

Fizică experimentală

Laborator 2

Victor E. Ambruș

*Facultatea de fizică, Universitatea de Vest din Timișoara,
Bd. Vasile Pârvan nr. 4, Timișoara, RO 300223, România*

7 noiembrie 2016

Rezumat

Scopul acestui laborator este de a oferi o introducere în utilizarea suitei R-Studio (având la bază limbajul R) pentru prelucrarea datelor statistice. Acest tip de prelucrări este foarte util pentru analiza rezultatelor experimentale obținute prin măsurători la disciplinele la care se fac lucrări de laborator.

1 Instalarea

Instalarea se face în două etape:

1. Pachetul de bază R se instalează de la adresa <https://cran.rstudio.com/>;
2. Suita R-Studio se instalează de la adresa <https://www.rstudio.com>.

2 Import - citirea datelor

În R se pot importa datele în diferite formate. Pentru acest laborator, vom folosi un fișier de tip text, în care datele sunt așezate în felul următor:

1. Pe prima linie avem antetul celor trei coloane: *Înălțime (m)* (pentru axa x), *Masă (kg)* (pentru axa y) și *Pondere*, separate folosind TAB.
2. Pe fiecare linie ulterioară vom avea câte trei valori de tip (înălțime, masă, pondere), separate prin TAB.

Importarea datelor se poate face în oricare din următoarele moduri:

1. utilizând butonul *Import Dataset* (vezi Fig. 1);
2. utilizând meniul: Tools → Import Dataset → From Local File.

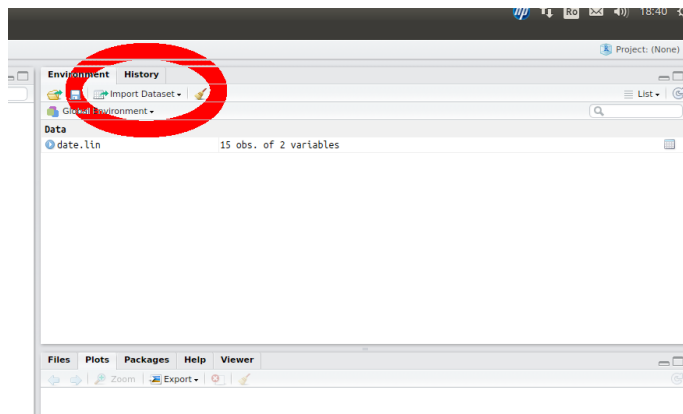


Figura 1: Butonul pentru importarea datelor din fișier este evidențiat cu roșu.

Inaltimea(m)	Masa(kg)	Pondere
1.47	52.21	1
1.5	53.12	1
1.52	54.48	2
1.55	55.84	2
1.57	57.2	3
1.6	58.57	6
1.63	59.93	8
1.65	61.29	6
1.68	63.11	4
1.7	64.47	4
1.73	66.28	6
1.75	68.1	8
1.78	69.92	6
1.8	72.19	4
1.83	74.46	2

Tabela 1: Date privind corelația dintre înălțime și greutate. Ultima coloană conține ponderea asociată fiecărei măsurători.

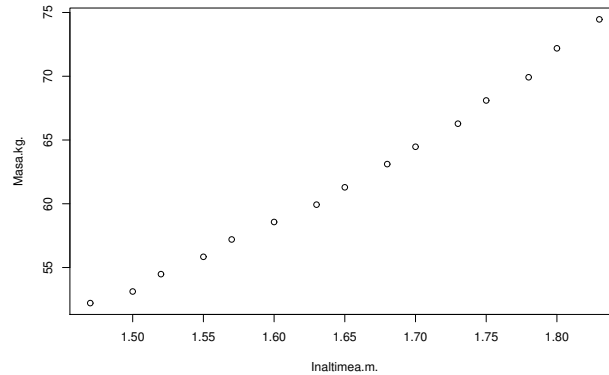


Figura 2: Reprezentarea grafică a variabilei `date.lin`

3. Folosind comanda

```
date.lin <- read.delim(
  "~/Documents/uvvt/cursuri/fizexp/lab2/date-lin.txt",
  comment.char="#")
```

