

Capitolul 3 – Analiza de frecvențe



Îmi amintesc că atunci când aveam în jur de 5 ani, mergeam cu părinții în concediu și obișnuiam să număr mașinile care ne depășeau. Numărătoarea pe care o făceam ținea cont și de marca mașinii (spre norocul meu, pe atunci, nu erau decât vreo 4-5 mărci de mașini la noi în țară). După vreo doi ani, mi-am dat seama că ar fi mai bine să le număr pe cele pe care le depășeam noi. Când ajungeam la destinație, tata mă întreba, cu umor, la ce număr am ajuns cu numărătoarea. Pe atunci nu știam, și nici nu mi-a explicat cineva, că acel exercițiu se numește analiză de frecvențe. Cred că și voi ați experimentat la un moment dat astfel de jocuri. Apoi, pe măsură ce a trecut timpul, am tot făcut analize de frecvențe: numărul de goluri înscrise de fiecare echipă la Campionatul Mondial sau la Campionatul European, notele colegilor la anumite discipline pentru a ști când vine rândul meu să „ies la lecție”, numărul de absențe la fiecare disciplină în perioada liceului etc. Prin urmare, acest capitol are rolul de descrie tipurile de frecvențe și cum poate fi făcută analiza de frecvențe cu ajutorul softurilor statistice.

3.1 Analiza de frecvențe simple

În primul capitol ne-am familiarizat cu noțiunile elementare ale statisticii, aceste concepte oferindu-ne o imagine globală asupra a ceea ce va urma în continuare. Analiza de frecvențe este o componentă a statisticii descriptive. Am afirmat că statistica este o colecție de tehnici și proceduri de prezentare a datelor obținute în urma unei investigații. Pentru a fi eficienți, trebuie să folosim o serie de proceduri care să prezinte valorile măsurate într-o manieră care să permită interpretarea lor.

Profesorul de statistică aplică un chestionar de evaluare a conștiințozității studenților săi. Tabelul 3.1 prezintă scorurile obținute de cei 50 de studenți care au completat chestionarul.

- Poți descrie pe scurt scorurile obținute de studenți?
- Care este cel mai mic scor, dar cel mai mare?
- Care este scorul care apare de cele mai multe ori, dar cel cu cea mai mică frecvență?

Tabel 3.1 – Scorurile pentru variabila conștiinciozitate obținute de cei 50 de studenți

57	59	51	64	66
55	68	67	71	61
61	54	52	57	61
59	67	55	74	72
64	65	51	72	76
52	73	53	67	57
51	45	52	59	70
66	48	48	71	71
51	75	74	45	71
66	75	54	63	63

După cum se poate observa, ne este foarte greu să interpretăm aceste scoruri și să răspundem la întrebările de mai sus. De aceea, ne-ar fi mult mai ușor dacă le-am prezenta într-o manieră succintă, în care să evidențiem aspectele importante ale scorurilor obținute de studenți. Datele prezentate în Tabelul 3.1 sunt corecte și nimeni nu ne poate acuza că am dori să îl păcălim sau că nu avem suficiente informații. Problema modalității de expunere a scorurilor constă în faptul că nimeni nu reușește să înțeleagă ceva din el. Astfel, avem nevoie de o prezentare care să nu îl obosească pe cititor și care să îi permită să înțeleagă cum se distribuie scorurile variabilei conștiinciozitate în rândul celor 50 de participanți. Soluția constă în prezentarea unui **tabel de frecvențe** precum în Tabelul 3.2.

