LAB5

Lubin

2024-11-01

Дисперсионный анализ

Загрузим данные

```
data <- read.csv("data/diet.csv", row.names=1)
summary(data)
```

```
##      gender            Age            Height        pre.weight
##  Min.   :0.0000   Min.   :16.00   Min.   :141.0   Min.   : 58.00
##  1st Qu.:0.0000   1st Qu.:32.25   1st Qu.:164.2   1st Qu.: 66.00
##  Median :0.0000   Median :39.00   Median :169.5   Median : 72.00
##  Mean   :0.4231   Mean   :39.15   Mean   :170.8   Mean   : 72.53
##  3rd Qu.:1.0000   3rd Qu.:46.75   3rd Qu.:174.8   3rd Qu.: 78.00
##  Max.   :1.0000   Max.   :60.00   Max.   :201.0   Max.   :103.00
##       Diet        weight6weeks
##  Min.   :1.000   Min.   : 53.00
##  1st Qu.:1.000   1st Qu.: 61.85
##  Median :2.000   Median : 68.95
##  Mean   :2.038   Mean   : 68.68
##  3rd Qu.:3.000   3rd Qu.: 73.83
##  Max.   :3.000   Max.   :103.00
```

Проведем некоторые преобразования над таблицей

```
colnames(data) <- c("gender", "age", "height", "initial.weight",
                    "diet.type", "final.weight")
data$diet.type <- factor(c("A", "B", "C")[data$diet.type])
data$weight.loss = data$initial.weight - data$final.weight
summary(data)
```

```
##      gender            age            height      initial.weight   diet.type
##  Min.   :0.0000   Min.   :16.00   Min.   :141.0   Min.   : 58.00   A:24
##  1st Qu.:0.0000   1st Qu.:32.25   1st Qu.:164.2   1st Qu.: 66.00   B:27
##  Median :0.0000   Median :39.00   Median :169.5   Median : 72.00   C:27
##  Mean   :0.4231   Mean   :39.15   Mean   :170.8   Mean   : 72.53
##  3rd Qu.:1.0000   3rd Qu.:46.75   3rd Qu.:174.8   3rd Qu.: 78.00
##  Max.   :1.0000   Max.   :60.00   Max.   :201.0   Max.   :103.00
##   final.weight     weight.loss
##  Min.   : 53.00   Min.   :-2.100
##  1st Qu.: 61.85   1st Qu.: 2.000
##  Median : 68.95   Median : 3.600
##  Mean   : 68.68   Mean   : 3.845
##  3rd Qu.: 73.83   3rd Qu.: 5.550
##  Max.   :103.00   Max.   : 9.200
```

Удаление выбросов

Сейчас нужно проверить данные в колонке “weight.loss” на наличие выбросов. Извлечем потенциальные выбросы на основе критерия IQR (межквартильного размаха), используя функцию “boxplot.stats()”.

```
boxplot.stats(data$weight.loss)$out
```

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

Если рассматривать данные целиком, то выбросов нет. Разделим данные на 3 группы по типу диеты и проверим на наличие выбросов каждую из них.

```
data.Adiet <- subset(data, diet.type == "A")
data.Bdiet <- subset(data, diet.type == "B")
data.Cdiet <- subset(data, diet.type == "C")

out.A <- boxplot.stats(data.Adiet$weight.loss)$out
out.B <- boxplot.stats(data.Bdiet$weight.loss)$out
out.C <- boxplot.stats(data.Cdiet$weight.loss)$out

out.A
```

```
## [1] 8.5 9.0
```

```
out.B
```

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

```
out.C
```

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

В итоге, потенциальные выбросы были обнаружены только в группе “A”. Избавимся от них.

```
data.new <- subset(data, weight.loss != 8.5 & weight.loss != 9.0)
boxplot.stats(subset(data.new, diet.type == "A")$weight.loss)$out
```

