Curs 6 – Verificarea datelor parametrice

Lect. Adrian Gorbănescu

Cele mai multe din procedurile statistice descrise în această carte reprezintă **teste parametrice**, care se bazează pe o distribuție normală (vezi capitolul 5.2). A utiliza un test parametric atunci când datele sunt neparametrice înseamnă a obține niște rezultate care, cel mai probabil, sunt incorecte. Astfel, este foarte important să verificăm îndeplinirea condițiilor necesare pentru aplicarea testelor parametrice înainte de a aplica testul pe care noi îl considerăm potrivit. În cazul în care nu sunt îndeplinite condițiile pentru date parametrice, vom aplica un **test neparametric**. Cele mai multe dintre testele parametrice necesită îndeplinirea a patru condiții. Mulți dintre studenți percep verificarea acestor condiții ca pe o acțiune obositoare și, de multe ori, devin confuzi atunci când trebuie să decidă dacă condițiile sunt îndeplinite sau nu. Cele patru condiții sunt:

1. condiția de normalitate
2. omogenitatea variațelor
3. datele trebuie măsurate pe scală de interval/raport (vezi secțiunea 1.5.1)
4. condiția de independență - se referă la faptul că atunci când măsurăm comportamentul unui participant acesta nu este influențat de comportamentul altui participant. Să ne imaginăm că avem două persoane care participă la un experiment în care trebuie să indice cuvintele pe care și le amintesc dintr-o listă prezentată în urmă cu 10 minute. Dacă unei persoane i se permite să asiste la enumerarea cuvintelor celeilalte persoane, atunci răspunsurile ei ar fi influențate de cuvintele auzite în momentul expunerii primului participant.

Un cercetător este interesat să studieze rolul de predictor al trăsăturilor de personalitate asupra comportamentul sexual riscant în rândul tinerilor. În acest sens aplică o serie de chestionare pe un eșantion de 1200 de participanți, iar datele obținute sunt înregistrate în baza de date bdcap6.xlsx.

6.1. Condiția de normalitate

În multe teste statistice (de exemplu, testele t) presupunem că distribuția de eșantionare este distribuită normal. Deoarece nu avem acces la această distribuție, ci la un set de date culese pe un eșantion selectat de noi, nu ne putem uita pur și simplu la forma distribuției și să decidem dacă este normală sau nu. Cu toate acestea, cunoaștem din teorema limitei centrale (vezi capitolul 5.5) că pe măsură ce un eșantion este suficient de mare (cel puțin 30 de participanți) distribuția lui se apropie de curba normală.

Când o distribuție îndeplinește condiția de normalitate?

Există mai multe modalități de a verifica dacă o distribuție îndeplinește condiția de normalitate și ele vor fi prezentate mai jos.

6.1.1 Skewness și Kurtosis

Atunci când skewness și kurtosis au valori cuprinse între **-1** și **1** distribuția este normală. Dacă valorile indicilor skewness și kutosis sunt în afara acestui interval, putem alege transformarea lor în scoruri z. Această transformare se face prin împărțirea scorului skewnees, respectiv kurtosis la valoarea erorii standard a indicatorului.

$$z_{skewness} = \frac{skewness}{eroare\ standard_{skewness}}$$

(formula 6.1)

$$z_{kurtosis} = \frac{kurtosis}{eroare\ standard_{kurtosis}}$$

(formula 6.2)

În continuare vom testa normalitatea distribuției formată din scorurile obținute pentru comportamentul sexual riscant (*SR*). Pentru a calcula indicii skewness și kurtosis vom folosi funcțiile **skew()**, respectiv **kurtosi**(). Astfel, se poate observa că valoarea lui skewness este 2.53, iar cea a lui kurtosis este aproximativ 9.68, valori care depășesc mult limitele intervalului [-1 → 1] și care se traduc într-o încălcare a condiției de normalitate.

```
skew(bdcap6$SR)
kurtosi(bdcap6$SR)
```

```
2.534257
9.675366
```

Așa cum a fost menționat mai sus, atunci când indicii skewness și kurtosis depășesc intervalul [-1 → 1] putem testa normalitatea prin intermediul scorurilor zskewness și zkurtosis. Pentru a calcula scorul zkewness vom folosi funcția **se.skew()** din pachetul **sur**, iar rezultatul indică faptul că eroarea standard a lui skewness este 0.07.

se.skew(bdcap6$SR), unde:

- se.skew este funcția care calculează eroarea standard a lui skewness;
- bdcap6 reprezintă baza de date;
- SR este variabila analizată.

