Midterm 2

Ly Nguyen

2023-11-07

```r
bcar <- read.csv("~/Documents/Personal Docs_East Bay/STAT 630/bcar.csv")
View(bcar)
```

Part 1: Data Cleaning

```r
bcar <- bcar %>%
  mutate(month = factor(month),
         dose = factor(dose),
         sex_f = factor(sex,
                        labels = c("Female", "Male"))
)
summary(bcar)
```

```
##       ptid        month       bcarot            vite        dose
##  Min.   : 1.00   0 :38   Min.   :  49.0   Min.   : 3.160   0 :16
##  1st Qu.:11.00   11:38   1st Qu.: 182.8   1st Qu.: 6.048   15:14
##  Median :24.00           Median : 349.0   Median : 7.615   30:16
##  Mean   :24.97           Mean   : 828.7   Mean   : 7.455   45:12
##  3rd Qu.:36.00           3rd Qu.:1499.0   3rd Qu.: 8.648   60:18
##  Max.   :57.00           Max.   :3489.0   Max.   :12.020
##       age             sex              bmi             chol          sex_f
##  Min.   :50.00   Min.   :0.0000   Min.   :19.68   Min.   :159.0   Female:36
##  1st Qu.:52.00   1st Qu.:0.0000   1st Qu.:23.18   1st Qu.:202.0   Male  :40
##  Median :56.00   Median :1.0000   Median :25.24   Median :216.0
##  Mean   :56.24   Mean   :0.5263   Mean   :25.58   Mean   :218.6
##  3rd Qu.:60.00   3rd Qu.:1.0000   3rd Qu.:27.63   3rd Qu.:236.5
##  Max.   :64.00   Max.   :1.0000   Max.   :31.68   Max.   :312.5
```

Part 2: Exploratory Data Analysis

2.

```r
#Calculate n(%) of sex

table(bcar$sex_f)
```

```
##
## Female   Male
##     36     40
```

```r
props <- prop.table(table(bcar$sex_f))
props
```

```
##
##    Female      Male
## 0.4736842 0.5263158
```

```r
#Calculate mean (sd) of bmi

mean_bmi <- round(mean(bcar$bmi),2)
mean_bmi
```

```
## [1] 25.58
```

```r
sd_bmi <- round(sd(bcar$bmi),2)
sd_bmi
```

```
## [1] 2.98
```

```r
#Calculate mean (sd) of chol

mean_chol <- round(mean(bcar$chol),2)
mean_chol
```

```
## [1] 218.57
```

```r
sd_chol <- round(sd(bcar$chol),2)
sd_chol
```

```
## [1] 31.25
```

```r
#Calculate n(%) of dose

table(bcar$dose)
```

```
##
##  0 15 30 45 60
## 16 14 16 12 18
```

```r
props_1 <- round((prop.table(table(bcar$dose))),2)
props_1
```

```
##
##    0   15   30   45   60
## 0.21 0.18 0.21 0.16 0.24
```

```r
#Calculate mean (sd) of age

mean_age <- round(mean(bcar$age),2)
mean_age
```

```
## [1] 56.24
```

```r
sd_age <- round(sd(bcar$age),2)
sd_age
```

```
## [1] 4.13
```

Variable	mean(sd) or n (%)
sex	Female: 36 (47.37%)
	Male: 40 (52.63%)
bmi	mean: 25.58 (sd: 2.98)
chol	mean: 218.57 (sd: 31.25)
dose	0 mg/day: 16 (21%)
	15 mg/day: 14 (18%)
	30 mg/day: 16 (21%)

Variable	mean(sd) or n (%)
	45 mg/day: 12 (16%)
	60 mg/day: 18 (24%)
age	mean: 56.24 (sd: 4.13)

