 개의 새로운 데이터셋이 생성된 것을 확인할 수 있다.



정규분포 분위수 대조도를 그리기 위해서는 qqnorm()을 사용하고 qqline()을 이용해서 데이터가 만족해야 하는 직선관계를 그릴 수 있다.

```
> qqnorm(bodydims.f$bii.di)
> qqline(bodydims.f$bii.di, col=2)
```

아래는 주어진 변수의 히스토그램과 정규분포 분위수를 나타낸 것이다.



예 2) 여성의 데이터셋을 이용하여 elb.di, che.de 변수에 대해서도 각각 히스토그램과 정규분포 분위수 대조도를 그려보자. 주어진 변수들은 정규분포를 따른다고 볼 수 있는가?

#### 4.5 자료를 이용한 예제 (ames.csv)

주어진 자료는 Iowa의 도시 Ames의 2006년부터 2010년 사이의 부동산 거래내역 자료이다. 5년 동안 이 지역에서 발생한 총 2930건의 부동산 거래내역이 모두 기록되어 있다. 자료에 대한 자세한 설명은 다음을 참조한다.

(<http://www.openintro.org/stat/data/?data=ames>)

예제 1. 현재 주어진 자료는 일정 기간동안 지역 내의 모든 부동산 거래를 기록한 자료이므로 일종의 모집단이라고 생각할 수 있다. SalePrice 변수에 대해 히스토그램을 그려보고 수치적 요약값을 구해보자. 모집단의 분포는 어떠한가?

예제 2. 이 지역에서 발생한 전체 부동산 거래 가격의 평균값( $\mu$ )을 추정해보려고 한다. 지금처럼 모집단 전체를 알게 되는 경우는 매우 드물기 때문에 대부분의 경우에는 모집단의 부분집합인 표본을 선택하여 모수를 추정하게 된다. SalePrice에서 50개의 랜덤 표본을 선택해보자. 이 때, 모평균의 추정값은 무엇인가?

예제 3. 예제 2의 과정을 5000번 반복하여 표본 평균의 표본 분포를 구해보자. 즉, 크기가 50인 랜덤 표본을 선택하여 표본평균을 구하는 과정을 5000번 반복하고 이 결과를 sample\_mean50이라는 이름의 벡터에 저장을 한다. sample\_mean50을 이용하여 히스토그램을 sample\_mean50 그려보자. 표본 평균의 분포는 어떠한가?

예제 4. 예제 3의 sample\_mean50의 평균과 분산을 계산해보자. sample\_mean50의 평균값은 모집단의 평균과 어떠한 관계가 있는가? sample\_mean50의 분산값은 모분산과 어떠한 관계가 있는가?

예제 5. 예제 3의 과정을 표본의 크기를 150으로 증가시켜 반복해보자. 이 결과는 sample\_mean150에 저장한다. 표본의 크기에 따른 표본 평균의 분포는 어떠한가?