Curs BiostatisticÃ„Æ' 2017 - Laborator 5

AnalizÃ„Æ' de varianÃ^â€ºÃ„Æ'

Analiză de varianță cu un factor (one-way ANOVA)

Exemplul 1

Vom analiza setul de date `Cushings` din pachetul `MASS`. Sindromul *Cushing* reprezintă o serie de semne și simptome ca urmare a expunerii organismului pentru o perioadă îndelungată de timp la o concentrație ridicată de cortizon (mai multe detalii aici și aici). Pentru fiecare individ din eșantion, ratele de excreție urinară a doi metaboliți steroizi sunt înregistrate: *Tetrahydrocortisone* și *Pregnanetriol*. Variabila *Type* arată tipul de sindrom Cushing, acesta putând lua una din următoarele patru categorii: *adenom* (a), *hiperplazia bilaterală* (b), *carcinom* (c) și *necunoscut* (u). Obiectivul este să investigăm dacă cele patru tipuri de sindrom sunt diferite în raport cu excreția urinară de *Tetrahydrocortisone*.

Începem prin a atașa setul de date `Cushings`:

```
library(MASS)
data("Cushings")
attach(Cushings)
```

Tetrahydrocortisone	Pregnanetriol	Type
3.1	11.70	a
3.0	1.30	a
1.9	0.10	a
3.8	0.04	a
4.1	1.10	a
1.9	0.40	a
8.3	1.00	b
3.8	0.20	b
3.9	0.60	b
7.8	1.20	b
9.1	0.60	b
15.4	3.60	b
7.7	1.60	b
6.5	0.40	b
5.7	0.40	b
13.6	1.60	b
10.2	6.40	c
9.2	7.90	c
9.6	3.10	c
53.8	2.50	c
15.8	7.60	c
5.1	0.40	u

Tetrahydrocortisone	Pregnanetriol	Type
12.9	5.00	u
13.0	0.80	u
2.6	0.10	u
30.0	0.10	u
20.5	0.80	u

Notăm cu Y excreția urinară de *Tetrahydrocortisone* (variabila răspuns) și cu X variabila *Type* (variabila factor), cu $X \in \{1, 2, 3, 4\}$ după cum $Type \in \{a, b, c, u\}$. Astfel obiectivul este de a investiga dacă media variabilei răspuns Y diferă pentru valori diferite ale nivelelor variabilei factor X. Dacă notăm observațiile individuale cu y_{ij} (excreția urinară de *Tetrahydrocortisone* a individului j cu tipul de sindrom i) atunci putem determina

- numărul de observații din fiecare grup (n_i)

```
n = length(Cushings$Tetrahydrocortisone)

# varianta 1 - nr de observatii pe grup
ng = table(Cushings$Type)
ng
```

```
##
##  a  b  c  u
##  6 10  5  6
```

```
# varianta 2 - nr de observatii pe grup
ng2 = tapply(Cushings$Tetrahydrocortisone, Cushings$Type, length)
ng2
```

```
##  a  b  c  u
##  6 10  5  6
```

- media fiecărui grup ($\bar{y}_i$)

```
# media globala
my = mean(Cushings$Tetrahydrocortisone)

# varianta 1 - media pe grup
myg = tapply(Cushings$Tetrahydrocortisone, Cushings$Type, mean)
myg
```

```
##         a         b         c         u
##  2.966667  8.180000 19.720000 14.016667
```

```
# varianta 2 - media pe grup
myg2 = aggregate(Cushings$Tetrahydrocortisone, by = list(Cushings$Type), mean)
myg2
```

```
##   Group.1         x
## 1       a  2.966667
## 2       b  8.180000
## 3       c 19.720000
## 4       u 14.016667
```

