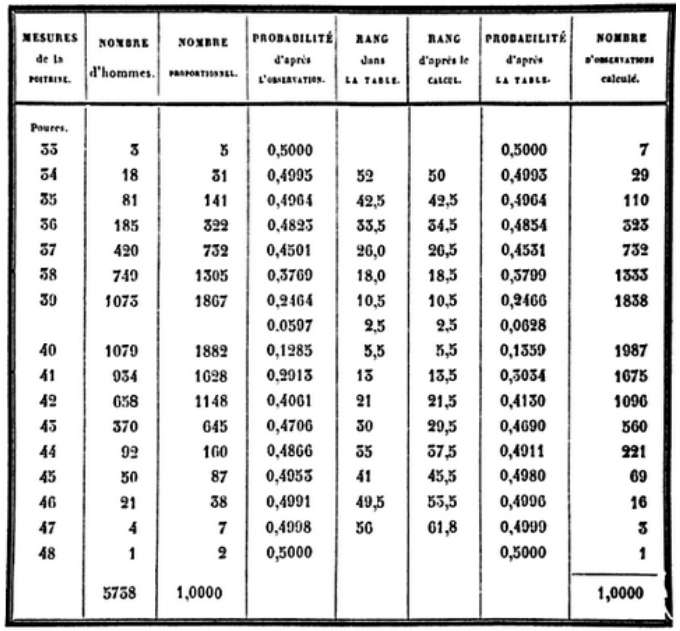
Fitting Normal Distribution

coop711

2016-03-14

Data

From Stigler's



MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE d'observations calculé.
Pouces.							
33	3	5	0,5000			0,5000	7
34	18	31	0,4995	52	50	0,4993	29
35	81	141	0,4964	42,5	42,5	0,4964	110
36	185	322	0,4893	33,5	34,5	0,4854	323
37	420	732	0,4501	26,0	26,5	0,4531	732
38	749	1305	0,3769	18,0	18,5	0,3799	1333
39	1073	1867	0,2464	10,5	10,5	0,2466	1838
			0,0597	2,5	2,5	0,0628	
40	1079	1882	0,1285	5,5	5,5	0,1359	1987
41	934	1628	0,2913	13	13,5	0,3034	1675
42	658	1148	0,4061	21	21,5	0,4130	1096
43	370	645	0,4706	30	29,5	0,4690	560
44	92	160	0,4866	35	37,5	0,4911	221
45	50	87	0,4953	41	45,5	0,4980	69
46	21	38	0,4991	49,5	53,5	0,4996	16
47	4	7	0,4998	56	61,8	0,4999	3
48	1	2	0,5000			0,5000	1
	5738	1,0000					1,0000

Frequency Table

- 케틀레가 작성한 스코틀랜드 군인 5738명의 가슴둘레(인치) 분포표를 옮기면

```
chest <- 33:48
freq <- c(3, 18, 81, 185, 420, 749, 1073, 1079, 934, 658, 370, 92, 50, 21, 4,
1)
(chest.table <- data.frame(Chest = chest, Freq = freq))
```

```
##    Chest Freq
## 1     33    3
## 2     34   18
## 3     35   81
## 4     36  185
## 5     37  420
## 6     38  749
## 7     39 1073
## 8     40 1079
## 9     41  934
## 10    42  658
## 11    43  370
## 12    44   92
## 13    45   50
## 14    46   21
## 15    47    4
## 16    48    1
```

```
str(chest.table)
```

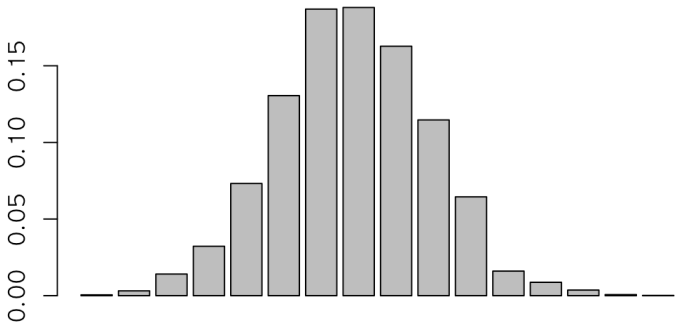
```
## 'data.frame':    16 obs. of  2 variables:
##  $ Chest: int  33 34 35 36 37 38 39 40 41 42 ...
##  $ Freq : num  3 18 81 185 420 ...
```