Tabelul 3.2 – Tabelul de frecvențe pentru scorurile din Tabelul 3.1

		Frecvența absolută (fa)	Frecvența cumulată (fc)	Frecvența procentuală (fr%)	Frecvența cumulată procentual (frc%)
Valid	45	2	2	4,0	4,0
	48	2	4	4,0	8,0
	51	4	8	8,0	16,0
	52	3	11	6,0	22,0
	53	1	12	2,0	24,0
	54	2	14	4,0	28,0
	55	2	16	4,0	32,0
	57	3	19	6,0	38,0
	59	3	22	6,0	44,0
	61	3	25	6,0	50,0
	63	2	27	4,0	54,0
	64	2	29	4,0	58,0
	65	1	30	2,0	60,0
	66	3	33	6,0	66,0
	67	3	36	6,0	72,0
	68	1	37	2,0	74,0
	70	1	38	2,0	76,0
	71	4	42	8,0	84,0
	72	2	44	4,0	88,0
	73	1	45	2,0	90,0
	74	2	47	4,0	94,0
	75	2	49	4,0	98,0

76	1	50	2,0	100,0
Total	50		100,0	

Acum este mult mai ușor să răspundem la întrebările expuse mai sus. Putem observa că scorul 45 este cel mai mic, iar scorul 76 este cel mai mare. De asemenea, putem vedea că scorurile 51 și 71 au fost obținute de cele de mai multe ori (de câte 4 studenți). Toate aceste afirmații sunt expresii ale analizei de frecvență.

Frecvența absolută (fa) este numărul de apariții al fiecărei valori în distribuție. Conform datelor din tabel, frecvența absolută a valorii 45 este 2 (doi studenți au obținut scorul 45), pentru valoarea 67 avem frecvența 3. Valorile cu frecvența 0 nu apar în tabel tocmai pentru a face mai ușoară interpretarea datelor. Suma tuturor frecvențelor absolute este egală cu numărul total de scoruri din distribuție (în cazul nostru, 50).

Frecvența cumulată (fc) reprezintă numărul total de valori începând de la cel mai mic scor din distribuție până la o anumită valoare, inclusiv. De exemplu, în Tabelul 3.2 avem opt valori până la scorul 51 sau 30 dintre studenții care au răspuns la chestionar au obținut un scor mai mic sau egal cu 65. Întotdeauna frecvența cumulată a ultimului scor din distribuție coincide cu suma frecvențelor absolute. În acest exemplu, frecvența cumulată pentru scorul 76 este 50.

Frecvența relativă (fr) indică probabilitatea de apariție a fiecărei valori din distribuție. Ea se calculează prin raportarea frecvenței absolute la numărul total de participanți. Suma tuturor frecvențelor relative este egală cu 1. În exemplul de mai sus, frecvența relativă a lui 71 este 0,08. Acest rezultat a fost obținut prin împărțirea lui 4 (frecvența absolută a lui 71) la 50 (numărul total de participanți).

Frecvența relativă procentuală (fr%) exprimă procentul care corespunde unei valori din cadrul distribuției. Aceasta se calculează prin înmulțirea cu 100 a raportului dintre frecvența absolută și suma tuturor frecvențelor absolute. Suma tuturor frecvențelor relative procentuale este întotdeauna egală cu 100%.

$$fr\% = \frac{fa}{\sum fa} * 100$$

(formula 3.1)

Analizând exemplul nostru observăm că 4% din studenți au obținut scorurile 45 și 48, iar 8% au obținut scorul 51 la chestionarul de conștiinciozitate.

Frecvența relativă cumulată procentuală (frc%) ne indică procentul cumulat al scorurilor din distribuție până la o anumită valoare, inclusiv. Pentru cel mai mare scor din distribuție întotdeauna frecvența cumulată procentuală este 100%. Pentru scorul 55 avem o frecvență cumulată procentuală de 32%. Acest rezultat se traduce prin faptul că **32% dintre studenți au un scor la chestionarul de conștiinciozitate mai mic sau egal cu 55**. De asemenea, 76% dintre studenți au obținut un scor mai mic sau egal cu 70.

Frecvența relativă cumulată procentuală se numește **rang percentil**. Scorul 64 are frecvența cumulată procentuală 58,0 și putem spune că îi corespunde rangul percentil 58. Valoarea corespunzătoare unui rang percentil poartă denumirea de **percentilă**. Astfel, scorul 64 este percentila 58.