3 Reprezentarea grafică

Fișierul `date-lin.txt` se citește în mod automat în variabila `date.lin`. Conținutul lui `date.lin` se poate vizualiza în fereastra **Environment**. În R, tipul acestei variabile este `data.frame`:

```
> class(date.lin)
[1] "data.frame"
```

Pentru a reprezenta grafic o variabilă de tip `data.frame`, se folosește funcție `plot`:

```
> plot(date.lin[,1],date.lin[,2])
```

Rezultatul se poate vedea în Fig. 3. Comanda `plot` poate lua diverși parametri care permit controlul graficului rezultat. Pentru a citi mai multe detalii despre o comandă, se poate utiliza sintaxa `?<comanda>`. În cazul comenzii `plot`, se poate folosi

```
> ?plot
```

4 Operații asupra datelor

Practic, acest tip de date se comportă ca o matrice:

```
> date.lin[4,]  
      Inaltimea.m. Masa.kg. Pondere  
4      1.55      55.84      2  
> date.lin[4,1]  
[1] 1.55  
> date.lin[4,2]  
[1] 55.84  
> date.lin[4,3]  
[1] 2
```

În general, instrucțiunea `date.lin[i,]` va returna linia `i` a setului de date, iar instrucțiunea `date.lin[i,j]` va returna valoarea corespunzătoare coloanei `j` de pe această linie.

Numărul de elemente din `date.lin` se poate afla cu comanda `nrow`:

```
> nrow(date.lin)  
[1] 15
```

Asupra elementelor lui `date.lin` se pot efectua operații uzuale, de tip înmulțire cu scalar, adunarea unei alte matrici, înmulțirea element cu element cu o altă matrice, adăugarea unui element, suprimarea unui element. Vom face aceste operații în ordinea de mai sus, obținând:

- `date2 <- date.lin * 2` crează o nouă variabilă de tip `data.frame` care conține aceleași elemente ca `date.lin`, dar înmulțite cu 2.
- `date3 <- date2 - date.lin`: fiecare element al lui `date3` va fi obținut făcând diferența dintre elementele corespunzătoare din `date2` și `date.lin`.
- `date4 <- date3 * date2`: simbolul `*` efectuează înmulțirea dintre operandi la nivel de element, adică `date4[i,j] = date3[i,j] * date2[i,j]`.
- `date5<-rbind(date3, c(2,100,3))` salvează în `date5` informația din `date3`, la care se adaugă elementele 2, 100 și 3, concatenate într-un vector folosind comanda `c`.
- `date6<-date3[c(1:6,8:nrow(date3)),]`: Această instrucțiune salvează în `date6` doar elementele din 3 de pe pozițiile 1–6 și 8–până la capătul vectorului.

5 Procesarea datelor

R reprezintă un limbaj de programare dezvoltat pentru prelucrări statistice. De aceea, R oferă nenumărate modalități de a procesa datele experimentale, care sunt foarte bine documentate atât în R în sine, cât și pe internet. Cu ocazia acestui laborator, vom efectua următoarele operații:

5.1 Media aritmetică

Pentru ușurință, ne vom referi la coloanele setului de date `date.lin` prin x_i (înălțimile), y_i (masele) și w_i (ponderile). Media lui x_i și a lui y_i se calculează după cum urmează:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

Acste operațiuni se fac în R folosind comanda `mean`:

```
> print(mx<-mean(date.lin[,1]))
[1] 1.650667
> print(my<-mean(date.lin[,2]))
[1] 62.078
```

Rezultatul a fost salvat în variabilele `mx` și `my`.

5.2 Media ponderată

Din moment ce fișierul nostru de date conține și o coloană de ponderi, se pot defini mediile ponderate ale lui x și y , după cum urmează:

$$\bar{x}_w = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N p_i}, \quad \bar{y}_w = \frac{\sum_{i=1}^N w_i y_i}{\sum_{i=1}^N p_i}.$$

În R, media ponderată se calculează folosind comanda `weighted.mean`:

```
> print(wmx<-weighted.mean(date.lin[,1],date.lin[,3]))
[1] 1.679365
> print(wmy<-weighted.mean(date.lin[,2],date.lin[,3]))
[1] 63.59
```

unde elementele de pe coloana a treia a lui `date.lin` au fost folosite ca și ponderi.

