Curs Biostatistică 2017 - Laborator 5&6

Analiză de varianță - ANOVA

# 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](http://www.csid.ro/boli-afectiuni/endocrinologie/sindromul-cushing-12821884/) și [aici](https://en.wikipedia.org/wiki/Cushing%27s_syndrome)). 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](http://www.sfatulmedicului.ro/dictionar-medical/adenom_119)), *hiperplazia bilaterală* ([b](http://www.sfatulmedicului.ro/Afectiunile-suprarenalelor/hiperplazia-congenitala-a-glandelor-suprarenale_8123)), *carcinom* ([c](http://www.sfatulmedicului.ro/Cancer/ce-este-carcinomul_15473)) 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 |
| 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 excreția urinară de *Tetrahydrocortisone* (variabila răspuns) și cu variabila *Type* (variabila factor), cu după cum . Astfel obiectivul este de a investiga dacă media variabilei răspuns diferă pentru valori diferite ale nivelelor variabilei factor . Dacă notăm observațiile individuale cu (excreția urinară de *Tetrahydrocortisone* a individului cu tipul de sindrom ) atunci putem determina

* numărul de observații din fiecare grup ()

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 ()

# 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

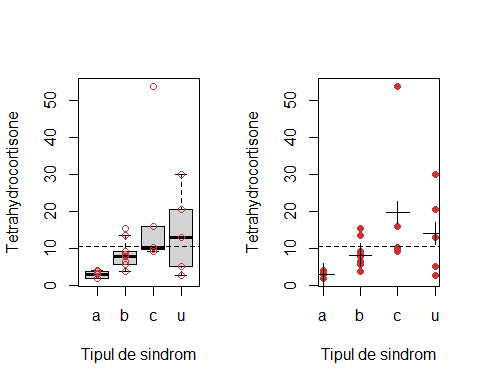
# varianta 1 - media pe grup   
syg = tapply(Cushings$Tetrahydrocortisone, Cushings$Type, sd)  
syg

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

# 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 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** ( este numărul de grupuri):

# 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):

y = Cushings$Tetrahydrocortisone # y\_{ij}  
ryi = rep(myg, ng)  
  
SS\_W = sum((y-ryi)^2)  
SS\_W

## [1] 2123.646

Calculăm **variabilitatea totală**:

# 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:

unde este media populației iar sunt mediile populațiilor din fiecare grup.

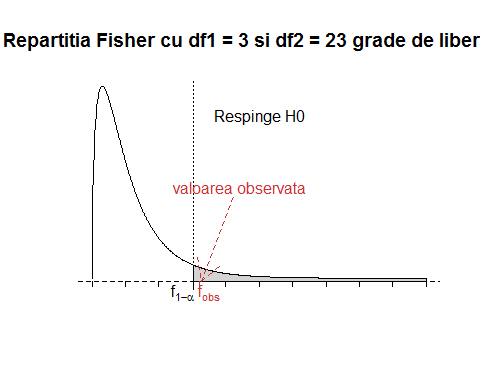
Statistica de test este:

unde și 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 este adevărată avem că .

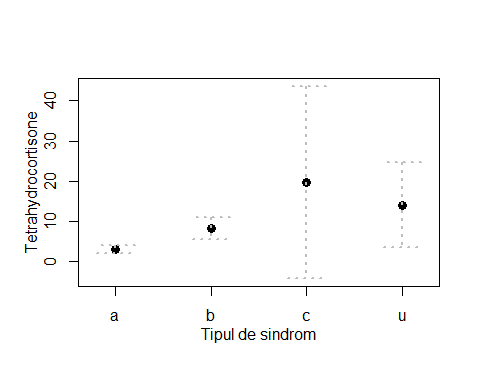
Avem modelul *ANOVA*:

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



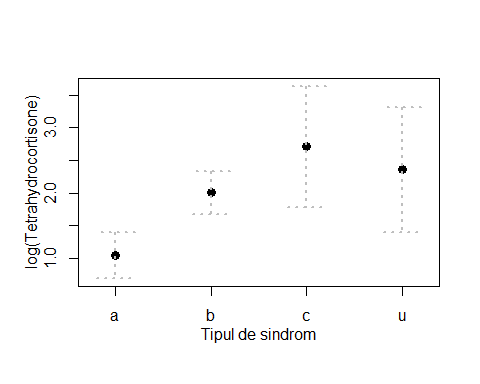
Aplicăm *testul lui Bartlett* pentru a testa homoscedasticitatea modelului (i.e. verificăm ):

