Proiect P&S - Analiza set date esoph

Cherim Erol, Roman Robert, Martinas Paul

28/1/2022

Descrierea setului de date

Proiectul trateaza partea teoretica si practica (grafice/tabele) a unei probleme de interes public pe baza setului de date “esoph” din R.

Scopul proiectului este stabilirea conexiunii intre rata cancerului esofagial si 3 variabile:

- Varsta
- Consum de alcool
- Consum de tutun

Setul de date “esoph” este alcatuit din 88 de intrari, impartite dupa urmatoarele criterii:

- Categorie de varsta (ani): 25-34, 35-44, 45-54, 55-64, 65-74, 75+
- Consum de tutun (g/zi): 0-9, 10-19, 20-29, 30+
- Consum de alcool (g/zi): 0-39, 40-79, 80-119, 120+

Fiecare combinatie unica dintre cele 3 variabile vine insotita de numarul de cazuri de cancer si numarul de cazuri de control.

Structura setului de date

Variabile control pentru fiecare coloana:

```
summary(esoph)
```

```
##    agegp          alcgp         tobgp        ncases         ncontrols
##  25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00
##  35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00
##  45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00
##  55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08
##  65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00
##  75+  :11                                Max.   :17.000   Max.   :60.00
```

Se poate observa structura datelor pentru primele 15 intrari, ce reprezinta datele pentru categoria de varsta 35-44 ani:

```
##    agegp     alcgp    tobgp ncases ncontrols
## 16 35-44 0-39g/day 0-9g/day      0        60
## 17 35-44 0-39g/day    10-19      1        14
## 18 35-44 0-39g/day    20-29      0         7
## 19 35-44 0-39g/day      30+      0         8
## 20 35-44     40-79 0-9g/day      0        35
## 21 35-44     40-79    10-19      3        23
## 22 35-44     40-79    20-29      1        14
## 23 35-44     40-79      30+      0         8
## 24 35-44    80-119 0-9g/day      0        11
## 25 35-44    80-119    10-19      0         6
## 26 35-44    80-119    20-29      0         2
## 27 35-44    80-119      30+      0         1
## 28 35-44      120+ 0-9g/day      2         3
## 29 35-44      120+    10-19      0         3
## 30 35-44      120+    20-29      2         4
```

Manipularea si analiza datelor

Pentru evidentierea celor mai expuse categorii de varsta trebuie realizata distributia procentajului de cazuri de cancer `ncases` normalizata in functie de suma numarului cazurilor de control pentru fiecare categorie `sum(ncontrols)` .

Din cauza naturii structurii setului, datele trebuie agregate in functie de categoria de varsta.

Obiectul `cases_by_age` reprezinta numarul total (`FUN=sum`) de cazuri de cancer `ncases` pentru fiecare cateogire de varsta, calculat cu functia `aggregate` .

```
cases_by_age <- aggregate(esoph$ncases, by=list(agegp = esoph$agegp), FUN = sum)
cases_by_age
```

```
##   agegp  x
## 1 25-34  1
## 2 35-44  9
## 3 45-54 46
## 4 55-64 76
## 5 65-74 55
## 6   75+ 13
```

Obiectul `controls_by_age` reprezinta numarul total de control `ncotrols` pentru fiecare categorie de varsta. Se calculeaza de asemenea cu ajutorul functiei `aggregate` .

```
controls_by_age <- aggregate(esoph$ncontrols, by=list(agegp = esoph$agegp), FUN = su
m)
controls_by_age
```

```
##   agegp   x
## 1 25-34 116
## 2 35-44 199
## 3 45-54 213
## 4 55-64 242
## 5 65-74 161
## 6   75+  44
```