5.3 Abaterea

Abaterea se definește cu formula:

$$y'_i = y_i - \bar{y},$$

unde \bar{y} poate fi media aritmetică sau media ponderată a variabilei y . În acest exemplu, vom folosi media ponderată. Putem construi setul de date corespunzător abaterii scăzând din `date.lin` elementul $(0, \bar{y})$:

```
> abatere<-sweep(date.lin,2,c(0,wmy,0))
```

5.4 Momente absolute

Momentul absolut de ordin k al variabilei x se definește ca:

$$M(x^k) = \frac{\sum_{i=1}^N w_i x_i^k}{\sum_{i=1}^N w_i}.$$

Pentru a afla momentul absolut de ordinul 2 al lui x , se poate folosi următoarea comandă:

```
> wmx2<-weighted.mean(date.lin[,1]^2,date.lin[,3])
```

Operatorul de ridicare la putere acționează asupra fiecărui element din `date.lin`.

Dintre cazurile particulare ale momentelor absolute de ordin k amintim:

- Media armonică $M_H = [M(x^{-1})]^{-1}$, corespunzătoare lui $k = -1$, care se obține cu comanda:

```
> print(wmxh<-(weighted.mean(date.lin[,1]^(-1),date.lin[,3]))^(-1))  
[1] 1.674746
```

- Media ponderată $M_A = M(x)$, corespunzătoare lui $k = 1$;
- Media pătratică $M_P = \sqrt{M(x^2)}$, corespunzătoare lui $k = 2$:

```
> print(wmxxp<-(weighted.mean(date.lin[,1]^2,date.lin[,3]))^(1/2))  
[1] 1.681632
```

5.5 Momente centrate

Momentul centrat de ordin k al variabilei x se definește ca fiind media lui $(x - \bar{x})^k$:

$$M[(x - \bar{x})^k] = \frac{\sum_{i=1}^N w_i (x_i - \bar{x})^k}{\sum_{i=1}^N w_i}.$$

Din definiție, se poate vedea că momentul centrat de ordinul 1 se anulează. Un caz particular extrem de important este dat de momentul centrat de ordinul 2, care are semnificația de **dispersie** a variabilei x :

$$D(x) = M[(x - \bar{x})^2].$$

Dispersia se obține în R folosind comenzile:

```
> print(dispx<-weighted.mean((date.lin[,1]-wmx)^2,date.lin[,3]))  
[1] 0.007618644  
> print(dispy<-weighted.mean((date.lin[,2]-wmy)^2,date.lin[,3]))  
[1] 30.48511
```

Abaterea medie pătratică, sau deviația standard, se definește ca:

$$\sigma(x) = \sqrt{D(x)} = \sqrt{x^2 - \bar{x}^2},$$

și poate fi calculată în R folosind comanda:

```
> print(devstx<-sqrt(dispx))
[1] 0.08728485
> print(devsty<-sqrt(dispy))
[1] 5.521332
```

Coeficientul de variație reprezintă raportul dintre deviația standard și media variabilei:

$$V(x) = \frac{\sigma(x)}{\bar{x}},$$

valoarea acestuia pentru cazul curent fiind:

```
> print(variatiax<-devstx/wmx)
[1] 0.05197491
> print(variatiay<-devsty/wmy)
[1] 0.08682705
```

5.6 Corelația variabilelor

În fine, să studiem corelația dintre variabilele x (înălțimea) și y (masa). Aceasta se definește ca media produsului dintre cele două variabile, raportate la valoarea lor medie:

$$C_{xy} = M[(x - \bar{x})(y - \bar{y})] = \overline{xy} - \bar{x}\bar{y}.$$

Pentru cazul nostru, C_{xy} are valoarea:

```
> print(corxy<-weighted.mean(date.lin[,1]*date.lin[,2],
                             date.lin[,3])-wmx*wmy)
[1] 0.4792206
```

Cu ajutorul corelației, se poate defini **coeficientul de corelație**:

$$\rho = \frac{C_{xy}}{\sigma(x)\sigma(y)},$$

unde $\sigma(x)$ și $\sigma(y)$ reprezintă deviațiile standard ale variabilelor x și y . Pentru cazul curent obținem:

```
> print(coefcor<-corxy/devstx/devsty)
[1] 0.9943809
```