Cushny and Peeble's Data

coop711

2018-03-31

Data Management

Data

R-base에서 제공하고 있는 `sleep` data 는 long form data frame 으로 주어져 있음.

```
library(knitr)
sleep
```

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

```
str(sleep)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ extra: num  0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2 ...
##  $ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ID   : Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

Long Form vs Wide Form

long form을 wide form으로 변환하고, 각각의 경우에 적절한 t-test를 시도해 볼 것임. 먼저 wide form 으로 변환하는 작업은 결국 data frame을 새로 구성하는 것일 뿐이므로 다음으로 완료됨.

```
sleep_wide <- data.frame(A = sleep[sleep$group == 1, 1],
                         B = sleep[sleep$group == 2, 1])
sleep_wide
```

```
##       A    B
## 1   0.7  1.9
## 2  -1.6  0.8
## 3  -0.2  1.1
## 4  -1.2  0.1
## 5  -0.1 -0.1
## 6   3.4  4.4
## 7   3.7  5.5
## 8   0.8  1.6
## 9   0.0  4.6
## 10  2.0  3.4
```

```
str(sleep_wide)
```

```
## 'data.frame':    10 obs. of  2 variables:
##  $ A: num  0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2
##  $ B: num  1.9 0.8 1.1 0.1 -0.1 4.4 5.5 1.6 4.6 3.4
```

```
sleep_wide$A
```

```
##  [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

```
sleep_wide[, "A"]
```

```
##  [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

```
sleep_wide[, 1]
```

```
##  [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

One Sample T test

Long Form Data Frame

long form 에서 각 수면제의 효과가 없다는 가설을 t-test 하려면

One sided t-test

```
t.test(sleep$extra[sleep$group == 1],
       alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  sleep$extra[sleep$group == 1]
## t = 1.3257, df = 9, p-value = 0.1088
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2870553        Inf
## sample estimates:
## mean of x
##      0.75
```

```
t.test(sleep$extra[sleep$group == 2],
       alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  sleep$extra[sleep$group == 2]
## t = 3.6799, df = 9, p-value = 0.002538
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.169334      Inf
## sample estimates:
## mean of x
##      2.33
```

`tapply()`

둘을 단번에 수행하려면 `tapply()` 를 이용하여

```
tapply(sleep$extra,
       INDEX = sleep$group,
       FUN = t.test, alternative = "greater")
```

```
## $`1`
##
##  One Sample t-test
##
## data:  X[[i]]
## t = 1.3257, df = 9, p-value = 0.1088
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2870553        Inf
## sample estimates:
## mean of x
##      0.75
##
##
## $`2`
##
##  One Sample t-test
##
## data:  X[[i]]
## t = 3.6799, df = 9, p-value = 0.002538
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.169334      Inf
## sample estimates:
## mean of x
##      2.33
```

Paired t-test

두 수면제 간의 효과에 차이가 없다는 가설을 검증하려면, paired 임을 유념하여야 함.

```
t.test(sleep$extra[sleep$group == 1], sleep$extra[sleep$group == 2],
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sleep$extra[sleep$group == 1] and sleep$extra[sleep$group == 2]
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

Formula form

formula 형식을 빌리면 다음과 같이 비교적 간결하게 기술할 수 있음.

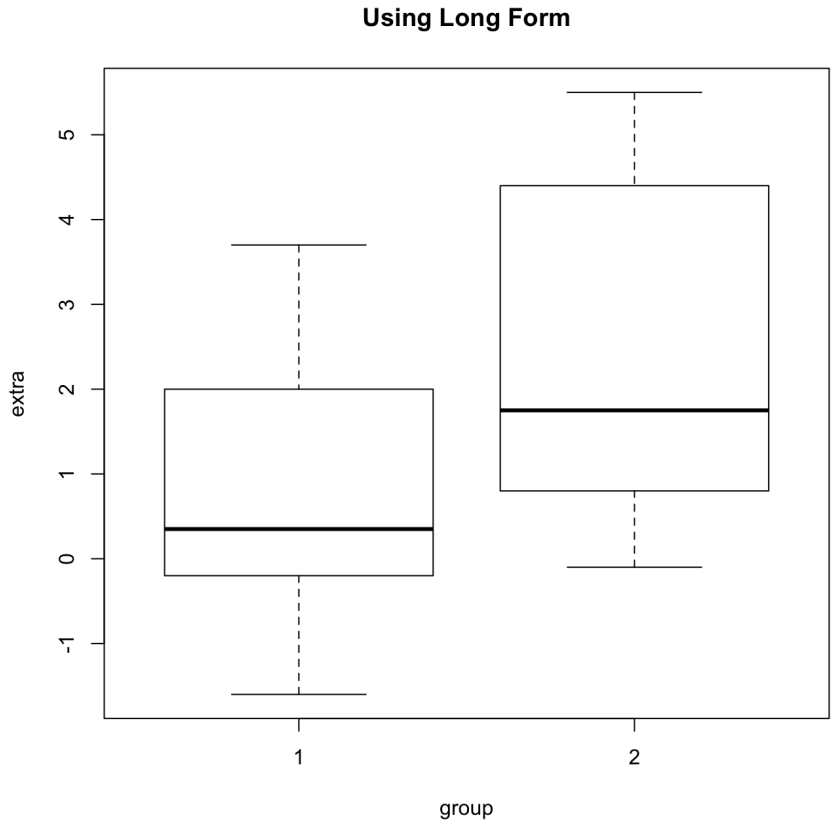
```
t.test(extra ~ group,
       data = sleep,
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  extra by group
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