Pentru a calcula eroarea standard a lui kurtosis vom folosi formula de mai jos:

$$se_{kurtosis} = \sqrt{\frac{24}{N}}$$

(formula 6.3)

- 24 reprezintă o constantă;
- N indică numărul de participanți.

Softul R poate fi folosit pentru a calcula eroarea standard a lui kurtosis în baza liniei de cod prezentată mai jos, iar rezultatul obținut ne va arată că eroarea standard a lui kurtosis este egală cu 0.14.

```
sekurtosis <- sqrt(24/1200)
sekurtosis
```

- sekurtosis reprezintă un obiect care va primi rezultatul operației matematice din formula 6.3;
- sqrt este funcția pentru radical;

- 24 este o constantă;
- 1200 se referă la numărul de participanți.

```
se.skew(bdcap6$SR)
```
```
0.0706225
```
```
sekurtosis <- sqrt(24/1200)
sekurtosis
```
```
0.1414214
```

Folosind formulele de mai sus, vom observa că zskewness este egal cu 36.14, în timp ce zkurtosis este egal cu 69.14.

```
zskewness <- 2.53/0.07
zskewness
```
```
36.14286
```
```
zkurtosis <- 9.68/0.14
zkurtosis
```
```
69.14286
```

Scorurile *z* obținute vor fi comparate cu valorile care corespund probabilității ca distribuția să aibă această formă ca urmare a întâmplării, șansei (Anexa 1). Astfel, atunci când scorul z (se interpretează în valoare absolută, în modul) are o valoare mai mare de 1,96 este semnificativ la $p < 0.05$; când este mai mare decât 2,58 este semnificativ la $p < 0,01$; când este mai mare decât 3,29 este semnificativ la $p < 0,001$. Cu alte cuvinte, distribuția este normală atunci când scorurile z pentru skewnees și kurtosis sunt cuprinse între -1,96 și 1,96 sau între -2,58 și 2,58 sau între -3,29 și 3,29. Intervalul va fi selectat în funcție de mărimea eșantionului. Astfel, în cazul unui eșantion mic vom alege valoarea 1,96, pentru eșantioanele mari vom crește la 2,58, iar în cazul eșantioanelor foarte mari, din cauza faptului că eroarea standard scade (vezi capitolul 5.4) nu se mai ține cont de nici un criteriu. De asemenea, în cazul eșantioanelor foarte mari (200 de participanți sau peste) distribuția va fi considerată normală ca urmare a teoremei limitei centrale și se recomandă analiza grafică a distribuției pentru vedea dacă aceasta îndeplinește condiția de normalitate.

În exemplul rezultatelor obținute de eșantionul de studenți la examenul de admitere (N = 179), $z_{skewness} = 0,060$, iar $z_{kurtosis} = -2,875$. Deoarece $z_{kurtosis}$ este în afara intervalului -2,58 și 2,58 vom considera că distribuția are o ușoară abatere de la condiția de normalitate. Totuși, nu trebuie să pierdem din vedere faptul că eșantionul este format din 179 de participanți.

6.1.2 Aplicarea testelor de normalitate

O altă metodă de a verifica dacă o distribuție îndeplinește condiția de normalitate este aceea de a analiza dacă ea diferă semnificativ statistic față de curba normală. Cele mai cunoscute teste pentru verificarea acestei condiții sunt **Kolmogorov-Smirnov** și **Shapiro-Wilk**. Aceste teste compară scorurile obținute la nivel de eșantion cu un set de scoruri distribuit normal care au aceeași medie și abatere standard. Dacă testul este nesemnificativ statistic ($p > 0,05$) înțelegem că distribuția nu diferă semnificativ de distribuția normală. Dacă testul este semnificativ statistic ($p < 0,05$) distribuția analizată diferă semnificativ de distribuția normală și, cel mai probabil, nu îndeplinește condiția de normalitate.

Deși aceste teste verifică foarte ușor condiția de normalitate, nu trebuie să neglijăm faptul că ele au o serie de limite: la eșantioane mari este foarte ușor să obținem teste semnificative statistic, chiar și atunci când distribuția se abate foarte puțin de la curba normală. Cu alte

cuvinte, testele de normalitate pot respinge condiția de normalitate, chiar dacă, în realitate, distribuția îndeplinește această condiție.