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

Probability Histogram

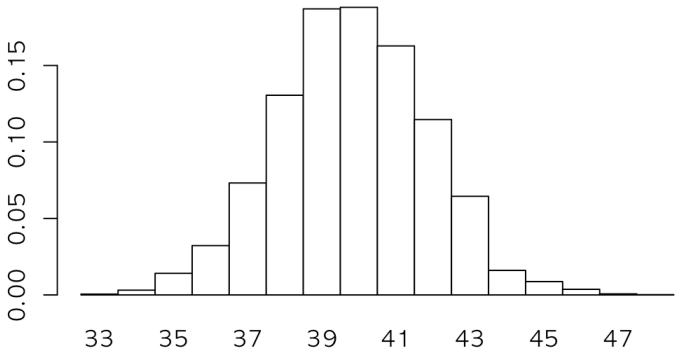
- `barplot(height, ...)` 은 기본적으로 `height` 만 주어지면 그릴 수 있음. 확률 히스토그램의 기둥 면적의 합은 1이므로, 각 기둥의 높이는 각 계급의 돗수를 전체 돗수, 5738(`sum(chest.table$Freq)`)로 나눠준 값임.

```
total <- sum(chest.table$Freq)
barplot(chest.table$Freq/total)
```



- 각 막대의 이름은 계급을 나타내는 가슴둘레 값으로 표현할 수 있고, 막대 간의 사이를 띄우지 않으며, 디폴트 값으로 주어진 회색 보다는 차라리 백색이 나으므로 이를 설정해 주면,

```
barplot(chest.table$Freq/total, names.arg = 33:48, space = 0, col = "white")
```



- 확률 히스토그램의 정의에 따라 이 막대들의 면적을 합하면 1이 됨에 유의.

Summary statistics and SD

- 33인치가 3명, 34인치가 18명 등을 한 줄의 긴 벡터로 나타내어야 평균과 표준편차를 쉽게 계산할 수 있으므로 long format으로 바꾸면,

```
chest.long <- rep(chest.table$Chest, chest.table$Freq)
str(chest.long)
```

```
##  int [1:5738] 33 33 33 34 34 34 34 34 34 34 ...
```

- `chest.long` 을 이용하여 기초통계와 표준편차를 계산하면,

```
summary(chest.long)
```

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

```
sd(chest.long)
```

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

Histogram

- 히스토그램을 직관적으로 그려보면 y축은 돗수가 기본값임을 알 수 있음.

```
hist(chest.long)
```

- 정규분포와 비교하기 위해서 y축을 확률로 나타내려면

```
hist(chest.long, probability = TRUE)
```

Inside the histogram

- 실제로 이 히스토그램을 그리는 데 계산된 값들은?

```
h.chest <- hist(chest.long, plot = FALSE)
list(breaks = h.chest$breaks, counts = h.chest$counts, density = h.chest$densit
y, mids = h.chest$mids)
```

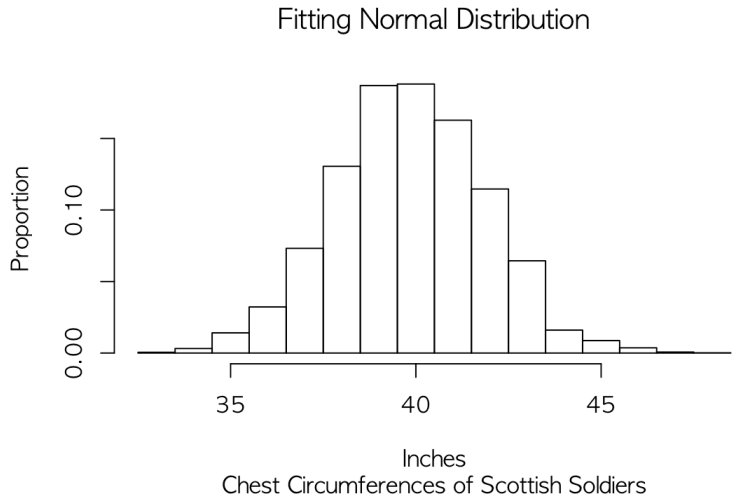
```
## $breaks
##  [1] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
##
## $counts
##  [1]   21   81  185  420  749 1073 1079  934  658  370   92   50   21    4
## [15]    1
##
## $density
##  [1] 0.0036598118 0.0141164169 0.0322411990 0.0731962356 0.1305332869
##  [6] 0.1869989543 0.1880446148 0.1627744859 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
```