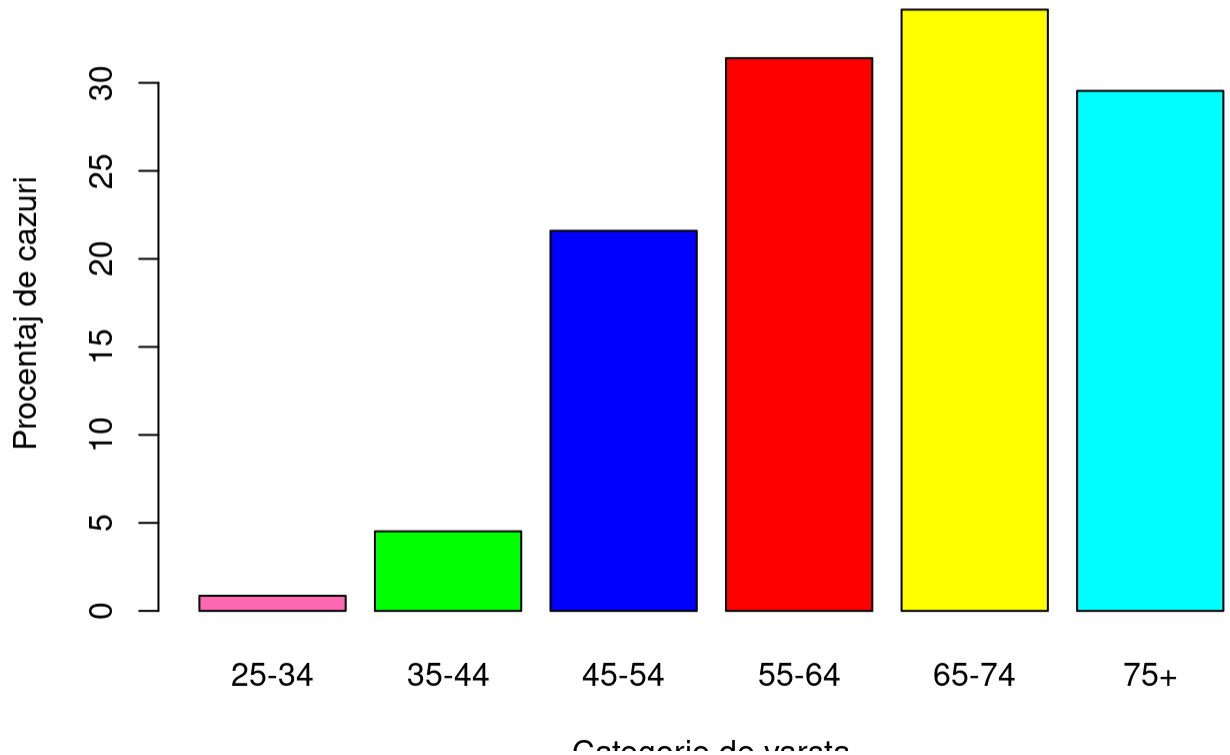
Vectorul `percentage_of_total` contine procentul de cazuri de cancer `ncases` din numarul de cazuri de control `ncontrols` .

```
percentage_of_total <- cases_by_age$x * 100 / controls_by_age$x
```

Distributia procentului cazurilor de cancer in functie de categoria de varsta

```
vect_col <- c("hotpink", "green", "blue", "red", "yellow", "cyan")

barplot ( height = percentage_of_total,
          names.arg = cases_by_age$agegp,
          xlab = "Categorie de varsta",
          ylab = "Procentaj de cazuri",
          col = vect_col,
          )
```



Categorie de varsta

Se observa cresterea cazurilor de cancer odata cu varsta, cele mai vulnerbaile categorii de varsta fiind 55-64, 65-74, 75+. Pentru observarea efectelor fiecarei substante, se realizeaza distributia consumului de alcool/tutun pentru fiecare categorie de varsta/categorie de consum.

Datele trebuie grupate in functie de categoria de varsta si consumul de tutun. Se genereaza subsetul `tob1` ce contine toate intrarile din setul de date ce se incadreaza in categoria de consum `"0-9g/day"` . Se pot observa primele 10 intrari pentru subsetul `tob1` .

```
tob1 <- subset(esoph, subset = tobgp == "0-9g/day")
head(tob1,10)
```

```
##    agegp     alcgp    tobgp ncases ncontrols
## 1  25-34 0-39g/day 0-9g/day      0        40
## 5  25-34     40-79 0-9g/day      0        27
## 9  25-34    80-119 0-9g/day      0         2
## 12 25-34      120+ 0-9g/day      0         1
## 16 35-44 0-39g/day 0-9g/day      0        60
## 20 35-44     40-79 0-9g/day      0        35
## 24 35-44    80-119 0-9g/day      0        11
## 28 35-44      120+ 0-9g/day      2         3
## 31 45-54 0-39g/day 0-9g/day      1        46
## 35 45-54     40-79 0-9g/day      6        38
```

Obiectul `cases_by_age_tobacco1` reuneste variabilele `ncases` , `ncontrols` , grupate in functie de fiecare categorie de varsta. Se folosesc functiile `cbind` si `aggregate` .

```
cases_by_age_tobacco1 <- aggregate(cbind(ncases=tob1$ncases, ncontrols= tob1$ncontrol
s), by=list(interv_varsta = tob1$agegp), FUN=sum)
cases_by_age_tobacco1
```

```
##   interv_varsta ncases ncontrols
## 1         25-34      0        70
## 2         35-44      2       109
## 3         45-54     14       104
## 4         55-64     25       117
## 5         65-74     31        99
## 6           75+      6        26
```

Se genereaza vectorul `proc_tob1` ce contine procentul de cazuri de cancer pentru fiecare categorie de varsta, pentru categoria de consum `"0-9g/day"` .

