````
---
title: "Fitting Normal Distribution"
author: "coop711"
date: "`r Sys.Date()`"
output: html_document
---

## Data

### From Stigler's

```{r, packages, echo = FALSE}
# install.packages(c("pander", "ggplot2"), repos = "https://cran.rstudio.com")
library(pander)
library(knitr)
# search()
```

<img src = "../pics/quetelet_soldiers.png" width = "480"/>

<!--![Quetelet's frequency table](../pics/quetelet_soldiers.png)-->


### Frequency Table

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

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

#### Extract Parts of an Object

```{r, extract parts }
chest.table$Freq
str(chest.table$Freq)
chest.table[, 2]
str(chest.table[, 2])
chest.table[, "Freq"]
str(chest.table[, "Freq"])
chest.table["Freq"]
str(chest.table["Freq"])
chest.table["Freq"]$Freq
str(chest.table["Freq"]$Freq)
chest.table["Freq"][[1]]
str(chest.table["Freq"][[1]])
chest.table[2]
str(chest.table[2])
chest.table[2]$Freq
str(chest.table[2]$Freq)
chest.table[2][[1]]
str(chest.table[2][[1]])
chest.table[[2]]
str(chest.table[[2]])
```


* 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, `r sum(chest.table$Freq)`명으로 나눠준 값임.

```{r, barplot first, fig.width = 6, fig.height = 4}
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,

```{r, barplot white, fig.width=6, fig.height = 4}
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,

```{r, long format data frame}
chest.long <- rep(chest.table$Chest, chest.table$Freq)
str(chest.long)
```

`rep()`

```{r, rep()}
rep(1:3, 3)
rep(1:3, each = 3)
rep(1:3, 1:3)
```

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

```{r, basic statistics}
summary(chest.long)
sd(chest.long)
```

Histogram

* 히스토그램을 직관적으로 그려보면 y축은 돗수가 기본값임을 알 수 있음.

```{r, frequency histogram, fig.width = 6, fig.height = 4}
hist(chest.long)
```

* 정규분포와 비교하기 위해서 y축을 확률로 나타내려면

```{r, probability histogram, fig.width = 6, fig.height = 4}
hist(chest.long, probability = TRUE)
```

Inside the histogram

* 실제로 이 히스토그램을 그리는 데 계산된 값들은?

```{r, histogram objects}
(h.chest <- hist(chest.long, plot = FALSE))
list(breaks = h.chest$breaks, counts = h.chest$counts, density = h.chest$density, mids = h.chest$mids)
```

* 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

```{r, move 0.5 inches, fig.width = 6, fig.height = 4}
hist(chest.long, probability = TRUE, breaks = 32.5:48.5)
```

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

```{r, annotations, fig.width = 6, fig.height = 4}
main.title <- "Fitting Normal Distribution"
# sub.title <- "Chest Circumferences of Scottish Soldiers"
sub.title <- ""
x.lab <- "Chest Circumferences (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 $\pm$ SD contains 2/3 of total number of counts

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

```{r, mean and sd, fig.width = 6, fig.height = 4}
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

```{r, coordinates of polygon}
h.chest.2 <- hist(chest.long, breaks = 32.5:48.5, plot = FALSE)
h.chest.2
h.chest.2$density[6:10]
y <- h.chest.2$density[6:10]
```

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

```{r, 5 rectangles, fig.width = 6, fig.height = 4}
mean.chest <- mean(chest.long)
sd.chest <- sd(chest.long)
x.lower <- mean.chest - sd.chest
x.upper <- mean.chest + sd.chest
x.coord<-rep(c(x.lower, 38.5:41.5, x.upper), each = 2)
y.coord<-c(0, rep(y, each = 2), 0)
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(x.lower, x.upper), lty = 2, col = "red")
polygon(x.coord, y.coord, density = 20)
```

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

```{r, area of shaded area}
options(digits = 2)
x.area <- c(x.lower, 38.5:41.5, x.upper)
y
diff(x.area)
diff(x.area) * y
sum(diff(x.area) * y)
```

Comparison with normal curve

* 이론적인 정규분포 밀도함수 곡선을 히스토그램에 덧붙여 그림.

```{r, normal curve added, fig.width = 6, fig.height = 4}
x.chest <- seq(32.5, 48.5, length = 1000)
y.norm <- dnorm(x.chest, 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)
abline(v = c(x.lower, x.upper), lty = 2, col = "red")
# abline(v = c(38, 42), lty = 2, col = "red")
polygon(x.coord, y.coord, density = 20)
lines(x.chest, y.norm, col = "red")
```

Changing tick marks of x axis

* default로 주어지는 x축의 눈금을 제대로 볼 수 있게 고치려면,

```{r, x axis, fig.width = 6, fig.height = 4}
hist(chest.long, breaks = 32.5:48.5, probability = TRUE, main = main.title, sub =
sub.title, xlab = x.lab, ylab = y.lab, axes = FALSE)
abline(v = c(x.lower, x.upper), lty = 2, col = "red")
polygon(x.coord, y.coord, density = 20)
lines(x.chest, y.norm, col = "red")
axis(side = 1, at = seq(32, 48, by = 2), labels = seq(32, 48, by = 2))
axis(side = 2)
```

ggplot

* data frame으로 작업.

Basic histogram

```{r, ggplots, fig.width = 6, fig.height = 4}
library(ggplot2)
(g1 <- 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"))
```

Mean $\pm$ SD

```{r, mean plus minus sd, , fig.width = 6, fig.height = 4}
(g2 <- g1 + geom_vline(xintercept = c(x.lower, x.upper), linetype = "dashed", colour =
"black"))
```

x-axis label and main title

```{r, xlab and ggtitle, fig.width = 6, fig.height = 4}
(g3 <- g2 + theme_bw() + xlab("Chest Circumferences (Inches)") + ggtitle("Quetelet's
Scottish Soldiers Data"))
```

Shading the area

```{r, polygon, fig.width = 6, fig.height = 4}
(g4 <- g3 + geom_polygon(aes(x = x.coord, y = y.coord), data = data.frame(x.coord,
y.coord), alpha = 0.5, fill = "grey"))
```

Normal curve added

```{r, normal curve, fig.width = 6, fig.height = 4}
x.curve <- seq(32.5, 48.5, length = 100)
y.curve <- dnorm(x.curve, mean = mean.chest, sd = sd.chest)
(g5 <- g4 + geom_line(aes(x = x.curve, y = y.curve), data = data.frame(x.curve, y.curve),
colour = "blue"))
```

x-axis tick marks

```{r, tick marks, fig.width = 6, fig.height = 4}
(g6 <- g5 + scale_x_continuous(breaks = seq(32, 48, by = 2), labels = seq(32, 48, by =
2)))
```

```
```