Trebuie menționat faptul că există trei percentile speciale:

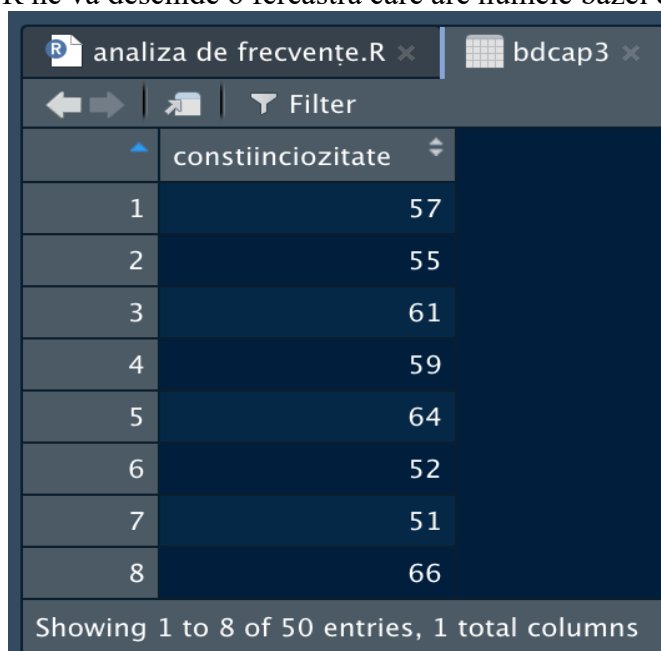
- **percentila 25** (*quartila 1*; corespunde rangului percentil 25).
- **percentila 50** (*quartila 2* sau *mediană*; corespunde rangului percentil 50).
- **percentila 75** (*quartila 3*; corespunde rangului percentil 75).

3.2 Analiza de frecvențe simple în R

Pentru a realiza analiza de frecvențe simple în R trebuie încărcat pachetul **psych**, precum în modelul de mai jos:

`library(psych)` – pentru încărcarea pachetului vom apăsa butonul **Run** sau combinația de taste **Ctrl+Enter/Comand+Enter**.

De asemenea, avem nevoie să încărcăm baza de date care conține variabilele pe care dorim să le analizăm. În cazul nostru, baza de date se numește **bdcap3** și este în format Microsoft Excel (**Import Dataset** → **From Excel** → **Browse** → **Import**). După ce baza de date a fost încărcată R ne va deschide o fereastră care are numele bazei de date.



	constiinciozitate
1	57
2	55
3	61
4	59
5	64
6	52
7	51
8	66

3.2.1 Frecvența absolută

Pentru a calcula frecvența absolută vom crea un obiect pe care îl vom numi sugestiv **faconst** (o denumire care să ne permită să înțelegem că este vorba despre frecvența absolută a variabilei conștiinciozitate). Acest obiect va primi rezultatele analizei de frecvență.

```
faconst <- table(bdcap3$constiinciozitate)
cbind(faconst)
```

- `table` – este funcția din R care calculează frecvențele absolute.
- `bdcap3` – reprezintă denumirea bazei de date.
- `constiinciozitate` – este variabila ale cărei frecvențe absolute dorim să le calculăm.
- `cbind` – reprezintă funcția care va genera un tabel cu frecvența absolută a fiecărei valori.

3.2.2 Frecvența cumulată

Așa cum am procedat în cazul frecvenței absolute, vom genera un obiect (**fcumconst**) care să primească rezultatele frecvențelor cumulate.

```
fcumconst <- cumsum(faconst)
cbind(faconst, fcumconst)
```

- cumsum – este funcția necesară pentru a calcula frecvențele cumulate.
- faconst – reprezintă frecvențele absolute ale variabilei conștiinciozitate, care vor fi adunate pentru a calcula frecvențele cumulate.
- cbind – va genera un tabel care va conține atât frecvențele absolute, cât și frecvențele cumulate ale variabilei conștiinciozitate.