3.

```
# calculate the mean and standard deviation of vitamin E (vite) stratified by month
vite_month0 <- bcar %>%
  filter(month == 0) %>%
  select(vite) %>%
  pull()


mean_vite_month0 <- mean(vite_month0)
mean_vite_month0
```

```
## [1] 8.221053
```

```
sd_vite_month0 <- sd(vite_month0)
sd_vite_month0
```

```
## [1] 1.478037
```

```
vite_month11 <- bcar %>%
  filter(month == 11) %>%
  select(vite) %>%
  pull()

mean_vite_month11 <- mean(vite_month11)
mean_vite_month11
```

```
## [1] 6.688684
```

```
sd_vite_month11 <- sd(vite_month11)
sd_vite_month11
```

```
## [1] 1.556867
```

```
# calculate the mean and standard deviation of beta-carotene (bcarot) stratified by month
bcarot_month0 <- bcar %>%
  filter(month == 0) %>%
  select(bcarot) %>%
  pull()


mean_bcarot_month0 <- mean(bcarot_month0)
mean_bcarot_month0
```

```
## [1] 240.6579
```

```
sd_bcarot_month0 <- sd(bcarot_month0)
sd_bcarot_month0
```

```
## [1] 126.479
```

```
bcarot_month11 <- bcar %>%
  filter(month == 11) %>%
```

```
  select(bcarot) %>%
  pull()

mean_bcarot_month11 <- mean(bcarot_month11)
mean_bcarot_month11
```

```
## [1] 1416.711
```

```
sd_bcarot_month11 <- sd(bcarot_month11)
sd_bcarot_month11
```

```
## [1] 908.1557
```

Variable	Beginning of the study	end of the study
	mean(sd)	mean(sd)
Vite	mean: 8.2210526 (sd: 1.478037)	mean: 6.6886842 (sd: 1.5568665)
bcarot	mean: 240.6578947 (sd: 126.4789858)	mean: 1416.7105263 (sd: 908.155687)

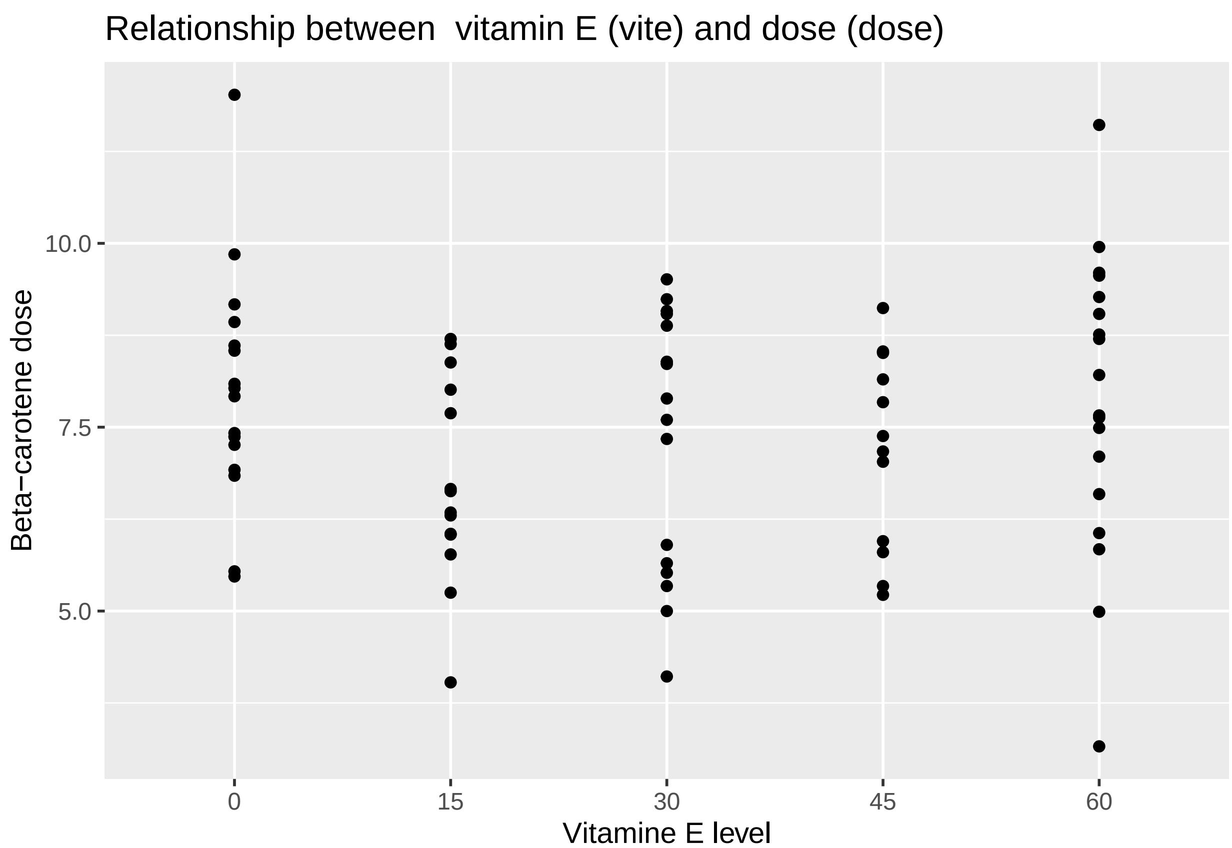
4.