Boxplot

두 수면제의 효과를 boxplot을 그려 비교하면(산점도를 그려 비교하려면 어떻게?)

```
plot(extra ~ group,
     data = sleep,
     main = "Using Long Form")
```



Wide Form Data Frame

wide form 으로 같은 작업을 수행하면

t.test()

```
t.test(sleep_wide$A,
       alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  sleep_wide$A
## t = 1.3257, df = 9, p-value = 0.1088
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2870553        Inf
## sample estimates:
## mean of x
##      0.75
```

```
t.test(sleep_wide$B,
       alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  sleep_wide$B
## t = 3.6799, df = 9, p-value = 0.002538
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.169334      Inf
## sample estimates:
## mean of x
##      2.33
```

apply()

`apply()` 를 이용해서 한번에 수행하면

```
apply(sleep_wide,
      MARGIN = 2,
      FUN = t.test, alternative="greater")
```

```
## $A
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 1.3257, df = 9, p-value = 0.1088
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -0.2870553        Inf
## sample estimates:
## mean of x
##      0.75
##
##
## $B
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 3.6799, df = 9, p-value = 0.002538
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.169334      Inf
## sample estimates:
## mean of x
##      2.33
```

Paired t-test

두 수면제 간의 효과 차이를 검증하려면

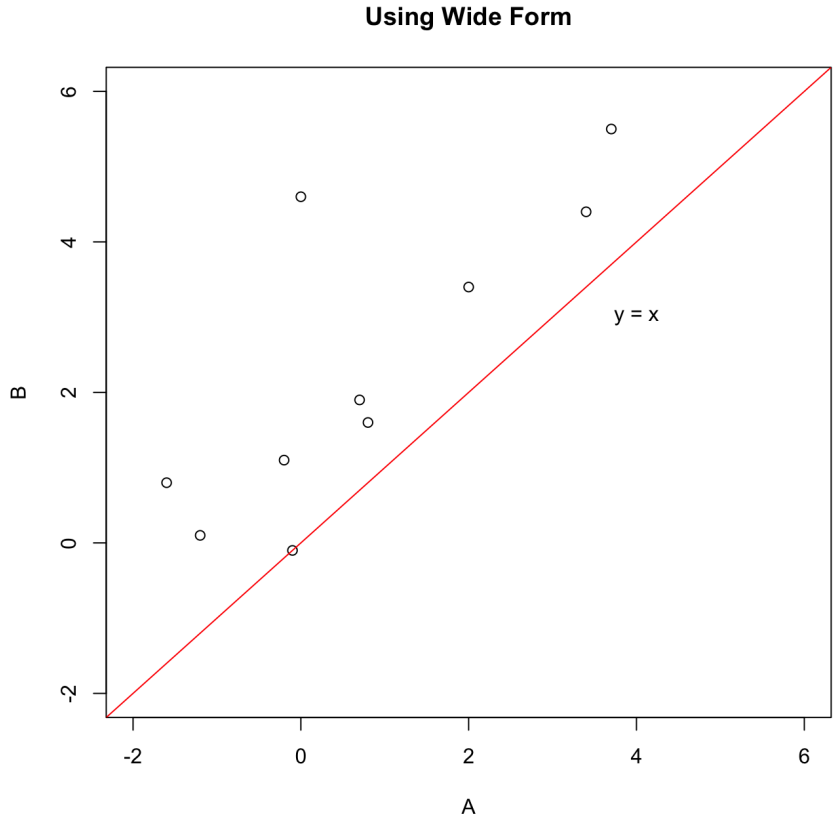
```
t.test(sleep_wide$A, sleep_wide$B,
       paired = T)
```

```
##
##  Paired t-test
##
## data:  sleep_wide$A and sleep_wide$B
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

Scatter Diagram

각각의 효과를 산점도를 그려 비교하면