3.2.3 Frecvența relativă

Și în cazul frecvenței relative trebuie creat un obiect care să primească rezultatele acestui tip de frecvență. Vom denumi acest obiect **frconst**. Reamintim faptul că frecvența relativă se referă la probabilitatea de apariție a fiecărei valori din distribuție.

```
frconst <- (faconst/50)
cbind(faconst, fcumconst, frconst)
```

- frconst – este obiectul care conține frecvențele relative ale variabilei conștiinciozitate.
- faconst/50 – este formula în funcție de care se calculează frecvența relativă; faconst este frecvența absolută, iar 50 indică numărul total de participanți.
- cbind – generează tabelul care cuprinde frecvențele absolute, frecvențele cumulate și frecvențele relative.

3.2.4 Frecvența relativă procentuală

Frpconst este obiectul pe care îl generăm pentru a primi frecvențele procentuale ale variabilei conștiinciozitate.

```
frpconst <- (faconst/50)*100
cbind(faconst, fcumconst, frconst, frpconst)
```

- frpconst – este obiectul care conține frecvențele procentuale ale variabilei analizate.
- (faconst/50)*100 - este formula în funcție de care se calculează frecvența procentuală (vezi formula 3.1).
- cbind – generează tabelul care cuprinde frecvențele absolute, frecvențele cumulate, frecvențele relative și frecvențele relative procentuale.

3.2.5 Frecvența relativă cumulată procentuală

Frcpconst este obiectul generat pentru a primi frecvențele relative cumulate procentuale ale variabilei conștiinciozitate.

```
frcpconst<-cumsum(frpconst)
cbind(faconst, fcumconst, frconst, frpconst, frcpconst)
```

- cumsum – este funcția necesară pentru a calcula frecvențele cumulate.
- frpconst – reprezintă frecvențele relative procentuale ale variabilei conștiinciozitate, care vor fi adunate pentru a calcula frecvențele cumulate procentuale.

- cbind – generează tabelul care cuprinde frecvențele absolute, frecvențele cumulate, frecvențele relative, frecvențele relative procentuale și frecvențele relative cumulate procentuale.

1	### Analiza de frecvențe ###				
2					
3	library(psych)				
4					
5	# Frecvența absolută				
6	faconst<-table(bdcap3\$constiinciozitate)				
7	cbind(faconst)				
8					
9	# Frecvența cumulată				
10	fcumconst <- cumsum(faconst)				
11	cbind(faconst, fcumconst)				
12					
13	# Frecvența relativă				
14	frconst <- (faconst/50)				
15	cbind(faconst, fcumconst, frconst)				
16					
17	#Frecvența relativă procentuală				
18	frpconst <- (faconst/50)*100				
19	cbind(faconst, fcumconst, frconst,frpconst)				
20					
21	#Frecvența relativă cumulată procentuală				
22	frcpconst<-cumsum(frpconst)				
23	cbind(faconst, fcumconst, frconst,frpconst,frcpconst)				
	faconst	fcumconst	frconst	frpconst	frcpconst
45	2	2	0.04	4	4
48	2	4	0.04	4	8
51	4	8	0.08	8	16
52	3	11	0.06	6	22
53	1	12	0.02	2	24
54	2	14	0.04	4	28
55	2	16	0.04	4	32
57	3	19	0.06	6	38
59	3	22	0.06	6	44
61	3	25	0.06	6	50
63	2	27	0.04	4	54
64	2	29	0.04	4	58
65	1	30	0.02	2	60
66	3	33	0.06	6	66
67	3	36	0.06	6	72
68	1	37	0.02	2	74
70	1	38	0.02	2	76
71	4	42	0.08	8	84
72	2	44	0.04	4	88
73	1	45	0.02	2	90
74	2	47	0.04	4	94
75	2	49	0.04	4	98
76	1	50	0.02	2	100

Rezultatele prezentate în imaginea de mai sus ne oferă informații numai cu privire la frecvențele valorilor existente în cadrul distribuției variabilei conștiinciozitate, fără a ne prezenta în mod direct percentilele speciale. În acest moment le putem estima pe baza rotunjirii frecvențelor cumulate procentual. Mai exact, analizăm care sunt rangurile percentile cele mai apropiate de 25, 50, respectiv 75 și atribuim valorilor asociate titulatura de **quartila 1 (Q1)**, **quartila 2 (Q2)**, respectiv **quartila 3 (Q3)**.