Dintre aceste două proceduri vom alege spre exemplificare testul Shapiro-Wilk deoarece oferă rezultate mai valide comparativ cu testul Kolmogorov-Smirnov. Testul Shapiro-Wilk poate fi aplicat prin intermediul funcției **shapiro.test()**. Rezultatele obținute ne indică faptul că distribuția nu îndeplinește condiția de normalitate, deoarece *p* este mai mic decât pragul de semnificație statistică 0.05.

shapiro.test(bdcap6$SR), unde:

- shapiro.test este funcția asociată testului Shapiro-Wilk;
- bdcap6 reprezintă baza de date;
- SR este variabila analizată.

```
shapiro.test(bdcap6$SR)

        Shapiro-Wilk normality test

data:  bdcap6$SR
W = 0.783, p-value < 2.2e-16
```

Tabelul **Shapiro-Wilk normality test** ne prezintă rezultatele Printre rezultate se regăsesc următorii indicatori:

- *W* – reprezintă valoarea calculată a testului și nu o vom folosi în stabilirea îndeplinirii condiției de normalitate. În cazul nostru *W* este egal cu 0.78.
- *p value* – indică probabilitatea asociată testului de normalitate. De valoarea lui *p* vom ține cont în luarea deciziei îndeplinirii condiției de normalitate. Dacă $p > 0.05$ acceptăm ipoteza de nul (H_0) și vom afirma că distribuția îndeplinește condiția de normalitate. Dacă $p \leq 0{,}05$ respingem ipoteza de nul și afirmăm că distribuția nu îndeplinește condiția de normalitate.

În cazul nostru, *p* este mai mic chiar decât 2.2 e-16 (ceea ce înseamnă 0.00000000000000022) și implicit este mai mic și decât 0.05. Acest rezultat se traduce în faptul că nu putem accepta ipoteza de nul și implicit nici condiția de normalitate.

6.1.1.3 Analiza vizuală a normalității distribuției

În secțiunea 3.4 am descoperit faptul că graficele de tip bară sau histogramele ne pot ajuta în vizualizarea formei distribuției și am prezentat modalitatea de obținere a acestora. În continuare vom observa faptul că aceste grafice pot fi utilizate și pentru a verifica îndeplinirea condiției de normalitate. Mai jos vom afișa graficul de tip histogramă bifând și opțiunea **Display normal curve** pentru rezultatele obținute de cei 179 de studenți la examenul de admitere.

Un alt grafic util în a analiza dacă distribuția îndeplinește condiția de normalitate este graficul **Q-Q plot**. Acesta compară la nivel grafic probabilitatea cumulativă a unei variabile cu probabilitatea cumulativă a unei distribuții specifice (în cazul nostru distribuția normală). Acest lucru se traduce în faptul că fiecărei valori i se atribuie un rang (lb. engleză, **rank**) și apoi valorile sunt sortate. Pentru fiecare rang se calculează scorul z corespunzător. Aceasta va fi valoarea așteptată pe care scorul ar trebui să o aibă în distribuția normală. În grafic, acest scor z este comparat cu scorul z așteptat. Dacă distribuția este normală, scorul z al valorii va fi

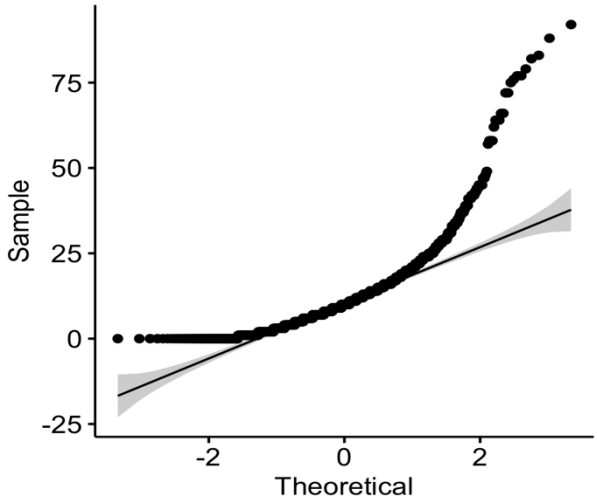
același cu scorul z așteptat și punctele de pe grafic vor forma o diagonală perfectă. Atunci când punctele se îndepărtează față de diagonală, distribuția se abate de la condiția de normalitate.

