

# Fitting Normal Distribution

coop711  
2015년 3월 8일

## 자료 생성

- 케틀레가 작성한 스코틀랜드 군인 5738명의 가슴둘레(인치) 분포표로부터

```
chest.circum<-c(rep(33,3),rep(34,18),rep(35,81),rep(36,185),rep(37,420), rep(38,749),rep(39,1073),rep(40,1079),rep(41,934),rep(42,658),rep(43,370),rep(44,92),rep(45,50),rep(46,21),rep(47,4),rep(48,1))
```

- 33인치인 사람이 3명, 34인치인 사람이 18명 등으로 기록되어 있으나 이는 구간의 가운데로 파악하여야 함.

## 기초통계와 표준편차 계산

```
summary(chest.circum)
```

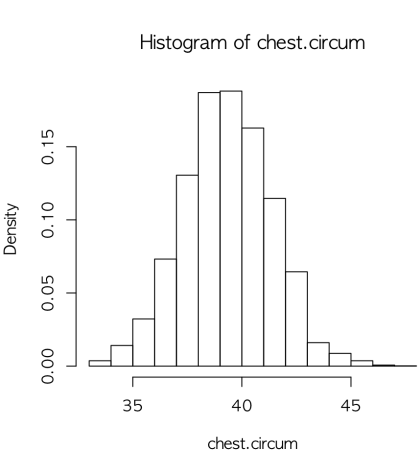
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	33.00	38.00	40.00	39.83	41.00	48.00

```
sd(chest.circum)
```

```
## [1] 2.049616
```

## 히스토그램으로 나타내기

- 히스토그램을 직관적으로 그려보면 y축은 숫자가 기본값임을 알 수 있음.

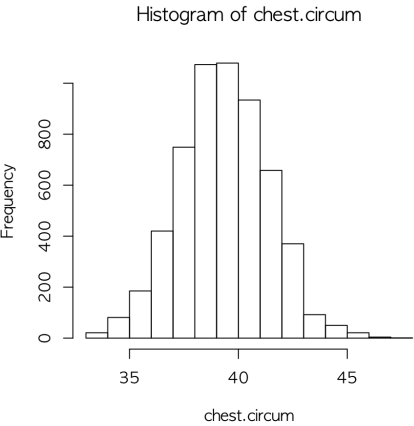


## 히스토그램의 내부 계산

- 실제로 이 히스토그램을 그리는 데 계산된 값들은?

```
h.circum<-hist(chest.circum, plot=F)  
h.circum
```

```
hist(chest.circum)
```



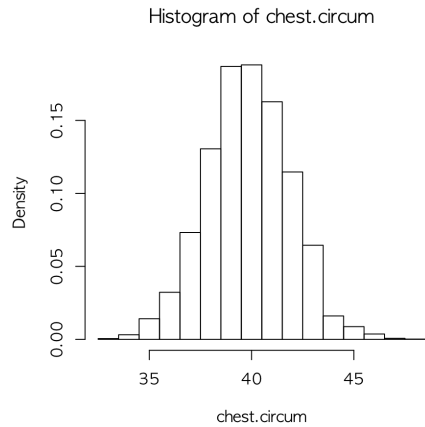
- 정규분포와 비교하기 위해서 y축을 확률로 나타내려면

```
hist(chest.circum,prob=T)
```

```
## $breaks  
## [1] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48  
##  
## $counts  
## [1] 21 81 185 420 749 1073 1079 934 658 370 92 50 21 4  
## [15] 1  
##  
## $density  
## [1] 0.0036598118 0.0141164169 0.0322411990 0.0731962356 0.1305332869  
## [6] 0.1869989543 0.1880446148 0.1627744859 0.1146741025 0.0644823980  
## [11] 0.0160334611 0.0087138376 0.0036598118 0.0006971070 0.0001742768  
##  
## $mids  
## [1] 33.5 34.5 35.5 36.5 37.5 38.5 39.5 40.5 41.5 42.5 43.5 44.5 45.5 46.5  
## [15] 47.5  
##  
## $xname  
## [1] "chest.circum"  
##  
## $equidist  
## [1] TRUE  
##  
## attr(,"class")  
## [1] "histogram"
```