The mean of vitamin E at the beginning of the study was more than that one at the ending. The sd of the vitamin E was quite the same in both of two period, so the variability stays unchanged. The average bcarot at the beginning was just 240, but at the ending it significantly came up to 1416. The sd of bcarot also went higher, so it shows that the variability of bcarot expanded more.

6. Compare the shape, center, and spread of both histograms in 2-3 complete sentences. Use the statistics you calculated in Question 3 to add to your comparison.

7. Using a plot of your choice, visualize the relationship between vitamin E (vite) and dose (dose).

```
ggplot(bcar, aes(x = dose, y = vite)) +
  geom_point() +
  labs(title = "Relationship between  vitamin E (vite) and dose (dose)",
       y = "Beta-carotene dose", x = "Vitamine E level")
```



Part 2: Data Analysis

```
bcar_wide <- bcar %>%
tidyr::pivot_wider(names_from = month,
values_from = c(bcarot, vite))
```

8. Create a 95% confidence interval for the true average difference in vitamin E level in month 11 (vite_11) minus month 0 (vite_0). Check that the necessary conditions are satisfied, compute the interval in R, and then interpret the interval in the context of the problem. Next, we are going to categorize vitamin E into low or high.

```
bcar_wide$vite_low_0 <- ifelse(bcar_wide$vite_0 < 6, "low vit E", "high Vit E")

bcar_wide$vite_low_11 <- ifelse(bcar_wide$vite_11 < 6, "low vit E", "high Vit E")

bcar_wide <- bcar_wide %>%
mutate_if(is.character, as.factor)

bcar_wide <- bcar_wide %>%
  mutate(vite_diff = vite_11 - vite_0)

mean_vite_diff <- mean(bcar_wide$vite_diff)
sd_vite_diff <- sd(bcar_wide$vite_diff)

# Calculate standard error
n <- length(bcar_wide$vite_diff)
se_vite_diff <- bcar_wide$sd_vite_diff / sqrt(n)
```

```r
# Set the confidence level
conf_level <- 0.95

# Calculate the critical t-value
t_stat <- qt((1 + 0.95) / 2, df = n - 1)

p_val <- pt(t_stat, df = n - 1, lower.tail = FALSE)
p_val
```

```
## [1] 0.025
```

```r
t_stat + c(-1, 1) * qt(0.975, df = n - 1) * se_vite_diff
```

```
## numeric(0)
```

Self - assessment: M (I try my best in the rush period of time, I am 90% confident that my result makes sense)