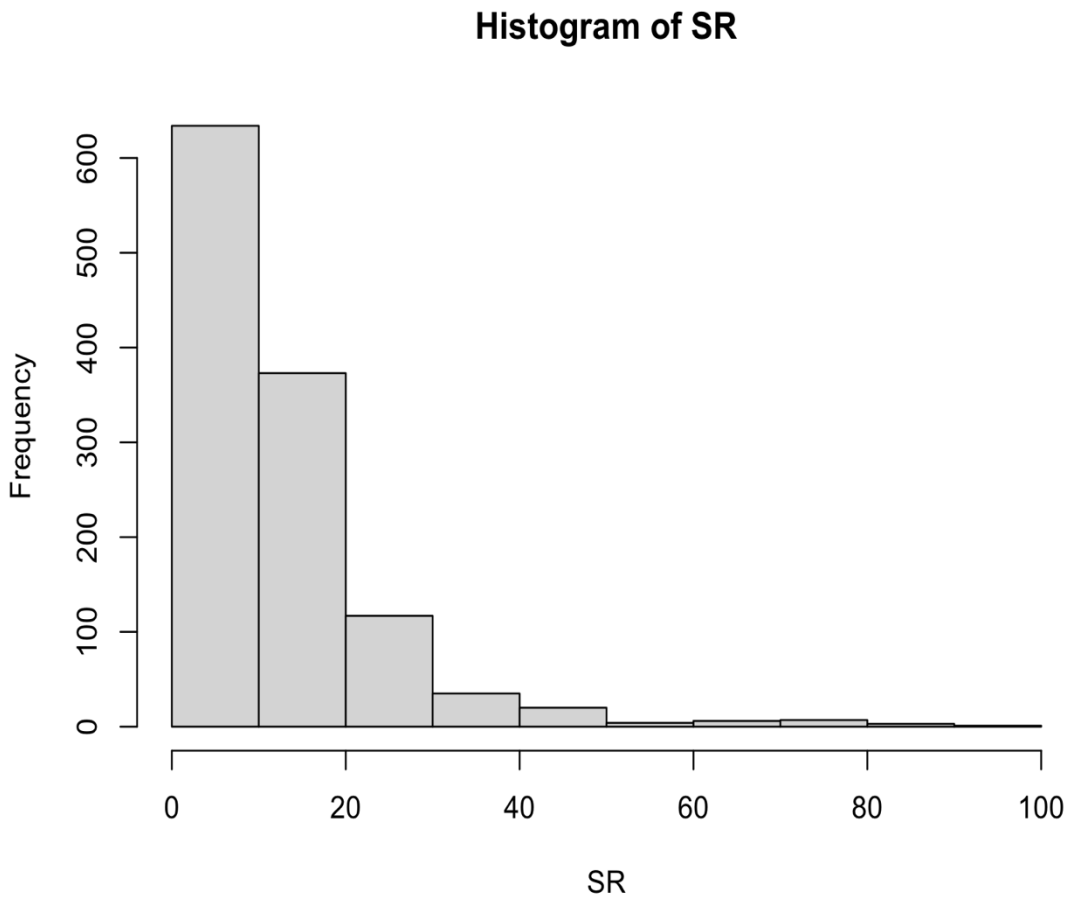
Pentru a obține graficul Q-Q plot vom folosi funcția **ggqqplot()** din pachetul **ggpubr**.

ggqqplot(bdcap6$SR), unde:

- ggqqplot este funcția asociată graficului Q-Q Plot;
- bdcap6 reprezintă baza de date;
- SR este variabila analizată.

```
ggqqplot(bdcap6$SR)
hist(bdcap6$SR, xlab="SR", main="Histogram of SR")
```