- deviația standard a fiecărui grup

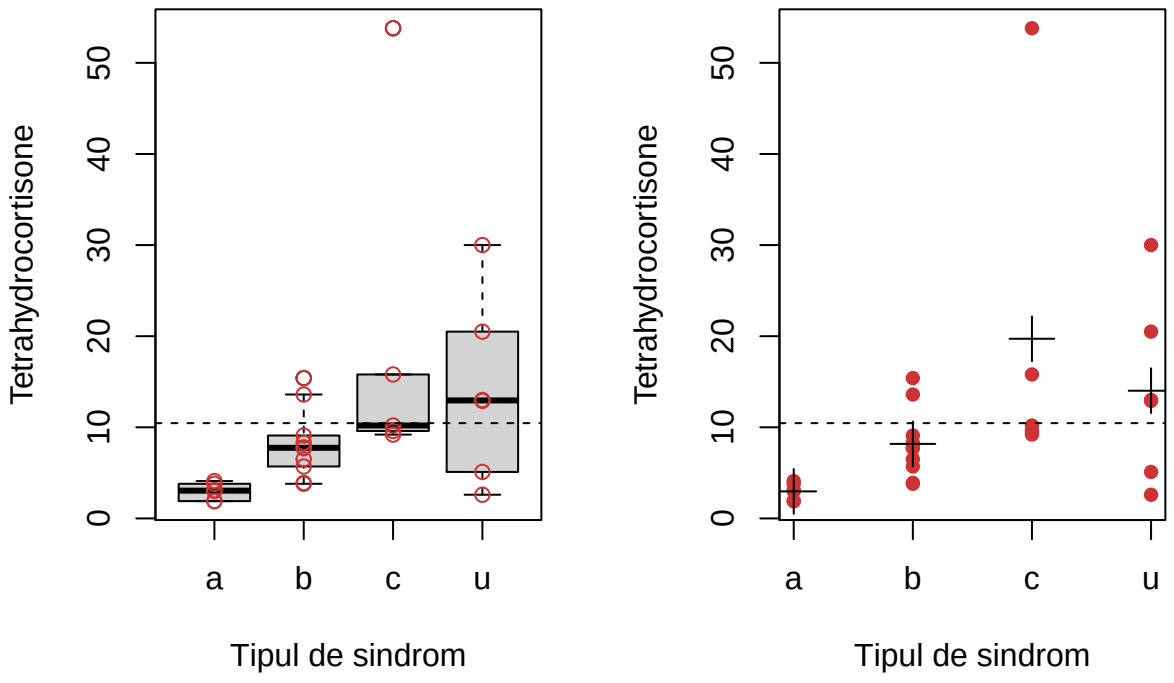
```r
# varianta 1 - media pe grup
syg = tapply(Cushings$Tetrahydrocortisone, Cushings$Type, sd)
syg
```

```
##         a         b         c         u
##  0.9244818  3.7891072 19.2388149 10.0958242
```

```r
# varianta 2 - media pe grup
syg2 = aggregate(Cushings$Tetrahydrocortisone, by = list(Cushings$Type), sd)
syg2
```

```
##   Group.1          x
## 1       a  0.9244818
## 2       b  3.7891072
## 3       c 19.2388149
## 4       u 10.0958242
```

Considerăm următorul grafic unde fiecare observație este reprezentată printr-un punct (gol în figura din stânga și plin în cea din dreapta) iar media globală este ilustrată printr-o linie punctată. În figura din stânga avem *boxplot*-ul pentru fiecare categorie a lui X iar în figura din dreapta (*stripchart*) mediile eșantioanelor din fiecare grup sunt ilustrate cu o cruce de culoare neagră:



Din figura de mai sus putem observa că avem o variație considerabilă între mediile grupurilor de-a lungul celor 4 categorii de sindrom *Cushing*. De asemenea, în interiorul grupurilor, avem grade diferite de variație a observațiilor (vezi figura din stânga). Ambele surse de variabilitate contribuie la variabilitatea totală a observațiilor în jurul mediei globale (linia punctată).