```
plot(sleep_wide,
     main = "Using Wide Form",
     xlim = c(-2, 6),
     ylim = c(-2, 6))
abline(a = 0, b = 1,
       col = "red")
text(x = 4, y = 3,
     labels = "y = x")
```



상관계수

```
cor(sleep_wide$A, sleep_wide$B)
```

```
## [1] 0.7951702
```

Tests of Normality

정규성에 대한 검증은 각자 수행해 볼 것.

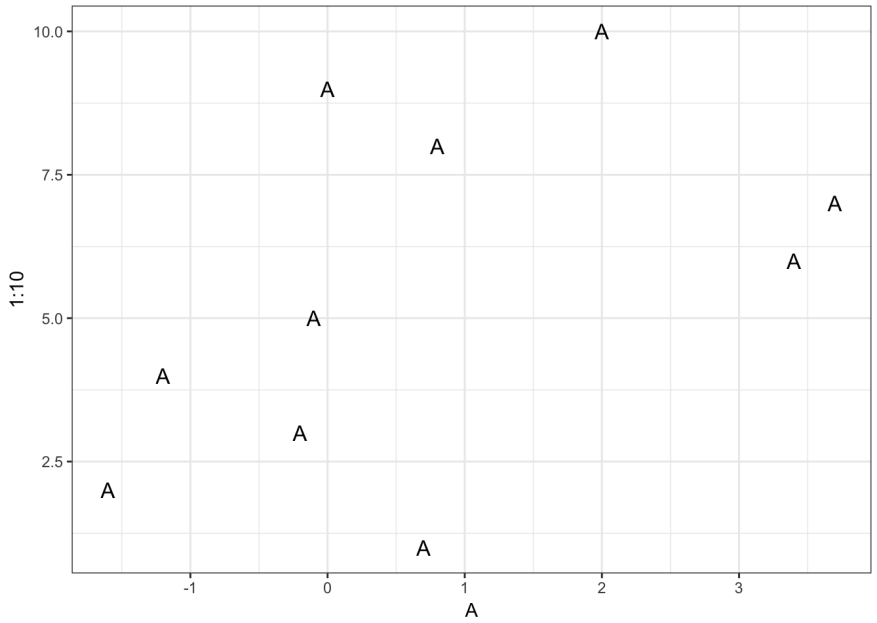
```
library(nortest)
kable(sapply(sleep_wide, ad.test))
```

	A	B
statistic	0.346906730810753	0.357157083362328
p.value	0.401927819514199	0.378470722436255
method	Anderson-Darling normality test	Anderson-Darling normality test
data.name	X[[i]]	X[[i]]

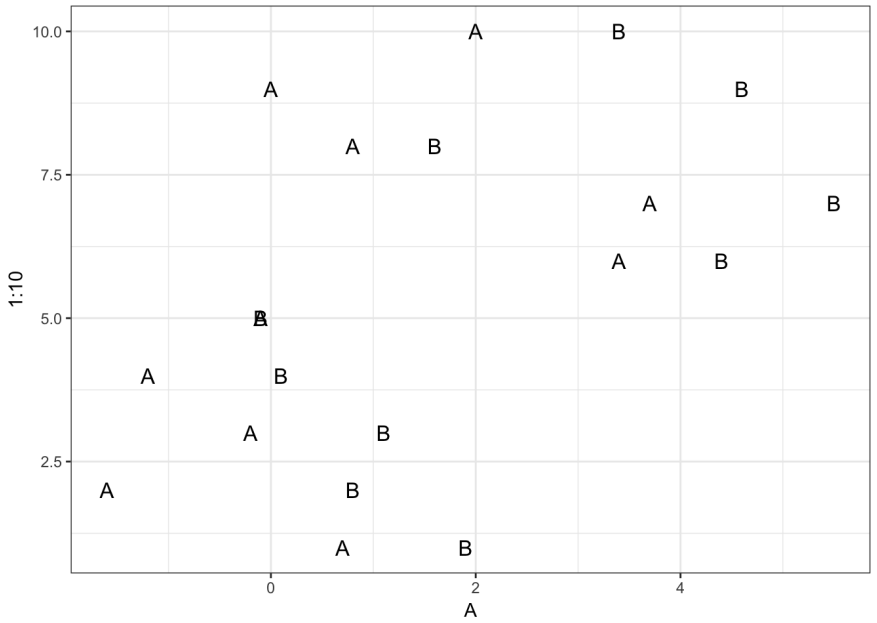
Dot Plot (ggplot)