Imaginea 6.1 – Graficele de tip Q-Q plot și histogramă

6.1.2 Testarea omogenității variațelor

Odată ce am verificat îndeplinirea condiției de normalitate va trebui să ne concentrăm atenția spre omogenitatea variațelor. Această condiție se referă la faptul că la niveluri diferite ale aceleași variabile, varianța (dispersia) nu trebuie să se modifice semnificativ. Această afirmație sugerează că dacă strângem date ale aceleași variabile din grupuri diferite de persoane, dispersia trebuie să fie aceeași pentru fiecare grup. Să rezumăm aceste explicații cu un exemplu. La examenul de admitere au participat 690 de candidați pe care îi putem împărți în patru grupuri în funcție de regiunile istorice din care provin: Muntenia, Transilvania, Dobrogea și Moldova. Omogenitatea variațelor se referă la faptul că, dacă am calcula dispersia numărului de răspunsuri corecte de la admitere pentru candidații din fiecare grup, ea ar trebui să fie aceeași. Același principiu se aplică dacă am împărți candidații în funcție de profilul liceului absolvit: uman, real, vocațional, tehnologic etc. Imaginea de mai jos ilustrează variațele rezultatelor de la admitere în funcție de profilul liceului absolvit.

Omogenitatea variațelor poate fi verificată cu ajutorul **testului Levene** și este necesar pentru a aplica o serie de teste statistice. În R putem aplica acest test folosind funcția **leveneTest()** din cadrul pachetului **car**.

Să ne imaginăm că ne dorim să testăm omogenitatea variațelor pentru variabila *SR* în funcție de genul participanților. Cu alte cuvinte, dorim să analizăm dacă scorurile comportamentului seuxal de risc obținute pe participanții de gen masculin sunt la fel de diversificate precum cele observate la participanții de gen feminin.

leveneTest(bdcap6$SR, bdcap6$Genul, center=mean), unde:

- leveneTest este funcția asociată testului Levene;
- bdcap6 reprezintă baza de date;
- SR este variabila analizată;
- Genul este variabila categorială în funcție de care se testează omogenitatea variațelor.

```
leveneTest(bdcap6$SR, bdcap6$Genul, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)
group   2  60.294 < 2.2e-16 ***
      1197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabelul **Levene's' Test for Homogeneity of Variances** prezintă rezultatele testului de omogenitate a variațelor.

- *Df* reprezintă numărul de grade de libertate;
- *F* value este valoarea calculată a testului și nu vom ține cont de ea atunci când vom lua decizia cu privire la omogenitatea variațelor;
- *Pr(>F)* este echivalentul lui *p*. La fel ca în cazul normalității, atunci când *p* este mai mare decât 0.05 vom accepta ipoteza de nul și vom concluziona că variațele sunt omogene. În opoziție, dacă *p* este mai mic decât 0.05 respingem ipoteza de nul și nu putem accepta condiția de omogenitate.

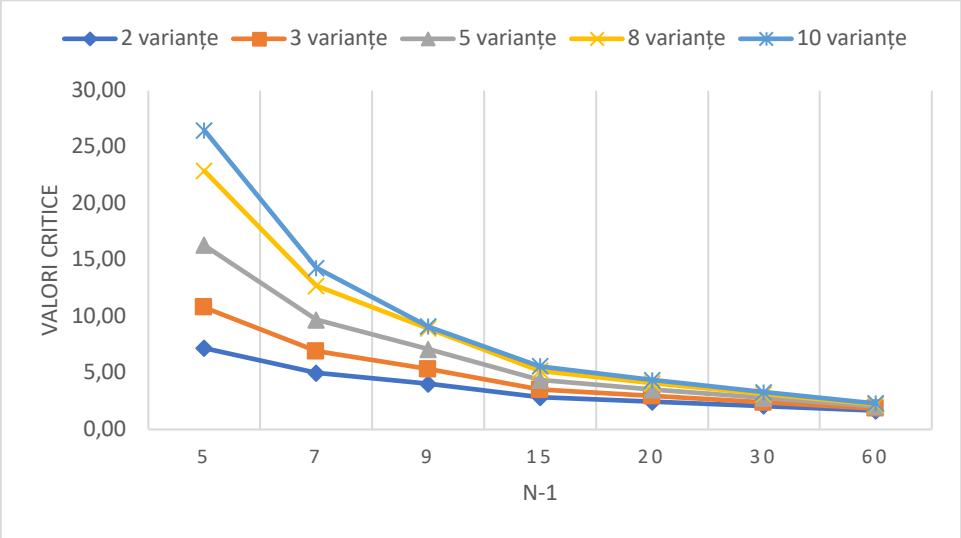
În cazul nostru, *p* este mult mai mic decât 0.05. Reamintim faptul că 2.2e-16 este o prescurtare a valorii 0.00000000000000022. În concluzie, testul Levene respinge omogenitatea variațelor și înțelegem că scorurile participanților de gen masculin nu au o împrăștiere similară cu cea a participanților de gen feminin.

Asemenea testelor de verificare a condiției de normalitate, rezultatele testului Levene pot fi afectate de volumul eșantionului. Astfel, pe măsură ce volumul eșantionului crește, testul Levene respinge omogenitatea variațelor, chiar dacă în realitate variațele grupurilor sunt egale. Acest rezultat este efectul faptului că odată cu creșterea numărului de participanți crește puterea testului (vezi capitolul 6.7). În aceste condiții este necesară o dublă verificare a condiției de omogenitate folosind o procedură cunoscută sub numele de **raportul variațelor** sau **Hartley's F_{Max}** (Pearson & Hartley, 1954). Această procedură presupune a calcula raportul dintre grupul cu varianța cea mai mare și grupul cu varianța cea mai mică, iar rezultatul obținut va fi comparat cu o valoare critică. Valoarea critică depinde de numărul de participanți din fiecare grup și de numărul variațelor comparate. Atunci când raportul variațelor este mai mic decât valoarea critică se consideră că variațele sunt omogene. De exemplu, dacă avem două grupuri a câte 10 participanți, pentru ca variațele să fie omogene raportul obținut trebuie să fie mai mic decât 4,03. În situația cu trei grupuri a câte 10 participanți raportul variațelor va fi comparat cu 5,34.

În R pot fi calculate variațele celor două grupuri folosind funcția **var()**. Astfel, se poate observa că varianța SR pentru participanții de gen masculin este 273.09, iar cea a participanților de gen feminin este 69.68. Dacă facem raportul dintre 273.09 și 69.68 obținem valoarea 3.92, care este mai mică decât 5. Astfel, putem concluziona că variațele sunt omogogene.