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 ():



Verificăm ipoteza de omogenitate (homoscedasticitatea):

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

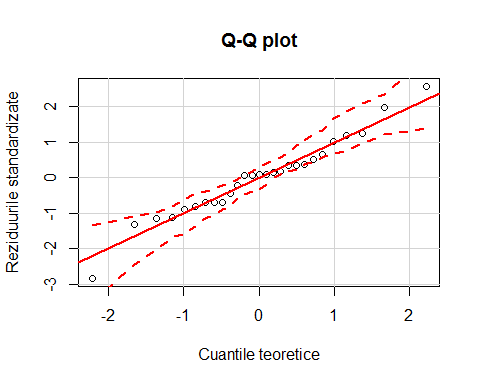
##   
## Bartlett test of homogeneity of variances  
##   
## data: log(Tetrahydrocortisone) by Type  
## Bartlett's K-squared = 5.7249, df = 3, p-value = 0.1258

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

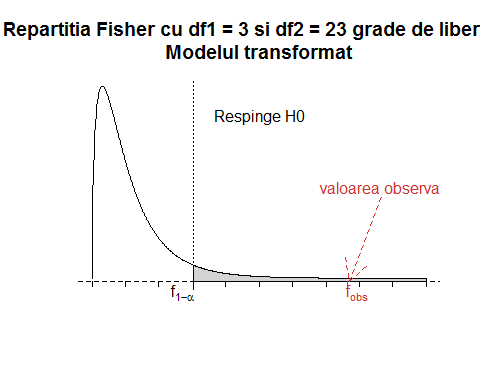
Verificăm normalitatea și grafic cu Q-Q Plot:



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](data/cholesterol.csv)). 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:

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 () și numărul de observații din fiecare grup ()

n = length(cholesterol$trt) # nr total de observații  
  
# nr de observatii pe grup  
ng = table(cholesterol$trt)

* media fiecărui grup ()

# 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

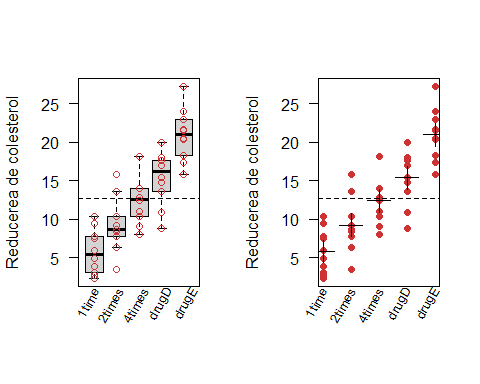
Se observă că drugE a produs (în medie) cea mai mare reducere a colesterolului pe când 1time a produs-o pe cea mai mică.

* abaterea standard a fiecărui grup

# 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.



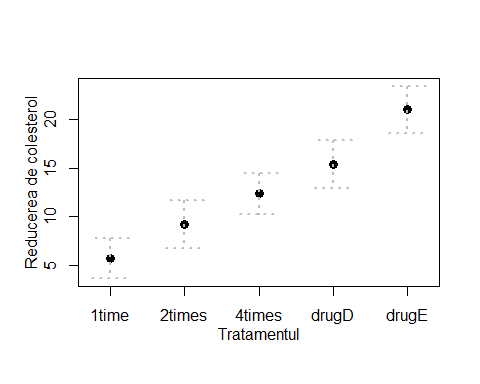
Avem tabelul *ANOVA*:

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

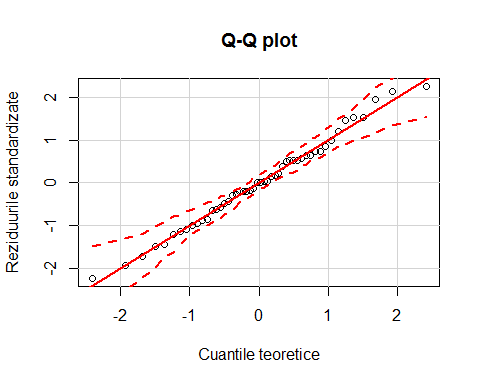
Testul ANOVA (F) pentru tratament (trt) este semnificativ (), ilustrând că cele 5 tratamente nu sunt la fel de eficiente.