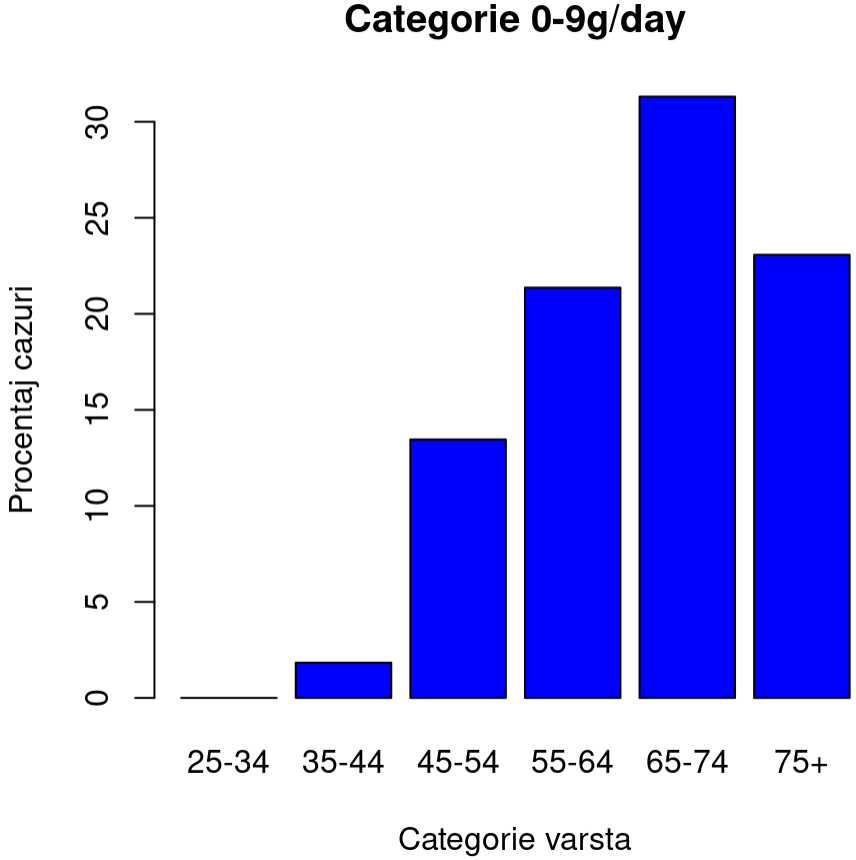
```
proc_tob1 <- cases_by_age_tobacco1$ncases*100/cases_by_age_tobacco1$ncontrol
proc_tob1
```

```
## [1]  0.000000  1.834862 13.461538 21.367521 31.313131 23.076923
```

Distributia procentajului cazurilor de cancer pentru consumatorii de tutun in fucntie de categoria de varsta

Se genereaza plotul ce evidentiaza procentul `proc_tob1` de cazuri de cancer raportat la categoria de varsta `agegp` pentru categoria de consum `"0-9g/day"` .

```
barplot(proc_tob1,
        names.arg = cases_by_age$agegp,
        main = "Categorie 0-9g/day",
        xlab = "Categorie varsta",
        ylab = "Procentaj cazuri",
        col = "blue",
        beside=TRUE,
        )
```



Analog si pentru celelalte categorii de consum:

- 10-19g/day
- 20-29g/day
- 30+g/day

```
tob2 <- subset(esoph, subset = tobgp == "10-19")
cases_by_age_tobacco2 <- aggregate(cbind(tob2$ncases, tob2$ncontrols), by=list(agegp
= tob2$agegp), FUN=sum)
cases_by_age_tobacco2
```

```
##   agegp V1 V2
## 1 25-34  1 19
## 2 35-44  4 46
## 3 45-54 13 57
## 4 55-64 23 65
## 5 65-74 12 38
## 6   75+  5 11
```

```
proctb2 <- cases_by_age_tobacco2$V1*100/cases_by_age_tobacco2$V2
proctb2
```

```
## [1]  5.263158  8.695652 22.807018 35.384615 31.578947 45.454545
```

```
tob3 <- subset(esoph, subset = tobgp == "20-29")
cases_by_age_tobacco3 <- aggregate(cbind(tob3$ncases, tob3$ncontrols), by=list(agegp
= tob3$agegp), FUN=sum)
cases_by_age_tobacco3
```

```
##   agegp V1 V2
## 1 25-34  0 11
## 2 35-44  3 27
## 3 45-54  8 33
## 4 55-64 12 38
## 5 65-74 10 20
## 6   75+  0  3
```

```
proctb3 <- cases_by_age_tobacco3$V1*100/cases_by_age_tobacco3$V2
proctb3
```

```
## [1]  0.00000 11.11111 24.24242 31.57895 50.00000  0.00000
```

```
tob4 <- subset(esoph, subset = tobgp == "30+")
cases_by_age_tobacco4 <- aggregate(cbind(tob4$ncases, tob4$ncontrols), by=list(agegp
= tob4$agegp), FUN=sum)
cases_by_age_tobacco4
```

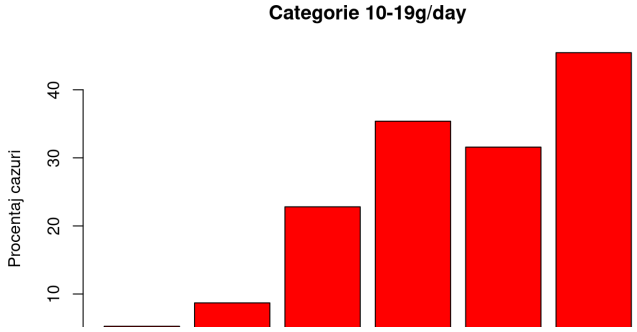
```
##   agegp V1 V2
## 1 25-34  0 16
## 2 35-44  0 17
## 3 45-54 11 19
## 4 55-64 16 22
## 5 65-74  2  4
## 6   75+  2  4
```

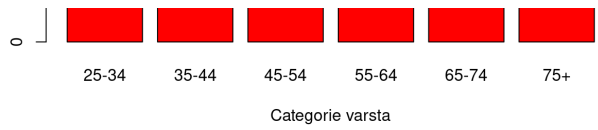
```
proctb4 <- cases_by_age_tobacco4$V1*100/cases_by_age_tobacco4$V2
proctb4
```

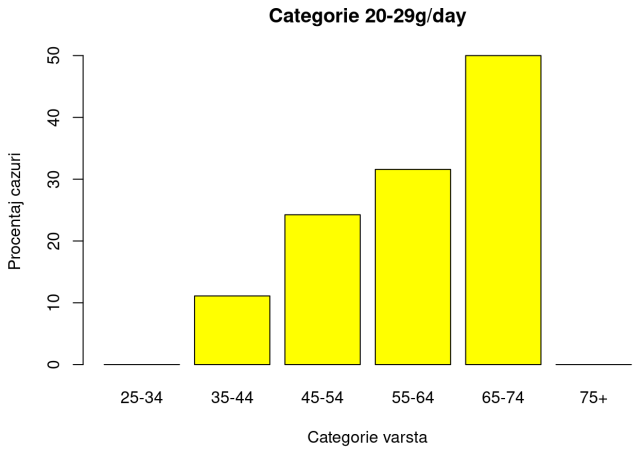
```
## [1]  0.00000  0.00000 57.89474 72.72727 50.00000 50.00000
```