- 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.long, probability = TRUE, breaks = 32.5:48.5)
```

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

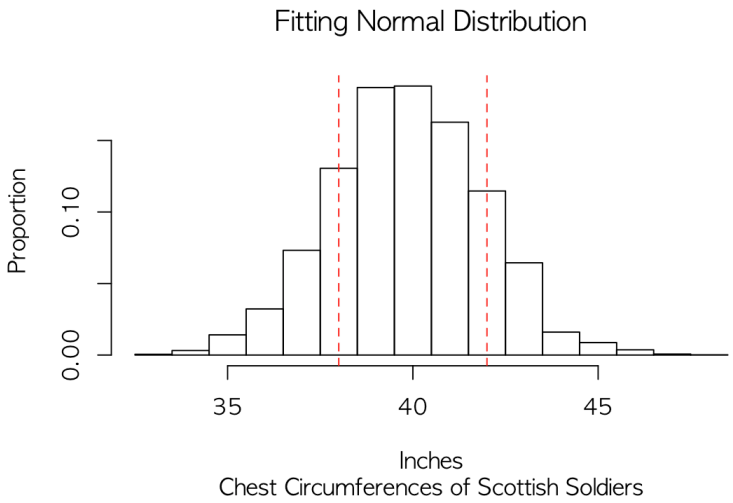
```
main.title <- "Fitting Normal Distribution"
sub.title <- "Chest Circumferences of Scottish Soldiers"
x.lab <- "Inches"
y.lab <- "Proportion"
hist(chest.long, breaks = 32.5:48.5, probability = TRUE, main = main.title, sub
= sub.title, xlab = x.lab, ylab = y.lab)
```



Mean ± SD contains 2/3 of total number of counts

- 평균을 중심으로 ±표준편차 만큼 떨어진 자료를 붉은 색 수직점선으로 표시.

```
hist(chest.long, breaks = 32.5:48.5, probability = TRUE, 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.2 <- hist(chest.long, breaks = 32.5:48.5, plot = FALSE)
h.chest.2
```

```
## $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.long"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

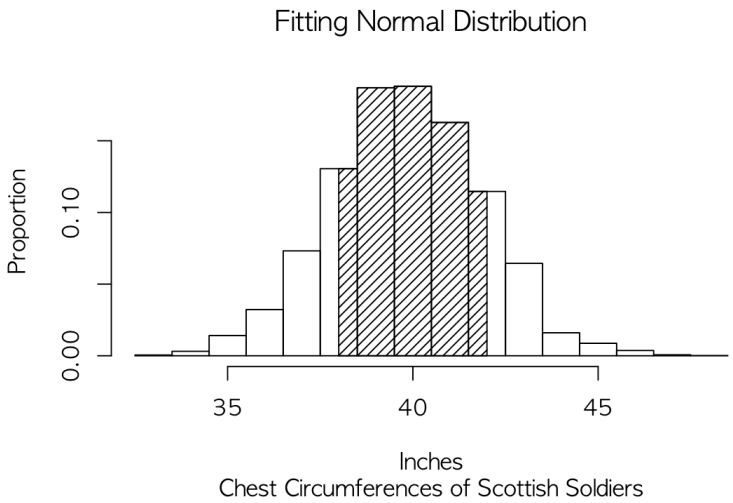
```
h.chest.2$density[6:10]
```

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

```
y <- h.chest.2$density[6:10]
```

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

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



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

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

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

```
y
```

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

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

```
## [1]  4.0  0.0 -0.5  0.0 -1.0
```

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

```
## [1]  0.52213315  0.00000000 -0.09402231  0.00000000 -0.11467410
```

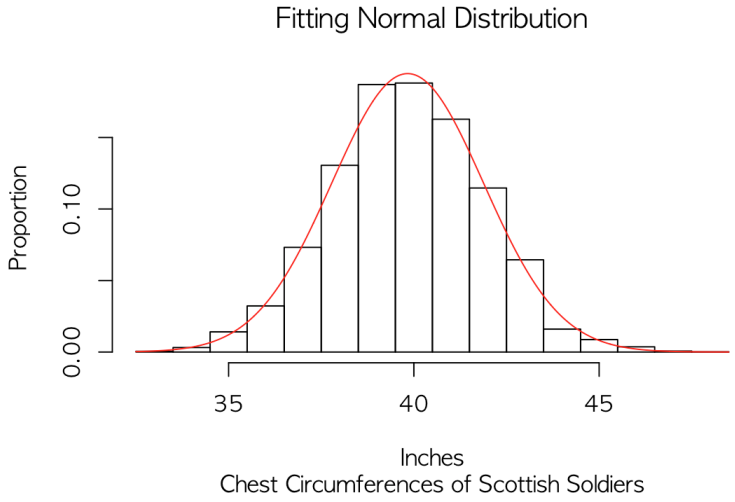
```
sum(diff(x.coord[1:6])*y)
```

```
## [1] 0.3134367
```