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



### Verificarea ipotezelor ANOVA

În ANOVA cu un factor, se presupune că variabila răspuns este repartizată normal cu aceeași varianță în fiecare grup. Pentru testarea normalității putem folosi ca metodă grafică Q-Q plot-ul:



De asemenea ipoteza de normalitate poate fi testată și cu testul *Shapiro-Wilks* sau *Shapiro-Francia*:

anova\_model\_chol = aov(response~trt, data = cholesterol)  
shapiro.test(residuals(anova\_model\_chol))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(anova\_model\_chol)  
## W = 0.98864, p-value = 0.9094

Pentru testarea ipotezei de homoscedasticitate aplicăm *testul lui Bartlett*(i.e. verificăm ):

bartlett.test(response~trt, data = cholesterol)

##   
## Bartlett test of homogeneity of variances  
##   
## data: response by trt  
## Bartlett's K-squared = 0.57975, df = 4, p-value = 0.9653

Testul lui Bartlett ne indică faptul că varianțele în cele 5 grupuri nu diferă semnificativ (). Pentru testarea ipotezei de omogenitate se mai pot folosi și alte teste printre care includem *testul lui Fligner-Killeen* (fligner.test) și *testul Brown-Forsythe* (funcția hov() din pachetul HH). Ambele teste întorc același rezultat:

fligner.test(response~as.factor(trt), data = cholesterol)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: response by as.factor(trt)  
## Fligner-Killeen:med chi-squared = 0.74277, df = 4, p-value = 0.946

hov(response~trt, data = cholesterol) # hov = homogeneity of variance

##   
## hov: Brown-Forsyth  
##   
## data: response  
## F = 0.075477, df:trt = 4, df:Residuals = 45, p-value = 0.9893  
## alternative hypothesis: variances are not identical

### Comparări multiple

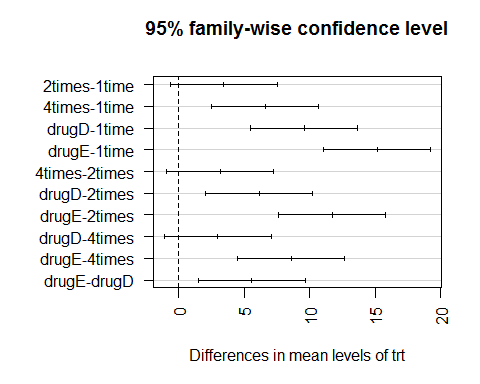
Testul F din ANOVA pentru tratamente ne spune că cele 5 tipuri de medicamente nu sunt la fel de eficiente, însă nu ne spune care dintre ele diferă față de celelalte. Pentru a răspunde la această întrebare vom folosi metodologia testării multiple. Ca exemplu vom folosi *Testul lui Tukey HSD* (Honestly Significant Difference), test care permite compararea tuturor perechilor de diferențe dintre mediile grupurilor:

TukeyHSD(anova\_model\_chol)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = response ~ trt, data = cholesterol)  
##   
## $trt  
## diff lwr upr p adj  
## 2times-1time 3.44300 -0.6582817 7.544282 0.1380949  
## 4times-1time 6.59281 2.4915283 10.694092 0.0003542  
## drugD-1time 9.57920 5.4779183 13.680482 0.0000003  
## drugE-1time 15.16555 11.0642683 19.266832 0.0000000  
## 4times-2times 3.14981 -0.9514717 7.251092 0.2050382  
## drugD-2times 6.13620 2.0349183 10.237482 0.0009611  
## drugE-2times 11.72255 7.6212683 15.823832 0.0000000  
## drugD-4times 2.98639 -1.1148917 7.087672 0.2512446  
## drugE-4times 8.57274 4.4714583 12.674022 0.0000037  
## drugE-drugD 5.58635 1.4850683 9.687632 0.0030633

Observăm că reducerea medie a colesterolului pentru tratamentele *1time* și *2times* nu este semnificativă () pe când reducerea medie a colesterolului pentru tratamentele *1time* și *4times* este semnificativă ().

Aceste diferențe se pot observa și grafic:



Trebuie menționat că sunt mai multe metode pentru comparări multiple: *metoda Bonferroni*, *metoda contrastelor liniare*, *metoda bazată pe statistici de rang*, *metoda Newman Keuls* etc.

# Analiză de varianță cu doi factori (two-way ANOVA)