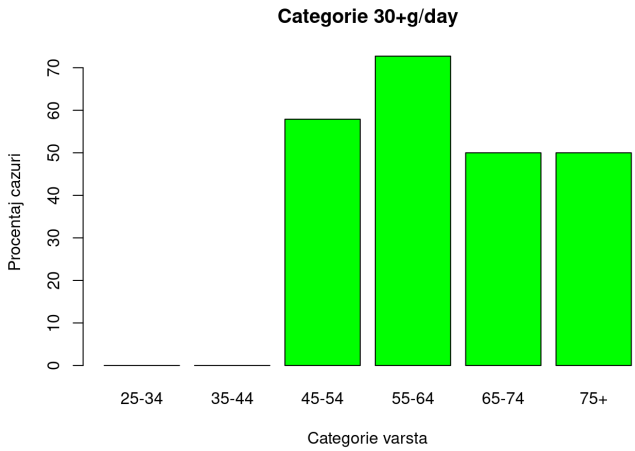
Si ploturile `barplot` aferente:

```
barplot(proctb2,
        names.arg = cases_by_age$agegp,
        main = "Categorie 10-19g/day",
        xlab = "Categorie varsta",
        ylab = "Procentaj cazuri",
        col = "red",
        beside=TRUE,
        )
barplot(proctb3,
        names.arg = cases_by_age$agegp,
        main = "Categorie 20-29g/day",
        xlab = "Categorie varsta",
        ylab = "Procentaj cazuri",
        col = "yellow",
        beside=TRUE,
        )
barplot(proctb4,
        names.arg = cases_by_age$agegp,
        main = "Categorie 30+g/day",
        xlab = "Categorie varsta",
        ylab = "Procentaj cazuri",
        col = "green",
        beside=TRUE,
        )
```









Dupa cum se poate observa din graficul `"30+g/day"` , in categoriile tinere de varsta, respectiv `25-34` si `35-44` nu se depisteaza cazuri de cancer esofagian, de unde se poate observa o crestere a consumului de tutun odata cu varsta. (Categoriile predispuse la un consum mai mare de tutun sunt cele mai in varsta).

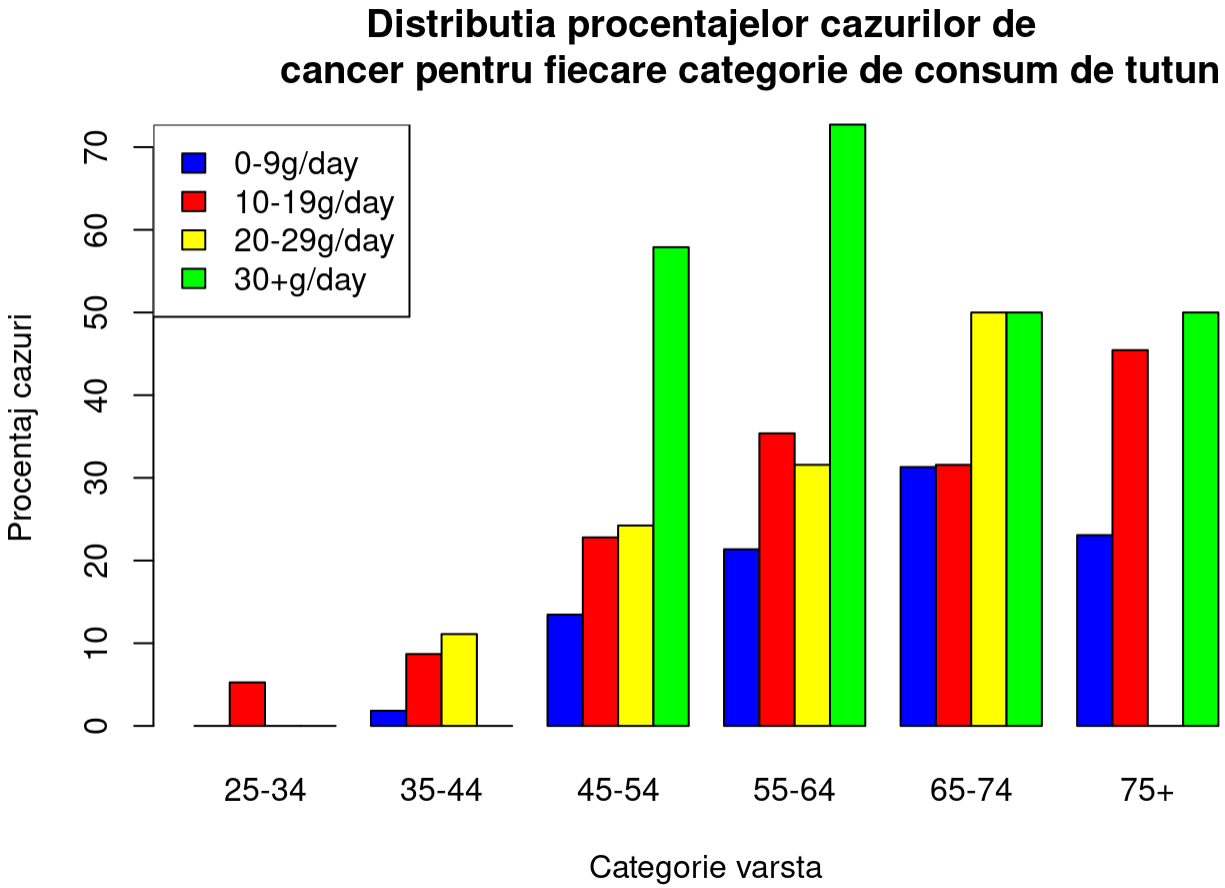
Pentru unificarea celor 4 grafice se foloseste functia `cbind` , in variabila `tob_t` .

```
tob_t <- cbind("0-9"=proc_tob1, "10-19"=proctb2, "20-29"=proctb3, "30+"=proctb4)
tob_t
```

```
##            0-9     10-19    20-29      30+
## [1,]  0.000000  5.263158  0.00000  0.00000
## [2,]  1.834862  8.695652 11.11111  0.00000
## [3,] 13.461538 22.807018 24.24242 57.89474
## [4,] 21.367521 35.384615 31.57895 72.72727
## [5,] 31.313131 31.578947 50.00000 50.00000
## [6,] 23.076923 45.454545  0.00000 50.00000
```