În R putem obține afișarea percentilelor speciale folosind liniile de cod de mai jos:

```
quartconst <- quantile(bdcap3$conștiinciozitate, probs = c(0.25, 0.50, 0.75))
quartconst
```

- `quartconst` – este obiectul creat pentru a primi valorile corespunzătoare celor trei quartile.
- `bdcap3$conștiinciozitate` – reprezintă numele bazei de date, respectiv variabila analizată.
- `probs` – indică cele trei valori corespunzătoare rangurilor percentile.
- `c` – permite cumularea mai multor valori în aceeași funcție.

După rularea celor două linii de cod vom putea observa că valoarea corespunzătoare quartilei 1 este 54, scorul 62 corespunde quartilei 2, iar scorul 69.5 corespunde quartilei 3.

25%	50%	75%
54.0	62.0	69.5

În cazul în care dorim să calculăm alte ranguri percentile putem utiliza liniile de cod de mai sus, modificând doar valorile probabilităților. Astfel, dacă dorim să calculăm percentilele 22, 60, 72 și 84 linia de cod devine:

```
quartconst <- quantile(bdcap3$conștiinciozitate, probs = c(0.22, 0.60, 0.72, 0.84))
quartconst
```

22%	60%	72%	84%
52.78	65.40	67.28	71.16

3.3 Reprezentări de tip grafic

Este foarte cunoscută acea maximă care spune că „o poză face cât o mie de cuvinte”. Atunci când ne referim la valori numerice expresia devine „un grafic face cât o mie de numere”. În funcție de tipul variabilei măsurate vom selecta graficul prin care vom prezenta datele. Majoritatea graficelor au două axe: **X** (abscisa) și **Y** (ordonata). Pe **axa X** sunt definite valorile variabilei analizate, iar pe **axa Y** ne este indicată frecvența fiecărei valori a variabilei respective.

Să ne imaginăm că cei 50 de studenți care au completat chestionarul de conștiinciozitate au răspuns și la întrebarea „În ce domeniu doriți să lucrați după absolvirea facultății?”, opțiunile fiind: 1) Psihoterapie, 2) Psihologie organizațională, 3) Securitate națională și 4) Psihologie

educațională. Astfel, 17 studenți au optat pentru psihoterapie, 11 pentru psihologie organizațională, 13 pentru securitate națională, iar 9 pentru psihologie educațională.

Graficul circular (de tip plăcintă) împarte datele în grupuri distincte sau în categorii. Acest grafic constă într-un cerc împărțit în felii, fiecare dintre ele reprezentând o anumită categorie. Mărimea fiecărei felii este dată de proporția sau frecvența fiecărui grup. Cu cât o categorie are o frecvență mai mare, cu atât felia ocupă o zonă mai mare din cerc. Atunci când se folosesc frecvențe procentuale prin adunarea tuturor frecvențelor se va obține 100%, iar în situația în care sunt utilizate frecvențele absolute prin însumare se va obține numărul total de persoane chestionate.

Pentru a afișa graficul este necesară pregătirea datelor. Astfel, trebuie să generăm în R un obiect care să conțină opțiunile studenților. Vom denumi acest obiect sugestiv **optiuni**. Deoarece fiecărui domeniu îi este atribuit un număr este necesar să facem legătura între denumirea domeniului și numărul alocat (1 = Psihoterapie, 2 = Psihologie organizațională, 3 = Securitate națională și 4 = Psihologie educațională). Prin urmare vom crea un obiect, **etichete**, care va conține această corespondență.