Comparison with normal curve

- 이론적인 정규분포 밀도함수 곡선을 히스토그램에 덧붙여 그림.

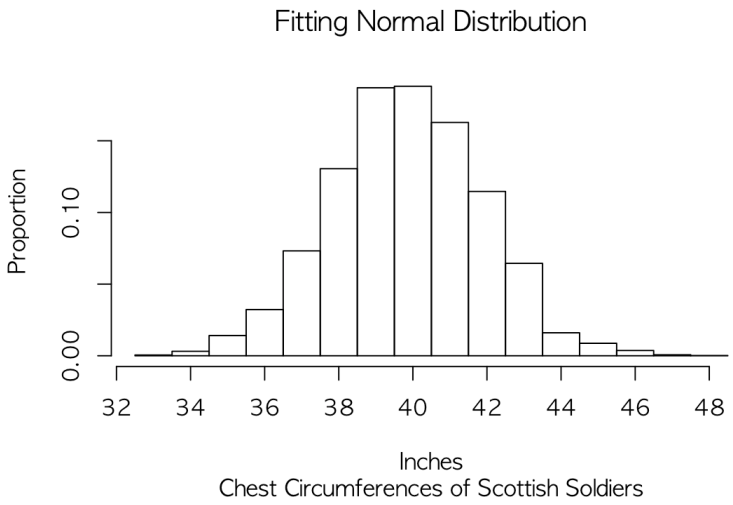
```
mean.chest <- mean(chest.long)
sd.chest <- sd(chest.long)
x <- seq(32.5, 48.5, length = 1000)
y.norm <- dnorm(x, mean = mean.chest, sd = sd.chest)
hist(chest.long, breaks = 32.5:48.5, probability = TRUE, main = main.title, sub
= sub.title, xlab = x.lab, ylab = y.lab)
lines(x, y.norm, col = "red")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

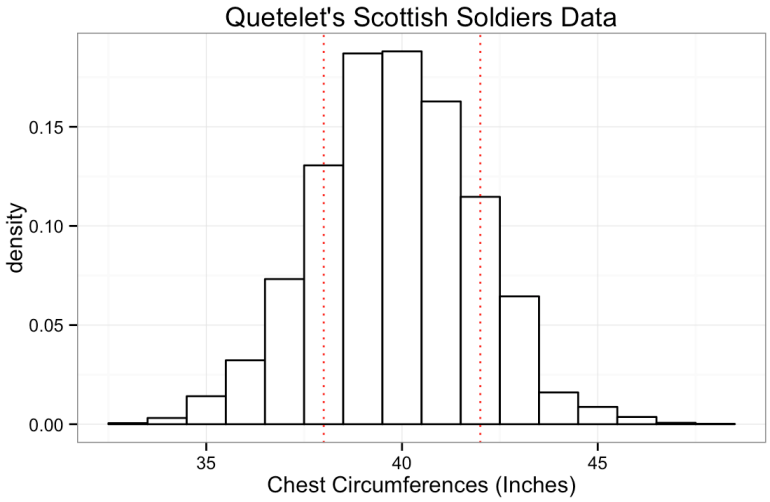


Changing tick marks of x axis

- default로 주어지는 x축의 눈금을 제대로 볼 수 있게 고치려면,

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





ggplot

- data frame으로 변환하여 접근

```
library(ggplot2)
ggplot(data.frame(chest.long), aes(x = chest.long)) + geom_histogram(aes(y = ..
density..), binwidth = 1, breaks = 32.5:48.5, fill = "white", colour = "black")
+
  geom_vline(xintercept = c(38, 42), linetype = "dotted", colour = "red") +
  theme_bw() + xlab("Chest Circumferences (Inches)") + ggtitle("Quetelet's Scot
tish Soldiers Data")
```