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

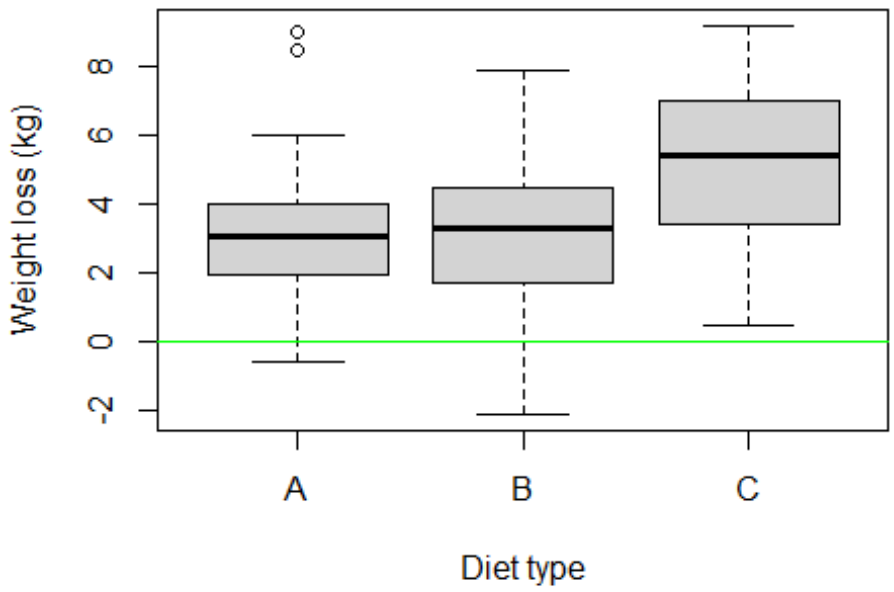
Тесты и построение графиков

Далее проведем все тесты, показанные в примере, и сравним результаты с выбросами и без.

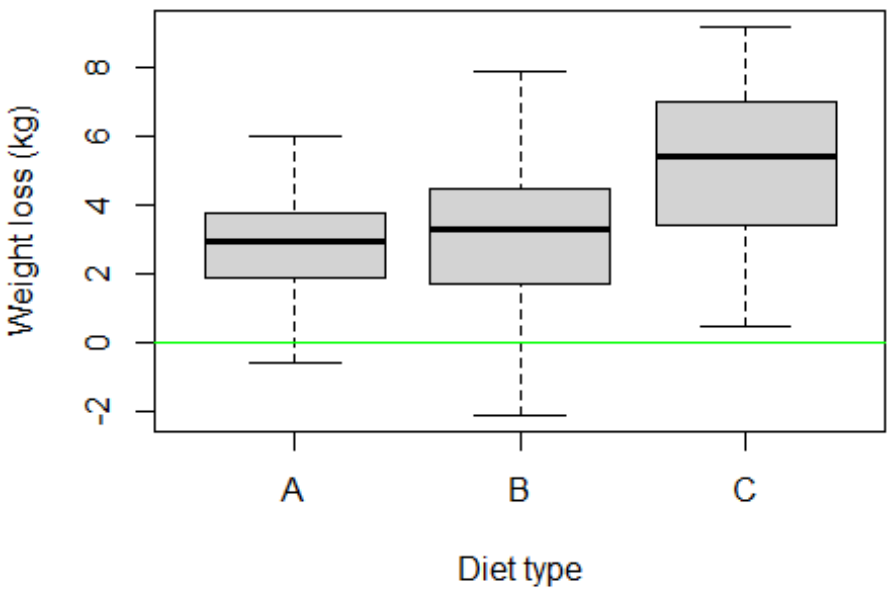
Диаграмма размаха

```
# С выбросами
boxplot(
  data=data,
  weight.loss~diet.type,
```

```
  col="light gray",
  ylab="Weight loss (kg)",
  xlab="Diet type"
)
abline(h=0, col="green")
```



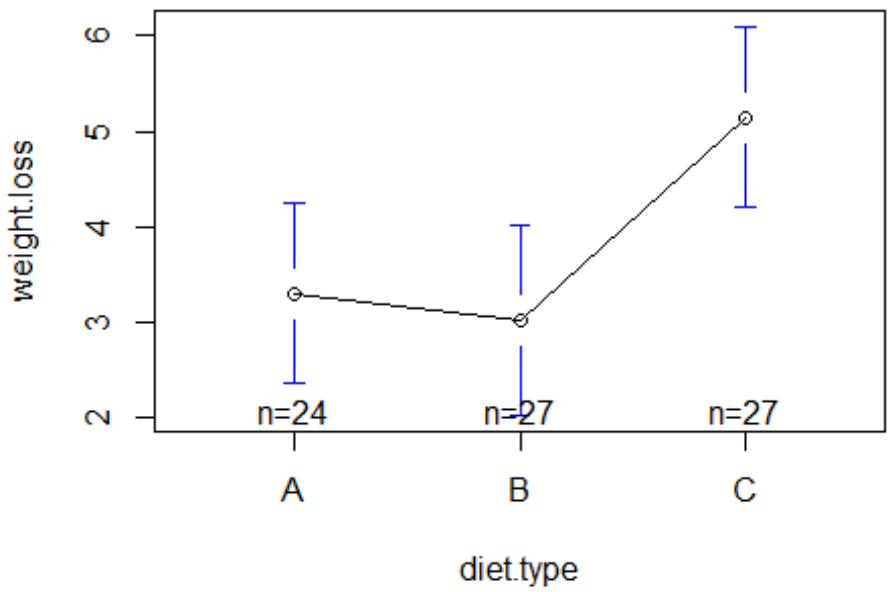

```
# Без выбросов
boxplot(
  data=data.new,
  weight.loss~diet.type,
  col="light gray",
  ylab="Weight loss (kg)",
  xlab="Diet type"
)
abline(h=0, col="green")
```



В случае без выбросов средняя потеря веса у людей, которые сидели на диете “A”, ниже чем в случае с выбросами. Также уменьшился верхний и нижний квантиль.