Calculăm **variabilitatea dintre grupuri** (r este numărul de grupuri):

$$SS_B = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2$$

```
# avem ng nr de observatii din fiecare grup, myg media lui y din fiecare grup si my media totala

SS_B = ng%*%(myg-my)^2 # unde %*% este produs de matrice
SS_B
```

```
##          [,1]
## [1,] 893.521
```

Calculăm **variabilitatea reziduală** (din grupuri):

$$SS_W = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

```
y = Cushings$Tetrahydrocortisone # y_{ij}
ryi = rep(myg, ng)

SS_W = sum((y-ryi)^2)
SS_W
```

```
## [1] 2123.646
```

Calculăm **variabilitatea totală**:

$$SS_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = SS_B + SS_W$$

```
# calculat cu SS_B+SS_W
SS_T = SS_B + SS_W
SS_T
```

```
##          [,1]
## [1,] 3017.167
```

```
# calculat cu sume (verificam formula)
SS_T2 = sum((y-my)^2)
SS_T2
```

```
## [1] 3017.167
```

Observăm că *variabilitatea totală poate fi atribuită parțial variabilității dintre grupuri și parțial variabilității din interiorul grupurilor.*

Considerăm ipoteza nulă:

$$H_0 : \mu_1 = \cdots = \mu_i = \mu$$

unde μ este media populației Y iar $\mu_1, \ldots, \mu_i$ sunt mediile populațiilor din fiecare grup.

Statistica de test este:

$$F = \frac{\frac{SS_B}{r-1}}{\frac{SS_W}{n-r}}$$

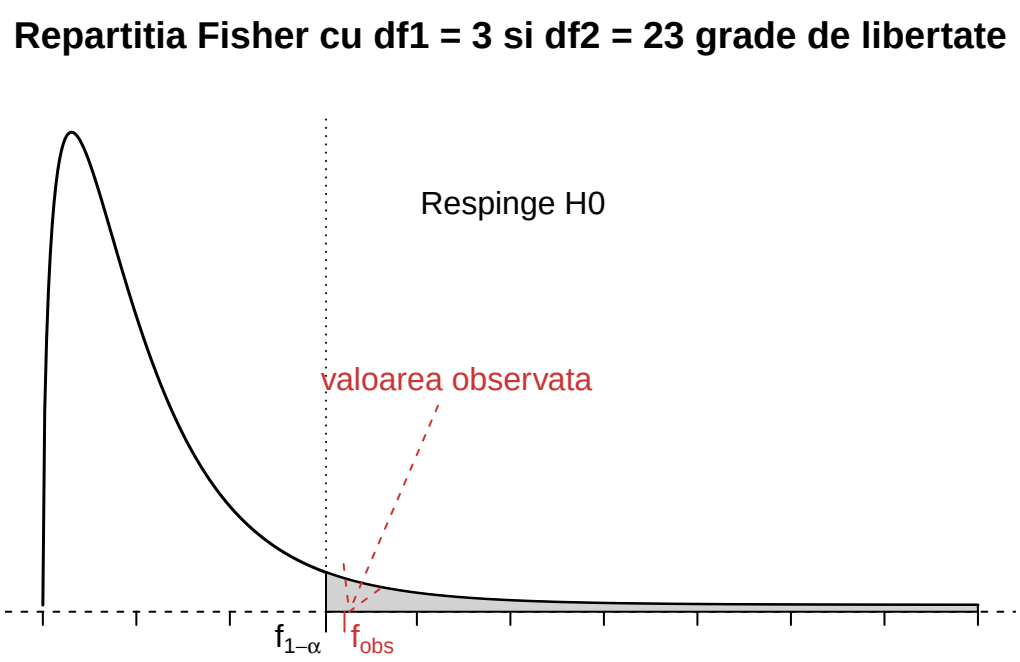
unde $\frac{SS_B}{r-1}$ și $\frac{SS_W}{n-r}$ sunt mediile pătrate pentru grupuri (mean square) și respectiv reziduri. Dacă condițiile *ANOVA* (datele din fiecare grup sunt i.i.d. și sunt normal distribuite) sunt satisfăcute și presupunând că H_0 este adevărată avem că $F \sim F(r-1, n-r)$.