```
library(ggplot2)
library(grid)
(g1 <- ggplot(data = sleep_wide) +
  theme_bw())
```

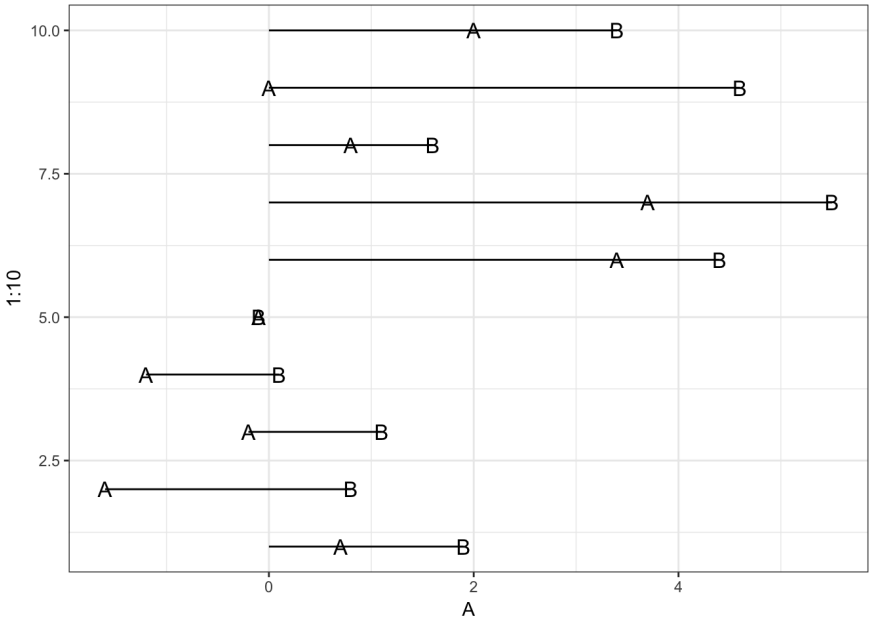
```
(g2 <- g1 +
    geom_point(mapping = aes(x = A, y = 1:10),
               shape = "A",
               size = 4))
```



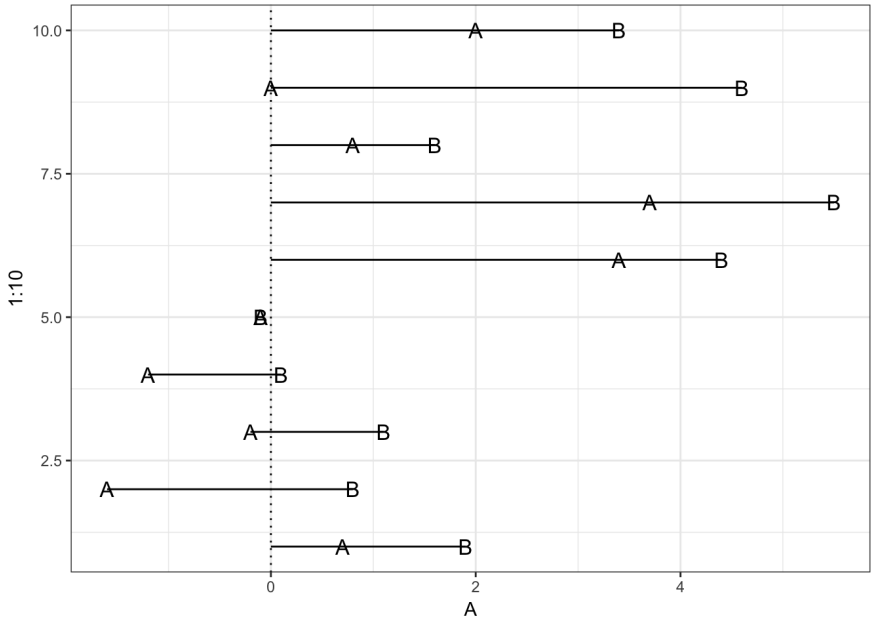