График групповых средних

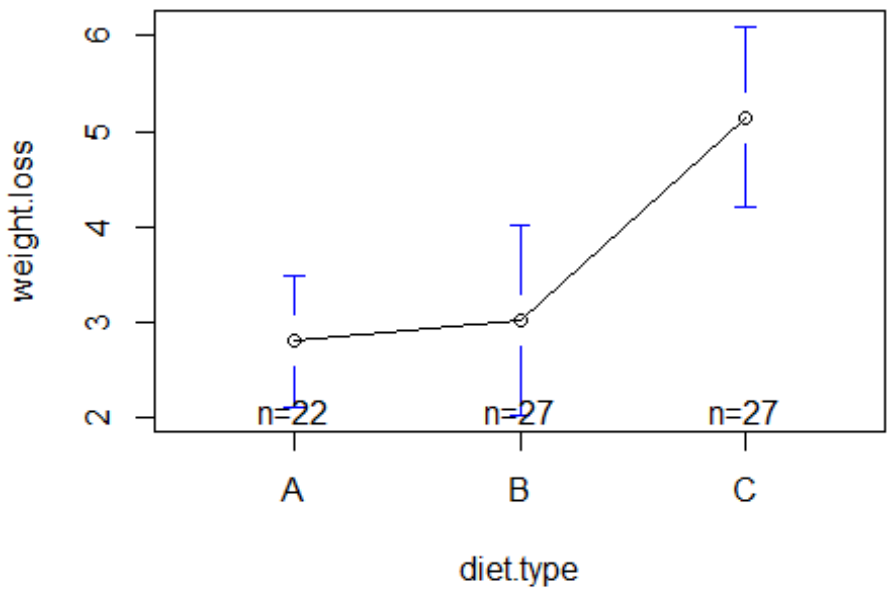
```
library(gplots)

# С выбросами
plotmeans(data=data, weight.loss~diet.type)
```




```
# Рассчет среднеквадратического отклонения для каждой группы
aggregate(data$weight.loss, by=list(data$diet.type), FUN=sd)
```

```
##   Group.1        x
## 1       A 2.240148
## 2       B 2.523367
## 3       C 2.395568
```

```
# Без выбросов
plotmeans(data=data.new, weight.loss~diet.type)
```




```
# Рассчет среднеквадратического отклонения для каждой группы
aggregate(data.new$weight.loss, by=list(data.new$diet.type), FUN=sd)
```

```
##   Group.1        x
## 1       A 1.550569
## 2       B 2.523367
## 3       C 2.395568
```

В случае без выбросов среднеквадратическое отклонение потери веса у людей, которые сидели на диете “A”, ниже чем в случае с выбросами.

Тест на межгрупповые различия

Для подгонки ANOVA модели используем функцию aov, частный случай линейной модели lm.

```
# С выбросами
fit <- aov(data=data, weight.loss ~ diet.type)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet.type    2   71.1   35.55   6.197 0.00323 **
## Residuals   75  430.2    5.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n\n")
```

```r
# Без выбросов
fit.new <- aov(data=data.new, weight.loss ~ diet.type)
summary(fit.new)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## diet.type    2   86.5   43.26   8.645 0.000427 ***
## Residuals   73  365.2    5.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

В случае без выбросов значение p-value меньше чем в случае с выбросами. Несмотря на это, в обоих случаях результат является статистически значимым. Это указывает на то, что по крайней мере одна группа отличается от других.

Попарные различия между средними значениями для всех групп

```r
# С выбросами
TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.loss ~ diet.type, data = data)
##
## $diet.type
##           diff        lwr      upr     p adj
## B-A -0.2740741 -1.8806155 1.332467 0.9124737
## C-A  1.8481481  0.2416067 3.454690 0.0201413
## C-B  2.1222222  0.5636481 3.680796 0.0047819
```