Se realizeaza graficul pentru obiectul `tob_t` , ce reuneste toate categoriile de consum.

```
vect_col4 <- c("blue", "red", "yellow", "green")
barplot(t(tob_t),
        main = "Distributia procentajelor cazurilor de
        cancer pentru fiecare categorie de consum de tutun",
        names.arg = cases_by_age$agegp,
        xlab = "Categorie varsta",
        ylab = "Procentaj cazuri",
        col = vect_col4,
        beside=TRUE,
        )
legend("topleft", legend=c("0-9g/day","10-19g/day","20-29g/day","30+g/day"), fill=vec
t_col4)
```



In graficul anterior, se pot observa urmatoarele:

- Riscul de cancer creste direct proportional cu varsta
- Riscul de cancer creste direct proportional cu cantitatea de tutun consumata
- Riscul de cancer se dubleaza pentru consumatorii de mai mult de 30 degrame de tutun pe zi

Distributia procentului cazurilor de cancer pentru fiecare categorie de consum de alcool

Analog se genereaza un grafic si pentru categoriile de consum de alcool

```
alc1 <- subset(esoph, subset = alcgp == "0-39g/day")
cases_by_age_alcohol1 <- aggregate(cbind(alc1$ncases, alc1$ncontrols), by=list(agegp
= alc1$agegp), FUN=sum)
cases_by_age_alcohol1
```

```
##   agegp V1 V2
## 1 25-34  0 61
## 2 35-44  1 89
## 3 45-54  1 78
## 4 55-64 12 89
## 5 65-74 11 71
## 6   75+  4 27
```

```
procalc1 <- cases_by_age_alcohol1$V1*100/cases_by_age_alcohol1$V2
procalc1
```

```
## [1]  0.000000  1.123596  1.282051 13.483146 15.492958 14.814815
```

```
alc2 <- subset(esoph, subset = alcgp == "40-79")
cases_by_age_alcohol2 <- aggregate(cbind(alc2$ncases, alc2$ncontrols), by=list(agegp
= alc2$agegp), FUN=sum)
cases_by_age_alcohol2
```

```
##   agegp V1 V2
## 1 25-34  0 45
## 2 35-44  4 80
## 3 45-54 20 81
## 4 55-64 22 84
## 5 65-74 25 53
## 6   75+  4 12
```

```
procalc2 <- cases_by_age_alcohol2$V1*100/cases_by_age_alcohol2$V2
procalc2
```

```
## [1]  0.00000  5.00000 24.69136 26.19048 47.16981 33.33333
```

```
alc3 <- subset(esoph, subset = alcgp == "80-119")
cases_by_age_alcohol3 <- aggregate(cbind(alc3$ncases, alc3$ncontrols), by=list(agegp
= alc3$agegp), FUN=sum)
cases_by_age_alcohol3
```

```
##   agegp V1 V2
## 1 25-34  0  5
## 2 35-44  0 20
## 3 45-54 12 39
## 4 55-64 24 43
## 5 65-74 13 29
## 6   75+  2  2
```

```
procalc3 <- cases_by_age_alcohol3$V1*100/cases_by_age_alcohol3$V2
procalc3
```

```
## [1]   0.00000   0.00000  30.76923  55.81395  44.82759 100.00000
```