Avem modelul *ANOVA*:

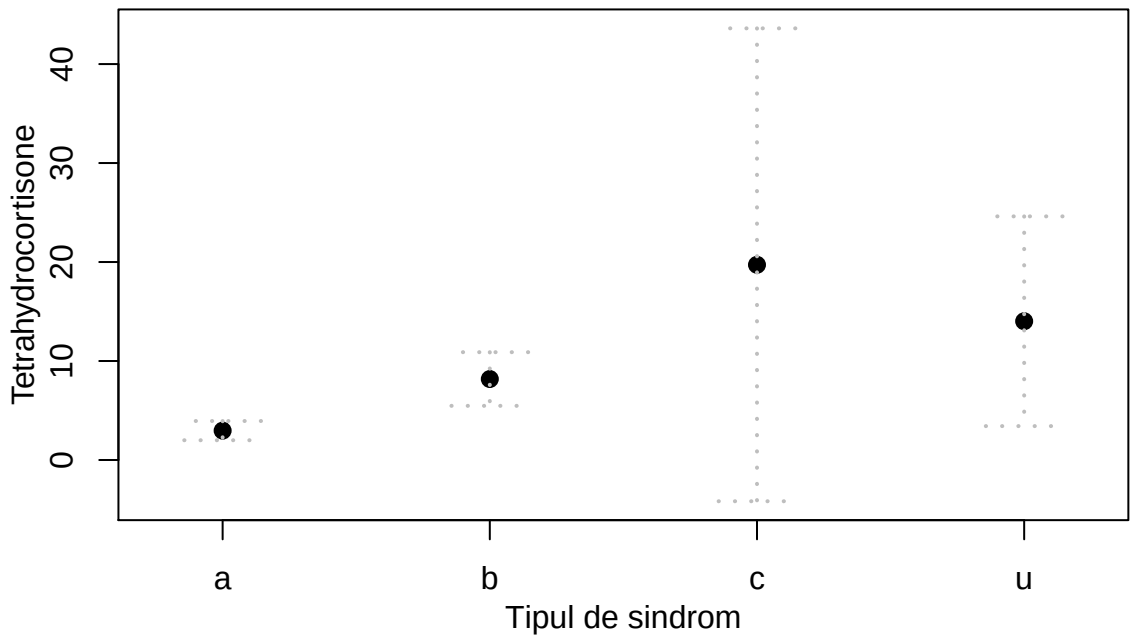
```
anova_model = aov(Tetrahydrocortisone~Type, data = Cushings)

summary(anova_model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Type         3  893.5  297.84   3.226 0.0412 *
## Residuals   23 2123.6   92.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Verificarea ipotezelor ANOVA



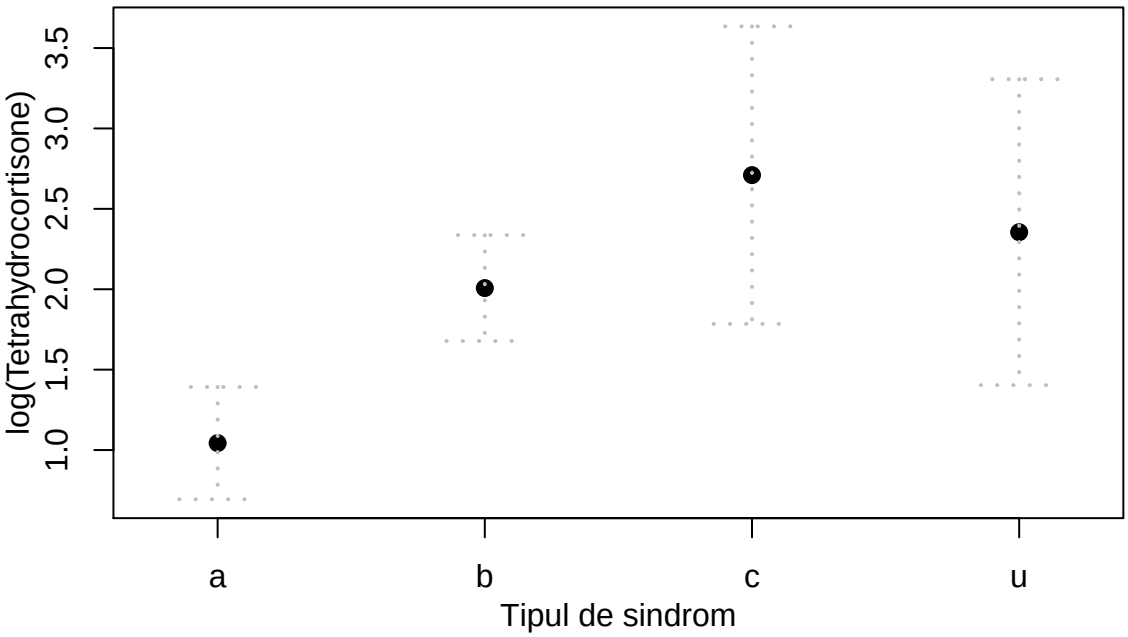
Aplicăm *testul lui Bartlett* pentru a testa homoscedasticitatea modelului (i.e. verificăm $H_0 : \sigma_1 = \cdots = \sigma_r$):