```r
cat("\n\n")
```

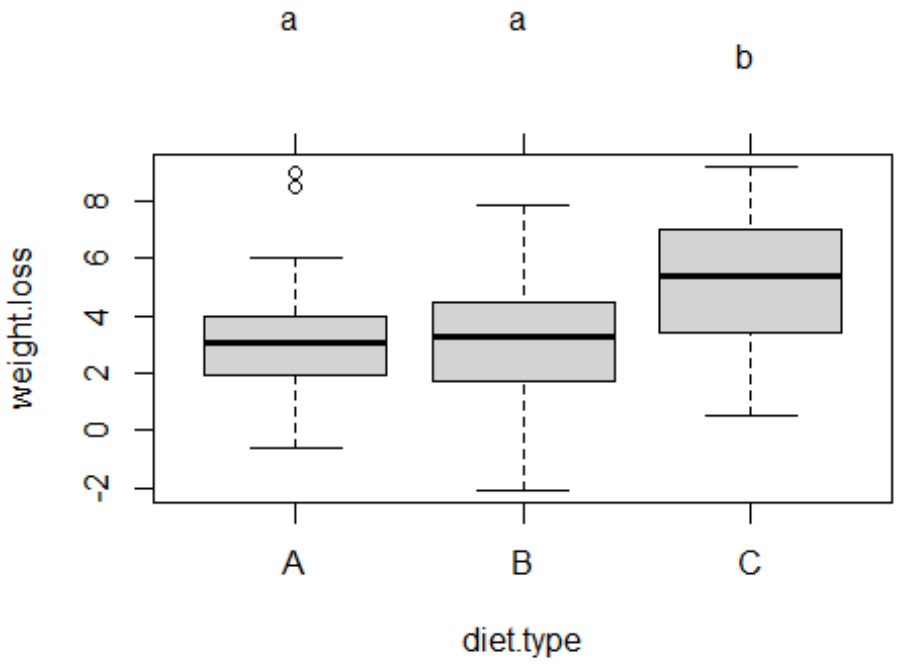
```r
# Без выбросов
TukeyHSD(fit.new)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight.loss ~ diet.type, data = data.new)
##
## $diet.type
##          diff        lwr      upr     p adj
## B-A 0.2213805 -1.3156340 1.758395 0.9367018
## C-A 2.3436027  0.8065882 3.880617 0.0014162
## C-B 2.1222222  0.6657364 3.578708 0.0023769
```

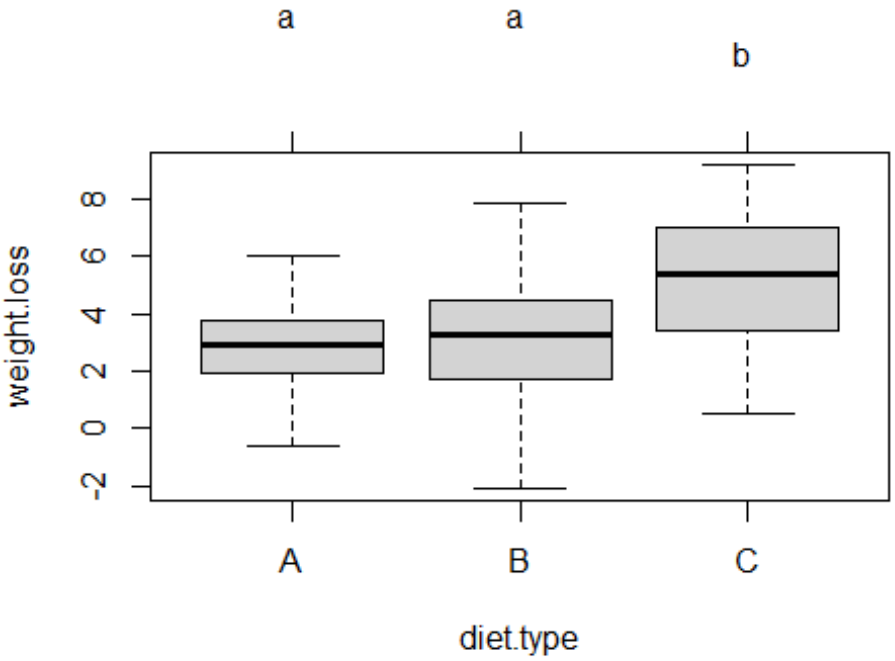
В случае без выбросов разница между группами A и B, A и C оказалась выше, чем в случае с выбросами.

Визуализация

```
library(multcomp)

par(mar=c(5,4,6,2))

# С выбросами
tuk <- glht(fit, linfct=mcp(diet.type="Tukey"))
plot(cld(tuk, level=.05),col="lightgrey")
```




```
# Без выбросов
tuk.new <- glht(fit.new, linfct=mcp(diet.type="Tukey"))
plot(cld(tuk.new, level=.05),col="lightgrey")
```


a
a
b
weight.loss
-2
0
2
4
6
8
A
B
C
diet.type