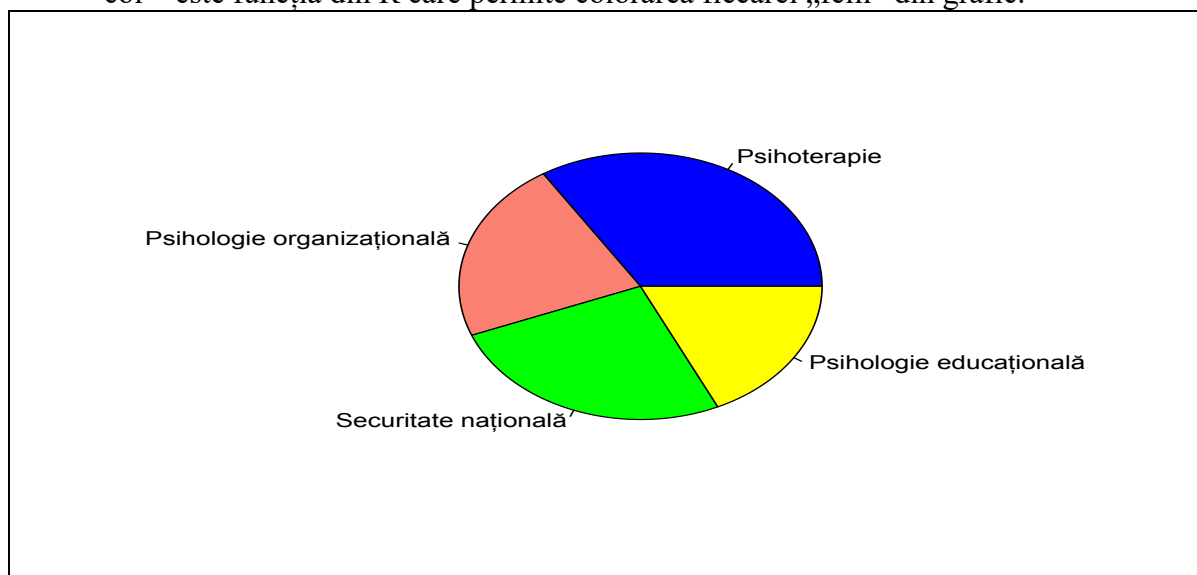
```
optiuni <- c(17, 11, 13, 9)
```

```
etichete<-c("Psihoterapie", "Psihologie organizațională", "Securitate națională", "Psihologie educațională")
```

Odată ce am pregătit datele necesare, putem scrie linia de cod care să ne afișeze graficul de tip plăcintă.

```
pie(optiuni, etichete, col = c("blue", "salmon", "green", "yellow"))
```

- pie – reprezintă funcția destinată graficului de tip plăcintă
- optiuni – este obiectul în care sunt cuprinse răspunsurile studenților
- etichete – este obiectul care face corespondența dintre valorile numerice și eticheta lor
- col – este funcția din R care permite colorarea fiecărei „felii” din grafic.



Graficul de tip bară este utilizat pentru a prezenta frecvența unor variabile categoricale sau calitative. În cazul acestui grafic se va lăsa un spațiu, acesta fiind semnul convențional care

ne indică faptul că variabilele sunt de ordin calitativ sau categorial. Spațiul lăsat între barele graficului evidențiază că nu există nici o legătură între valorile variabilei. Printre variabile care pot fi reprezentate într-un grafic de tip bară sunt: genul, starea civilă, nivelul de studii etc. Înălțimea fiecărei bare exprimă frecvența de apariție a acelei valori în cadrul distribuției. Cu cât o bară este mai înaltă, cu atât frecvența acelei valori este mai mare.

Graficul de tip bară este similar unui pahar cu apă. Când nivelul lichidului este ridicat afirmăm că paharul este plin, că are multă apă. Folosind același principiu, o bară înaltă ne indică o frecvență mare de apariție a unei valori în cadrul distribuției.