```
bartlett.test(Tetrahydrocortisone~Type, data = Cushings)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Tetrahydrocortisone by Type
## Bartlett's K-squared = 31.595, df = 3, p-value = 6.37e-07
```

Observăm că ipoteza de omogenitate este respinsă în favoarea alternativei prin urmare ipoteza de omogenitate din ANOVA este invalidată.

Transformăm variabila răspuns ($\log(Y) = \log(Tetrahydrocortisone)$):



Verificăm ipoteza de omogenitate (homoscedasticitatea):

```
bartlett.test(log(Tetrahydrocortisone)~Type, data = Cushings)
```

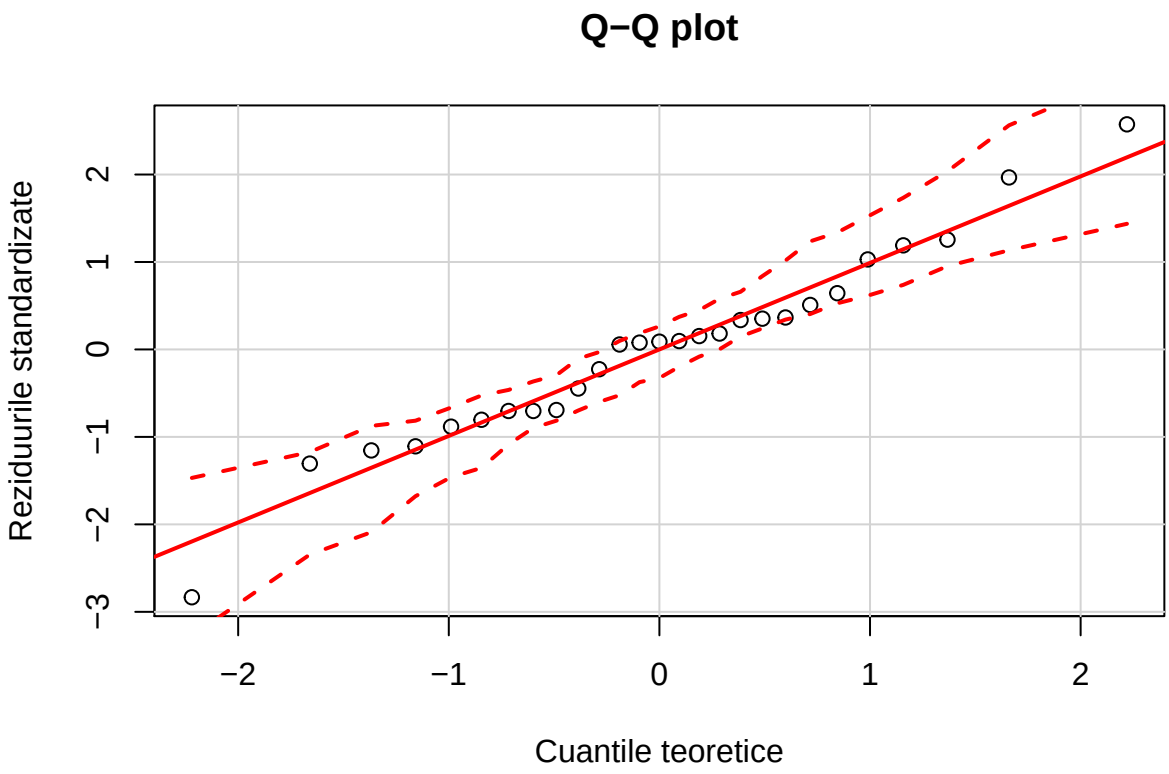
```
##
##  Bartlett test of homogeneity of variances
##
## data:  log(Tetrahydrocortisone) by Type
## Bartlett's K-squared = 5.7249, df = 3, p-value = 0.1258
```

Testăm normalitatea modelului transformat (*testul lui Shapiro-Wilks* sau *Shapiro-Francia*):