```
var(bdcap6$SR [bdcap6$Genul==1])   273.0929
```

```
var(bdcap6$SR [bdcap6$Genul==2])  69.68492
```



Imaginea 6.2 – Valorile critice ale testului Hartley's F_{Max}

6.1.3 Identificarea valorilor extreme

Atunci când am discutat despre media aritmetică am subliniat faptul că aceasta poate fi afectată de existența valorilor extreme, motiv pentru care este imperios necesară identificarea acestora. Cele mai simple metode de identificarea a valorilor extreme sunt reprezentate de graficul Boxplot și transformarea în scoruri standardizate z.

Graficul Boxplot este caracteristic valorilor cantitative și este important în identificarea valorilor extreme. Caseta graficului boxplot cuprinde 50% din valorile distribuției, fiind delimitată de percentila 25 (Q1) și percentila 75 (Q3). Distanța dintre cele două quartile reprezintă ***înălțimea*** graficului (H). Având în vedere că înălțimea graficului este diferența dintre valoarea corespunzătoare quartilei 3 și valoarea corespunzătoare quartilei 1, ea se mai numește **abatere interquartilă**. Linia îngroșată din interiorul casetei reprezintă percentila 50 (Q2). Mustățile graficului (limitele) au o înălțime egală cu **1,5*H**. Orice valoare mai mare sau mai mică decât aceste mustăți reprezintă valori extreme.

Folosind analiza de frecvențe sau funcția **quantile()** aflăm percentilele 25, 50 și 75. Astfel, pentru variabila *SR*, percentila 25 (Q_1) este scorul 5, percentila 50 (Q_2) este scorul 10, iar percentila 75 (Q_3) este scorul 16.

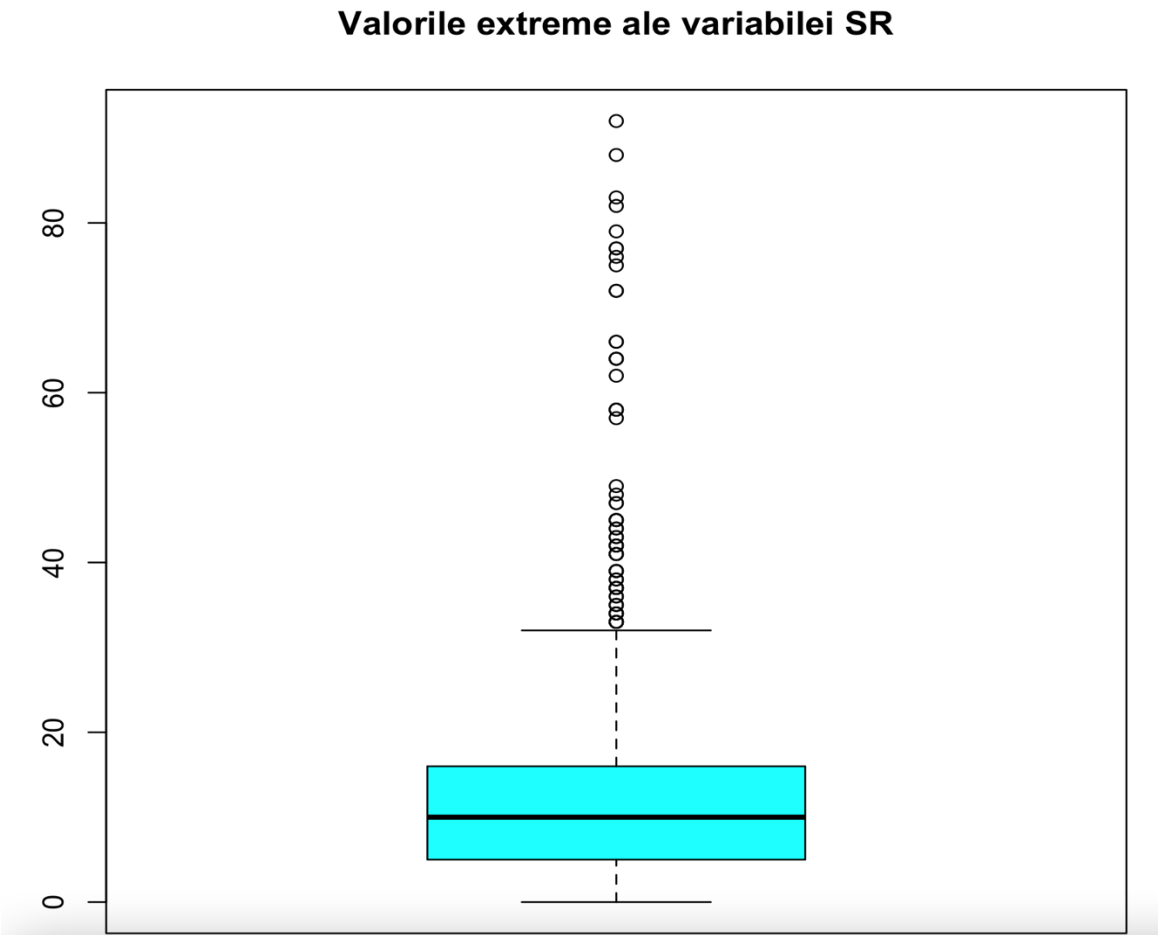
- $Q_1 = 5$; $Q_3 = 16$.
- **H** = $Q_3 - Q_1 \rightarrow H = 16 - 5 \rightarrow H = 11$
- **Limita inferioară** = $Q_1 - 1,5*H \rightarrow$ limita inferioară = $5 - 1,5*11 \rightarrow$ limita inferioară = $33 - 40,5 \rightarrow$ limita inferioară = -11,5. Deoarece în distribuția noastră nu există această valoare, limita inferioară a graficului va fi trasată în dreptul valorii minime (0).
- **Limita superioară** = $Q_3 + 1,5*H \rightarrow$ limita superioară = $16 + 1,5*11 \rightarrow$ limita superioară = 32,5.