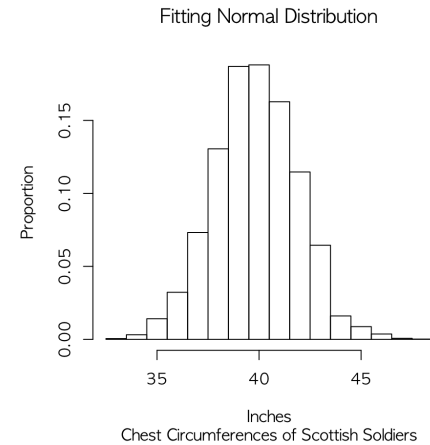
- 평균값과 표준편차로부터 히스토그램의 위치가 0.5만큼 왼쪽으로 치우쳐 있다는 것을 알 수 있음. 제자리에 옮겨 놓기 위해서 breaks 매개변수를 32.5부터 48.5까지 1간격으로 설정

```
hist(chest.circum,prob=T,breaks=32.5:48.5)
```



- 히스토그램을 보기 쉽게 하기 위해서 메인 타이틀과 서브 타이틀, x축 라벨, y축 라벨 설정

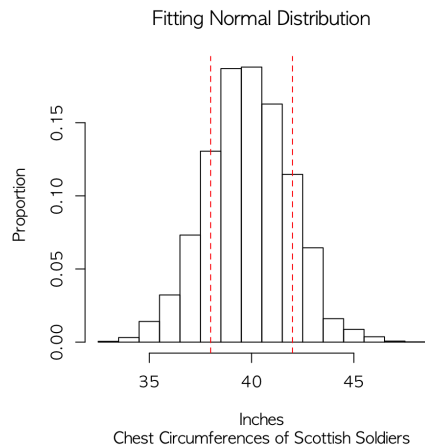
```
main.title<-"Fitting Normal Distribution"
sub.title<-"Chest Circumferences of Scottish Soldiers"
x.lab<-"Inches"
y.lab<-"Proportion"
hist(chest.circum,breaks=32.5:48.5,prob=T,main=main.title,sub=sub.title,xlab=x.lab,ylab=y.lab)
```



## 평균 $\pm$ 표준편차의 의미

- 평균을 중심으로  $\pm$  표준편차 만큼 떨어진 자료를 붉은 색 수직점선으로 표시.

```
hist(chest.circum,breaks=32.5:48.5,prob=T,main=main.title,sub=sub.title,xlab=x.lab,ylab=y.lab)
abline(v=c(38,42),lty=2,col="red")
```



- 그 사이의 영역을 빗금으로 표시하기 위하여 다각형의 좌표를 계산

```
h.chest<-hist(chest.circum, breaks=32.5:48.5, plot=F)
h.chest
```

```
## $breaks
## [1] 32.5 33.5 34.5 35.5 36.5 37.5 38.5 39.5 40.5 41.5 42.5 43.5 44.5 45.5
## [15] 46.5 47.5 48.5
##
## $counts
## [1] 3 18 81 185 420 749 1073 1079 934 658 370 92 50 21
## [15] 4 1
##
## $density
## [1] 0.0005228303 0.0031369815 0.0141164169 0.0322411990 0.0731962356
## [6] 0.1305332869 0.1869989543 0.1880446148 0.1627744859 0.1146741025
## [11] 0.0644823980 0.0160334611 0.0087138376 0.0036598118 0.0006971070
## [16] 0.0001742768
##
## $mids
## [1] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
##
## $xname
## [1] "chest.circum"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
h.chest$density[6:10]
```

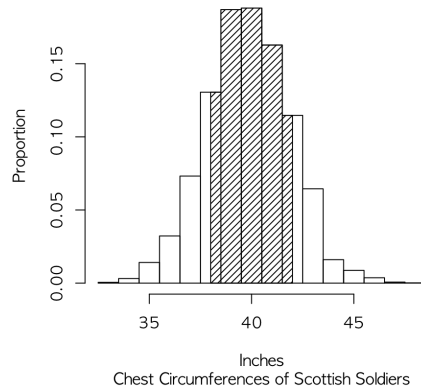
```
## [1] 0.1305333 0.1869990 0.1880446 0.1627745 0.1146741
```

```
y<-h.chest$density[6:10]
```

- 5개의 직사각형으로 파악하고 향후 면적 계산을 쉽게 하기 위하여 다음과 같이 좌표 설정

```
x.coord<-c(38,38.5:41.5,42,42,rep(41.5:38.5,each=2),38,38)
y.coord<-c(rep(0,6),rep(rev(y),each=2),0)
hist(chest.circum,breaks=32.5:48.5,prob=T,main=main.title,sub=sub.title,xlab=x.lab,ylab=y.lab)
polygon(x.coord,y.coord,density=20)
```

Fitting Normal Distribution



- 이론적으로 빗금친 부분의 면적은  $\text{pnorm}(1) - \text{pnorm}(-1) = 0.6826895$ 에 가까울 것으로 예상. 5개의 직사각형의 면적을 구하여 합하는 과정은 다음과 같음.

```
x.coord[1:6]
```

```
quetelet.chest<-data.frame(chest,freq)
quetelet.chest$prob<-quetelet.chest$freq/sum(quetelet.chest$freq)
```