```
anova_model_tr = aov(log(Tetrahydrocortisone)~Type, data = Cushings)
shapiro.test(residuals(anova_model_tr))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_model_tr)
## W = 0.97953, p-value = 0.8515
```

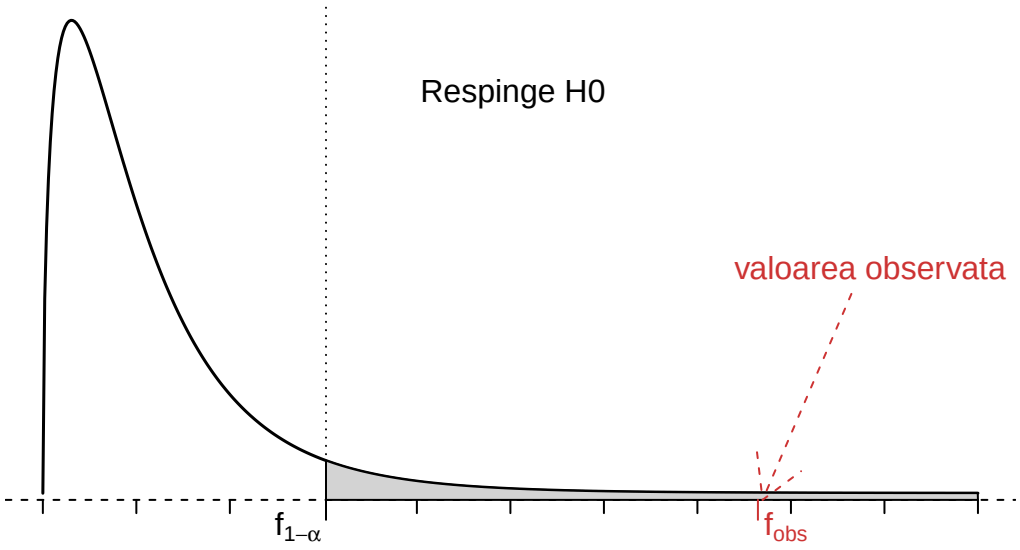
Verificăm normalitatea și grafic cu `Q-Q Plot`:



ANOVA pentru modelul transformat:

```
summary(anova_model_tr)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Type         3  8.766  2.9220   7.647 0.00102 **
## Residuals   23  8.789  0.3821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Exemplul 2

În acest exemplu vom folosi setul de date *Cholesterol* din pachetul `multcomp` (datele se pot descărca de aici). Datele prezintă cu cât s-a redus nivelul de colesterol (variabila *response*) la 50 de pacienți ce au urmat 5 tratamente de reducere a colesterolului. Trei dintre tratamente au implicat același medicament administrat în moduri diferite: 20 mg o dată pe zi (*1time*), 10 mg de două ori pe zi (*2time*) sau 5 mg de patru ori pe zi (*4time*). Celelalte două tratamente au constat din medicamente alternative diferite (*drugD* și *drugE*). Care tratament a produs cea mai mare reducere a colesterolului ?