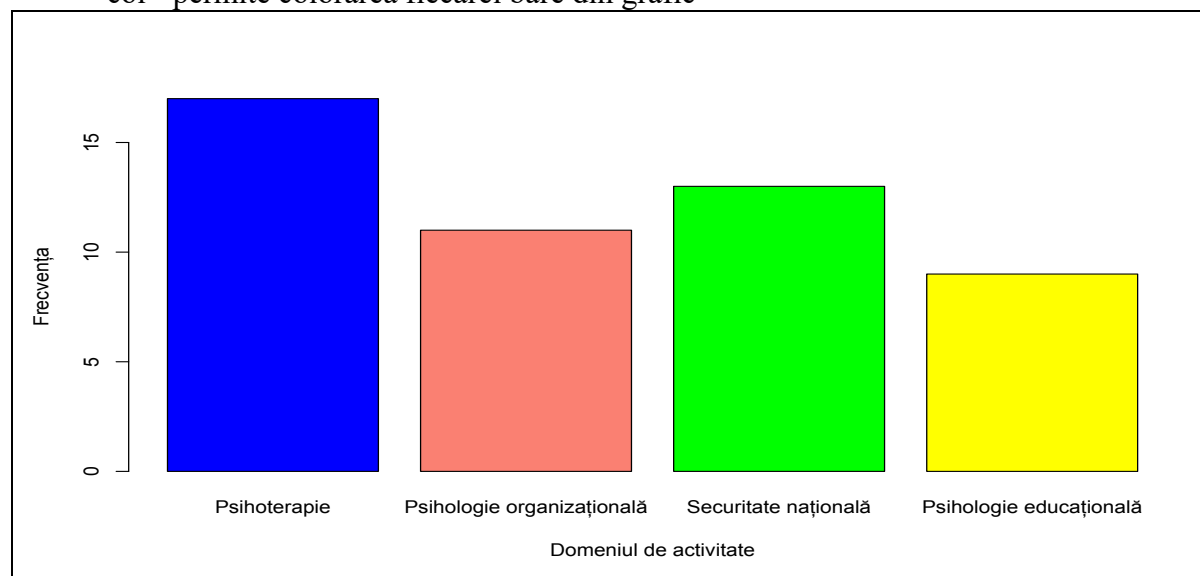
Frecvența =



Pentru graficul de tip bară putem folosi datele pregătite în pasul anterior. Mai jos este prezentată linia de cod necesară pentru a afișa graficul de tip bară.

```
barplot(optiuni, names.arg=etichete, xlab="Domeniul de activitate",
        ylab="Frecvența", col = c("blue", "salmon", "green", "yellow"))
```

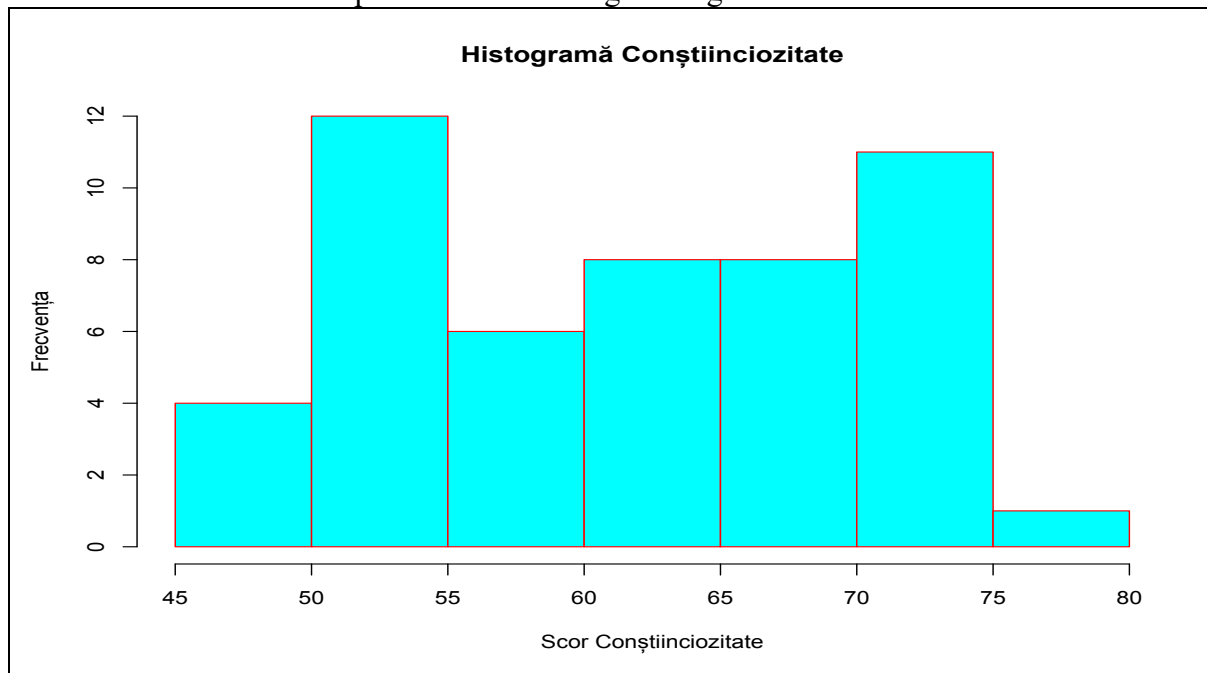
- barplot – reprezintă funcția din R care generează graficul de tip bară
- optiuni – face trimitere la obiectul în care sunt cuprinse răspunsurile studenților
- names.arg – este un vector care va permite etichetarea fiecărei bare din grafic
- xlab – permite denumirea axei X
- ylab – permite denumirea axei Y
- col - permite colorarea fiecărei bare din grafic