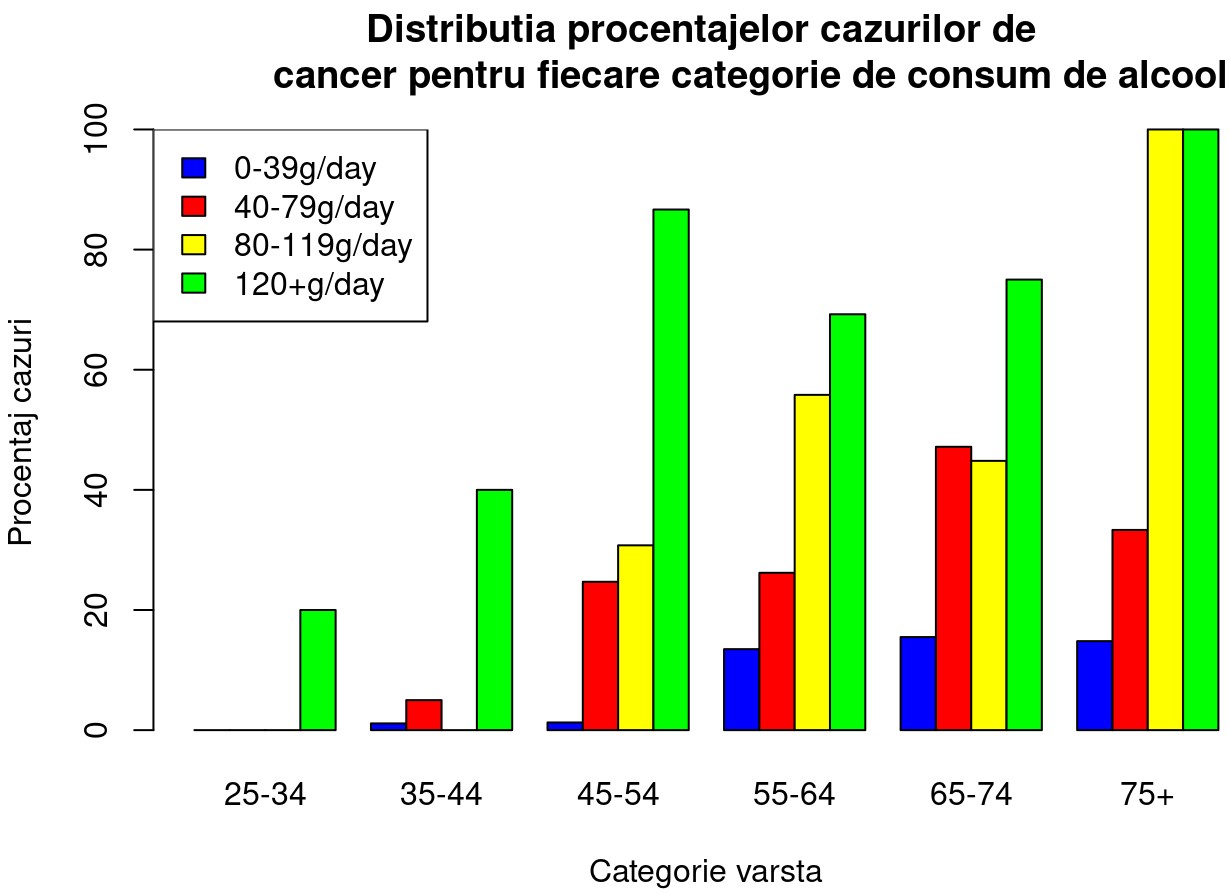
```
alc4 <- subset(esoph, subset = alcgp == "120+")
cases_by_age_alcohol4 <- aggregate(cbind(alc4$ncases, alc4$ncontrols), by=list(agegp
= alc4$agegp), FUN=sum)
cases_by_age_alcohol4
```

```
##   agegp V1 V2
## 1 25-34  1  5
## 2 35-44  4 10
## 3 45-54 13 15
## 4 55-64 18 26
## 5 65-74  6  8
## 6   75+  3  3
```

```
procalc4 <- cases_by_age_alcohol4$V1*100/cases_by_age_alcohol4$V2
procalc4
```

```
## [1]  20.00000  40.00000  86.66667  69.23077  75.00000 100.00000
```

```
#
-------------------------------------------------------------------------------
alc_t <- cbind("0-39"=procalc1, "40-79"=procalc2, "80-119"=procalc3, "120+"=procalc4)
vect_col5 <- c("blue", "red", "yellow", "green")
barplot(t(alc_t),
        main = "Distributia procentajelor cazurilor de
        cancer pentru fiecare categorie de consum de alcool",
        names.arg = cases_by_age$agegp,
        xlab = "Categorie varsta",
        ylab = "Procentaj cazuri",
        col = vect_col5,
        beside=TRUE,
)
legend("topleft", legend=c("0-39g/day","40-79g/day","80-119g/day", "120+g/day"), fill
=vect_col4)
```