Începem prin a citi setul de date:

```
cholesterol = read.csv("data/cholesterol.csv", stringsAsFactors = FALSE)
head(cholesterol)
```

```
##     trt response
## 1 1time   3.8612
## 2 1time  10.3868
## 3 1time   5.9059
## 4 1time   3.0609
## 5 1time   7.7204
## 6 1time   2.7139
```

Vedem câte observații avem pentru fiecare tratament:

```r
table(cholesterol$trt)
```

```
##
##  1time 2times 4times  drugD  drugE
##     10     10     10     10     10
```

Observăm că fiecare tratament a fost administrat la câte 10 pacienți (suntem în contextul unui *plan de experiență echilibrat*).

Calculăm:

- numărul total de observații (n) și numărul de observații din fiecare grup (n_i)

```r
n = length(cholesterol$trt) # nr total de observații

# nr de observatii pe grup
ng = table(cholesterol$trt)
```

- media fiecărui grup ($\bar{y}_i$)

```r
# media globala
my = mean(cholesterol$response)

# media pe grup
myg = tapply(cholesterol$response, cholesterol$trt, mean)
myg
```

```
##    1time   2times   4times    drugD    drugE
##  5.78197  9.22497 12.37478 15.36117 20.94752
```

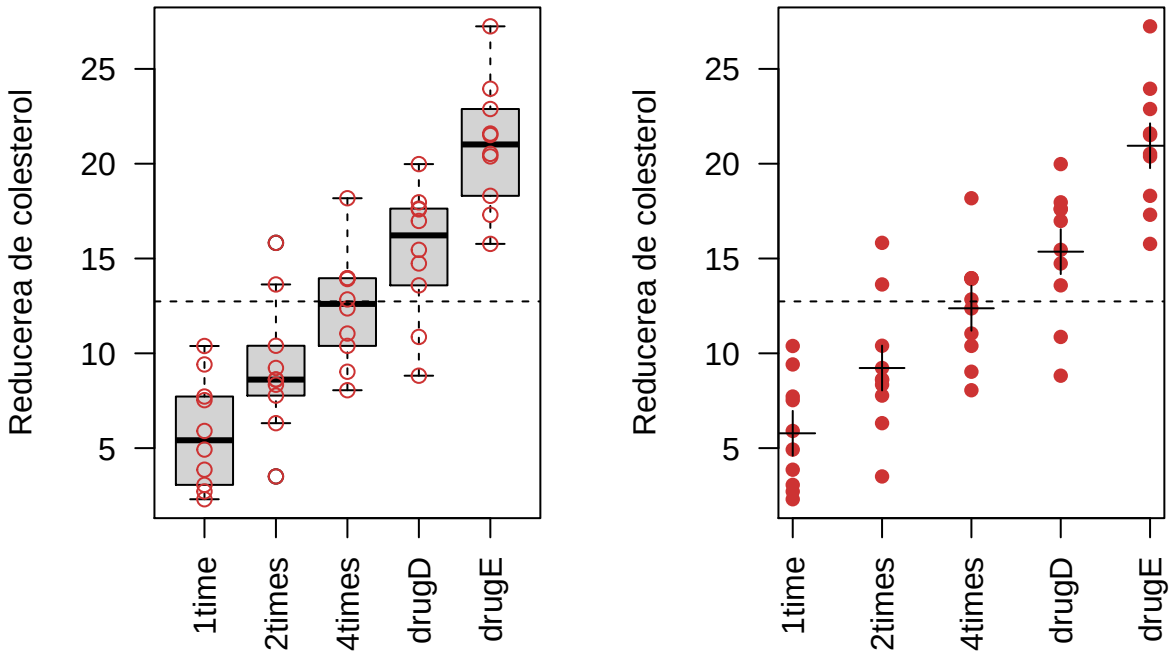
Se observă că drugE a produs (în medie) cea mai mare reducere a colesterolului pe când 1time a produs-o pe cea mai mică.