Histograma este un alt grafic utilizat pentru date cantitative. Fiecărui scor îi este atribuită o bară a cărei înălțime variază în funcție de frecvența absolută a scorului respectiv. Spre deosebire de graficul de tip bară, în cazul histogramei nu există spații între bare. Acest aspect ne arată caracterul continuu al valorilor din distribuție.

```
hist(bdcap3$conștiinciozitate, xlab="Scor Conștiinciozitate", ylab="Frecvența",  
col="cyan", border="red", main="Histogramă Conștiinciozitate")
```

- hist – funcția necesară pentru a genera graficul de tip histogramă
- bdcap3\$conștiinciozitate – reprezintă baza de date și variabila analizată
- xlab – permite denumirea axei X
- ylab – permite denumirea axei Y
- col – funcția necesară pentru a colora graficul
- border – permite colorarea conturului histogramelor
- main – este utilizat pentru a da un titlu general graficului



Exerciții

În Statistics City se organizează în fiecare an competiția *Standard Deviation Marathon*. La înscriere participanții sunt rugați să specifice vârsta și numărul de zile alocate antrenamentelor, datele fiind sintetizate în tabelul de mai jos:

Vârsta					Zile de antrenament				
63	21	58	27	27	18	21	21	23	24
22	36	38	47	18	23	18	15	16	22
31	33	19	25	49	23	15	15	24	24
63	30	56	19	43	23	17	18	24	17
34	64	59	38	30	21	24	25	19	17
42	42	21	47	22	17	18	16	18	21
50	37	40	65	18	22	18	18	17	25
29	20	25	45	24	24	21	18	21	24
54	36	60	35	46	21	23	19	22	18
28	36	35	26	22	18	20	25	22	19

Realizați o bază de date cu cele două variabile folosind datele din tabel și apoi rezolvați exercițiile de mai jos:

- Pentru variabila **vârstă** precizați:
 - Frecvența absolută pentru 20, 54, 61 și 64.
 - Frecvența procentuală pentru 17, 23, 35 și 47.
 - Precizați frecvența cumulată procentual pentru 38, 44, 50 și 58. Cum interpretăm frecvențele cumulate procentual ale acestor scoruri?
 - Precizați percentilele 25, 50 și 75.
- Pentru **variabila zile de antrenament** realizați graficul de tip box histogramă.
- Răspundeți la următoarele întrebări:
 - Care este obiectivul analizei de frecvențe?
 - Ce este frecvența absolută? Dar frecvența cumulată procentual?
 - Cum se mai numește frecvența cumulată procentual?
 - În ce situații folosim graficul de tip bară sau graficul circular?
 - Care sunt graficele recomandate în cazul variabilelor cantitative?