Orice scor mai mare de 32,5 sau mai mic de -7,5 ar fi considerat valoare extremă. După cum se poate observa din graficul afișat, în cazul acestei distribuții există un număr destul de mare de valori extreme. Graficul Boxplot se obține în R utilizând funcția **boxplot()**.

boxplot(bdcap6$SR, col = "cyan", main = "Valorile extreme ale variabilei SR"), unde:

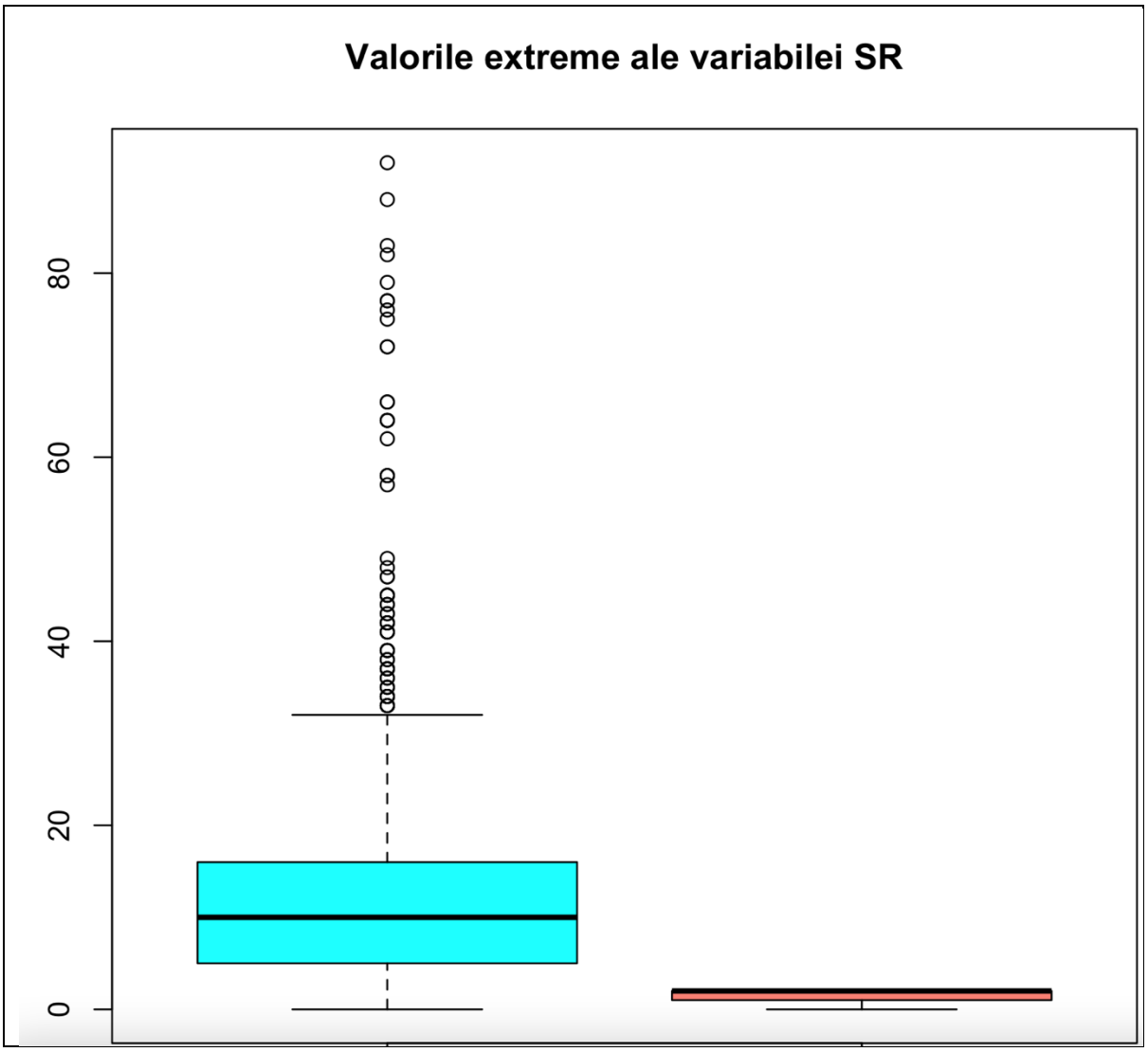
- boxplot() reprezintă funcția care generează graficul;
- bdcap reprezintă baza de date;
- SR este variabila analizată;
- col ne permite să colorăm graficul;
- main este utilizat pentru a stabili titlul graficului.

```
boxplot(bdcap6$SR, col ="cyan", main = "Valorile extreme ale variabilei SR")
```



Dacă dorim să afișăm graficul boxplot separat pentru participanții de gen masculin și participanții de gen feminin, vom folosi linia de cod de mai jos. Graficul din partea stângă reprezintă valorile extreme pentru distribuția scorurilor obținute de bărbați, iar cel din stânga ne indică faptul că în distribuția scorurilor obținute de femei nu există valori extreme.

```
boxplot(bdcap6$SR, bdcap6$Genul, col =c("cyan", "salmon"), main = "Valorile extreme ale variabilei SR")
```


Valorile extreme ale variabilei SR
0
20
40
60
80