```
(g3 <- g2 +
    geom_point(mapping = aes(x = B, y = 1:10),
               shape = "B",
               size = 4))
```



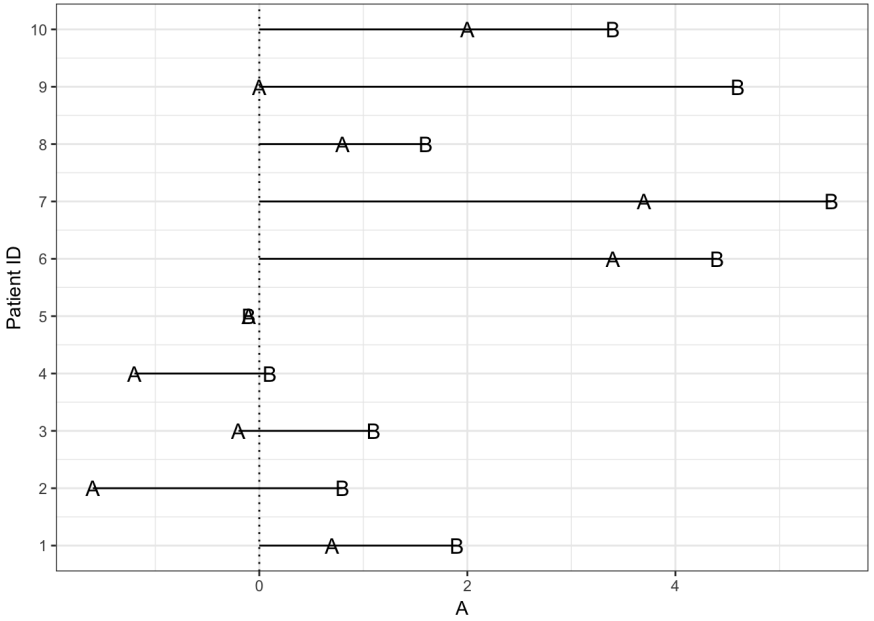
```
(g4 <- g3 +
    geom_segment(mapping = aes(x = ifelse(A >= 0, 0, B),
                               y = 1:10,
                               xend = ifelse(A >= 0, B, A),
                               yend = 1:10),
                 size = 0.5,
                 linetype = 1))
```



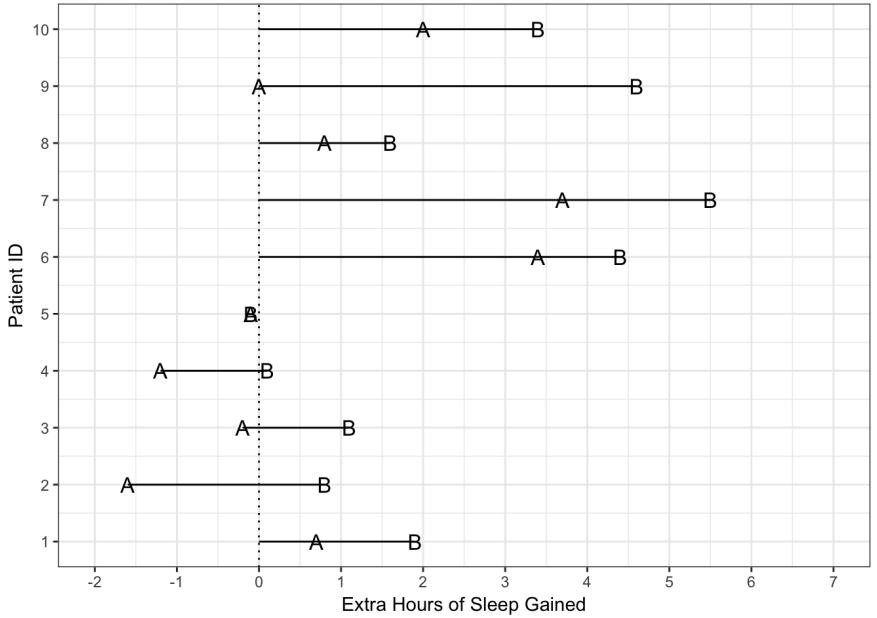