- 위와 같이 하나의 벡터로 모아서 히스토그램 작업을 하려면 다음과 같이 loop 이용

```
quetelet.chest.vec<-rep(33,3)
for (i in 34:48) {
  quetelet.chest.vec<-c(quetelet.chest.vec,rep(i,quetelet.chest$freq[i-32]))
}
length(quetelet.chest.vec)
```

```
## [1] 5738
```

- 5738개의 자료를 확인하고 종전과 같은 작업 수행

```
hist(quetelet.chest.vec,breaks=32.5:48.5,prob=T,main=main.title,sub=sub.title,xlab=x.lab,ylab=y.lab)
abline(v=c(38,42),lty=2,col="red")
```

```
## [1] 38.0 38.5 39.5 40.5 41.5 42.0
```

```
y
```

```
## [1] 0.1305333 0.1869990 0.1880446 0.1627745 0.1146741
```

```
diff(x.coord[1:6])
```

```
## [1] 0.5 1.0 1.0 1.0 0.5
```

```
diff(x.coord[1:6])*y
```

```
## [1] 0.06526664 0.18699895 0.18804461 0.16277449 0.05733705
```

```
sum(diff(x.coord[1:6])*y)
```

```
## [1] 0.6604217
```

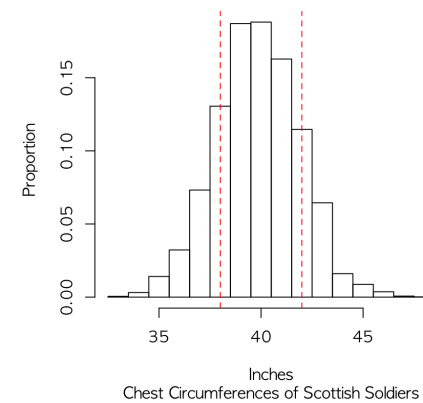
## data.frame으로 관리하기

- 자료 관리의 편의성을 높이기 위하여 가습틀레 자료를 **data.frame**으로 만드는 방법은

```
chest<-33:48
freq<-c(3,18,81,185,420,749,1073,1079,934,658,370,92,50,21,4,1)
sum(freq)
```

```
## [1] 5738
```

Fitting Normal Distribution

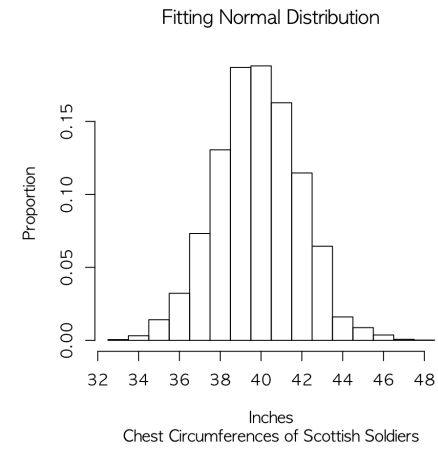
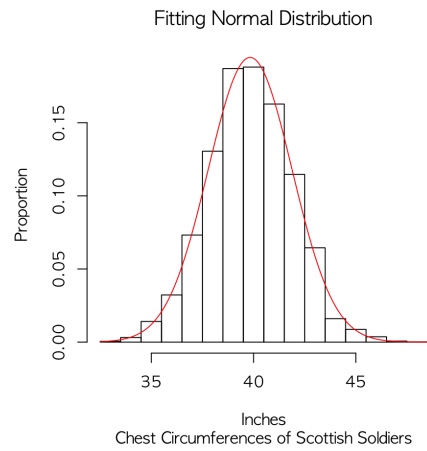


## 정규분포 밀도함수와 비교

- 이론적인 정규분포 밀도함수 곡선을 히스토그램에 덧붙여 그림.

```
mean.chest<-mean(quetelet.chest.vec)
sd.chest<-sd(quetelet.chest.vec)
x<-seq(32.5,48.5,length=1000)
y.norm<-dnorm(x,mean=mean.chest,sd=sd.chest)
hist(quetelet.chest.vec,breaks=32.5:48.5,prob=T,main=main.title,sub=sub.title,xlab=x.lab,ylab=y.lab)
lines(x,y.norm,col="red")
```

```
hist(chest.circum,breaks=32.5:48.5,prob=T,main=main.title,sub=sub.title,xlab=x.lab,ylab=y.lab,axes=F)
axis(side=1,at=seq(32,48,by=2),labels=paste(seq(32,48,by=2)))
axis(side=2)
```



x축의 눈금 조정