In graficul anterior, se pot observa urmatoarele:

- Riscul de cancer creste direct proportional cu varsta
- Riscul de cancer creste direct proportional cu cantitatea de alcool consumata
- Riscul de cancer se tripleaza pentru consumatorii de mai mult de 120 de grame de alcool pe zi pentru categoria de varsta `45-54`
- Consumul moderat de alcool la categoriile de varsta `25-34` , `35-44` are un efect minim asupra riscului de cancer esofagian

Risk-Calculator

Aplicatia Shiny foloseste functia `probabilitate` pentru estima riscul aparitiei cancerului unui individ conform setului de date, in functie de urmatorul input:

- varsta
- consum alcool zilnic
- consum tutun zilnic

```
probabilitate <- function (age, alc, tob)
{
  if(age<25)
  {
    stop("Varsta minima este 25")
  }
  if(age>=25&&age<=34)
  {
    vage <- subset(esoph, subset = agegp == "25-34")
  }
  if(age>=35&&age<=44)
  {
    vage <- subset(esoph, subset = agegp == "35-44")
  }
  if(age>=45&&age<=54)
  {
    vage <- subset(esoph, subset = agegp == "45-54")
  }
  if(age>=55&&age<=64)
  {
    vage <- subset(esoph, subset = agegp == "55-64")
  }
  if(age>=65&&age<=74)
  {
    vage <- subset(esoph, subset = agegp == "65-74")
  }
  if(age>=75)
  {
    vage <- subset(esoph, subset = agegp == "75+")
  }


  if(alc<0)
  {
    stop("Cantatitea de alcool nu poate fi negativa")
  }
  if(alc>=0&&alc<=39)
  {
    valc <- subset(vage, subset = alcgp == "0-39g/day")
  }
  if(alc>=40&&alc<=79)
  {
    valc <- subset(vage, subset = alcgp == "40-79")
  }
  if(alc>=80&&alc<=119)
  {
    valc <- subset(vage, subset = alcgp == "80-119")
  }
  if(alc>=120)
  {
    valc <- subset(vage, subset = alcgp == "120+")
```

```
  }

  if(tob<0)
  {
    stop("Cantatitea de tutun nu poate fi negativa")
  }
  if(tob>=0&&tob<=0)
  {
    vtob <- subset(valc, subset = tobgp == "0-9g/day")
  }
  if(tob>=10&&tob<=19)
  {
    vtob <- subset(valc, subset = tobgp == "10-19")
  }
  if(tob>=20&&tob<=29)
  {
    vtob <- subset(valc, subset = tobgp == "20-29")
  }
  if(tob>=30)
  {
    vtob <- subset(valc, subset = tobgp == "30+")
  }
  if(length(vtob$ncases)>0)
  {
    controlsage <- sum(vage$ncontrols)
    result <- vtob$ncases*100/controlsage
    return(result)
  }
  else
  {
    warning("Nu exista memorata combinatia de valori")
  }
}
```

Spre exemplu, conform setului de date, sansa aparitiei cancerului esofagian pentru un individ cu urmatoarele caracteristici:

- `varsta = 64` ani
- `consum_alcool = 100g/zi`
- `consum_tutun = 40g/zi`

este de:

```
probabilitate(64, 100, 18)
```

```
## [1] 3.305785
```

%

Probleme întâlnite:

Pe versiunea de RStudio a colegilor mei (cea de Windows), setul de date `esoph` contine date eronate pentru ultimele intrari (Categoria de varsta 75+). Numarul de ncases>ncontrols (se depisteaza 1 caz din 0 pacienti), lucru care afecteaza procentul cazurilor depistate. Spre exemplu, pt categoria de varsta 75+, exista 8 ncases depistate din 5 ncontrols, deci procentajul de cazuri depistate este 160% (lucru care nu este posibil). Acest lucru afecteaza manipularea si analiza datelor.