```
(g5 <- g4 +
    geom_vline(xintercept = 0,
               linetype = 3))
```



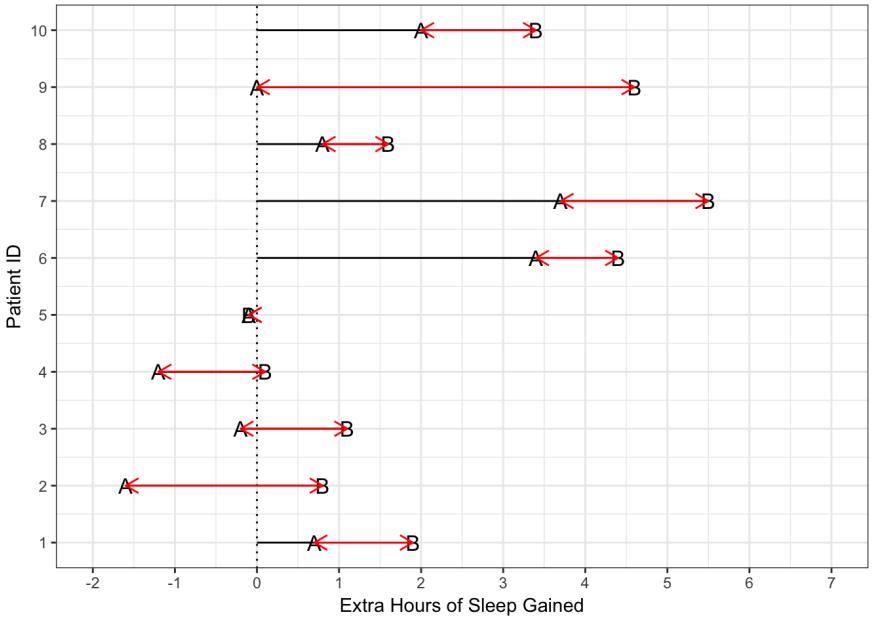
```
(g6 <- g5 +
    scale_y_continuous(name = "Patient ID",
                       breaks = 1:10,
                       labels = 1:10))
```



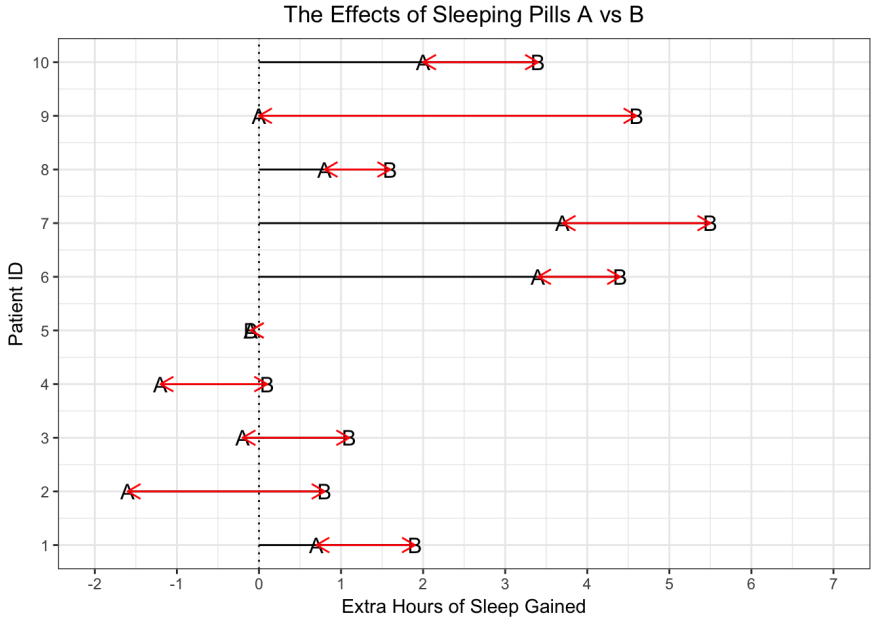

```
(g7 <- g6 +
    scale_x_continuous(name = "Extra Hours of Sleep Gained",
                       breaks = -2:7,
                       labels = -2:7,
                       limits = c(-2, 7)))
```



```
(g8 <- g7 +
    annotate("segment",
             x = sleep_wide$A,
             xend = sleep_wide$B,
             y = 1:10,
             yend = 1:10,
             col = "red",
             size = 0.5,
             arrow = arrow(length = unit(0.3, "cm"),
                           ends = "both")))
```




```
(g9 <- g8 +
    ggtitle("The Effects of Sleeping Pills A vs B") +
    theme(plot.title = element_text(hjust = 0.5)))
```



작업 디렉토리 정리

```
saveRDS(sleep_wide,
        file = "sleep_wide.RDS")
```