- abaterea standard a fiecărui grup

```r
# sd pe grup
syg = tapply(cholesterol$response, cholesterol$trt, sd)
syg
```

```
##    1time   2times   4times    drugD    drugE
## 2.878113 3.483054 2.923119 3.454636 3.345003
```

Se observă că abaterile standard sunt relativ constante pentru cele 5 tratamente, luând valori între 2.9 și 3.5.



Avem tabelul *ANOVA*:

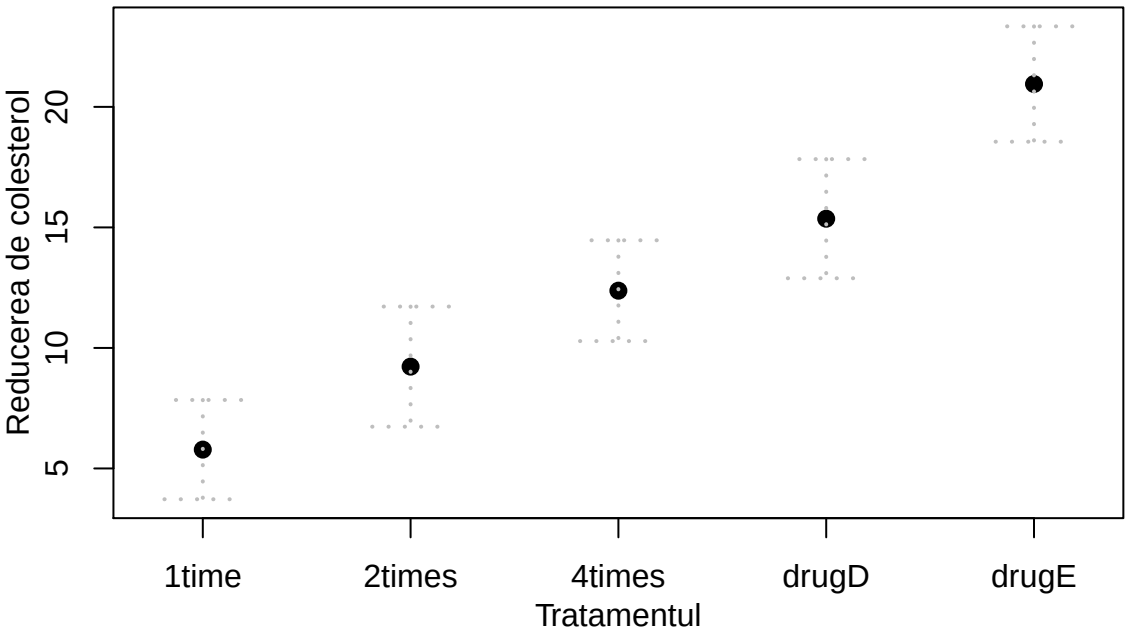
```
anova_model = aov(response~trt, data = cholesterol)

summary(anova_model)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## trt          4 1351.4   337.8   32.43 9.82e-13 ***
## Residuals   45  468.8    10.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testul ANOVA (F) pentru tratament (trt) este semnificativ ($p < 0.001$), ilustrând că cele 5 tratamente nu sunt la fel de eficiente.

Reducerea medie de colesterol pentru cele 5 tratamente împreună cu intervalele de încredere de nivel de încredere de 95% corespunzătoare:


Reducerea de colesterol
5
10
15
20
1time
2times
4times
drugD
drugE
Tratamentul