# VERIFICAREA DATELOR PARAMETRICE

***Lect. univ. dr. Adrian Gorbanescu***

# INTRODUCERE

- Cele mai multe dintre procedurile statistice sunt dedicate datelor parametrice.
- Testele parametrice folosesc date măsurate pe scală de interval/raport.
- Testele neparametrice folosesc date măsurate pe scală nominală sau ordinală.
- Este foarte important să verificăm îndeplinirea condițiilor necesare pentru aplicarea testelor parametrice înainte de a aplica testul pe care noi îl considerăm potrivit.

# INTRODUCERE

- Cele mai multe dintre testele parametrice necesită îndeplinirea a patru condiții.

1. condiția de normalitate
2. omogenitatea varianțelor
3. datele trebuie măsurate pe scală de interval/raport și fără valori extreme.
4. condiția de independență - se referă la faptul că atunci când măsurăm comportamentul unui participant acesta nu este influențat de comportamentul altui participant.

# CONDITIA DE NORMALITATE

***1. Skewness și Kurtosis***

- Atunci când skewness și kurtosis au valori cuprinse între **-1** și **1** distribuția este normală.

```
#1. Skewness and kurtosis

skew(bd$DERS)
kurtosi(bd$DERS)
```

```
> skew(bd$DERS)
[1] 0.4191354
> kurtosi(bd$DERS)
[1] -0.261033
>
```

- Skewness = 0,41; Kurtosis = -0,26
- Deoarece ambii coeficineți sunt cuprinși între -1 și 1 distribuția DERS îndeplinește condiția de normalitate.

# CONDITIA DE NORMALITATE

***2.* Analiza grafică a normalității**

- Graficul Q-Q plot este utilizat pentru a testa condiția de normalitate.
- Distribuția este normală atunci când punctele sunt/sunt apropiate în/de zona gri.
- Pe măsură ce punctele se îndepărtează de linie înțelegem că distribuția nu îndeplinește condiția de normalitate.
- Acest grafic solicită pachetul **ggpubr**.

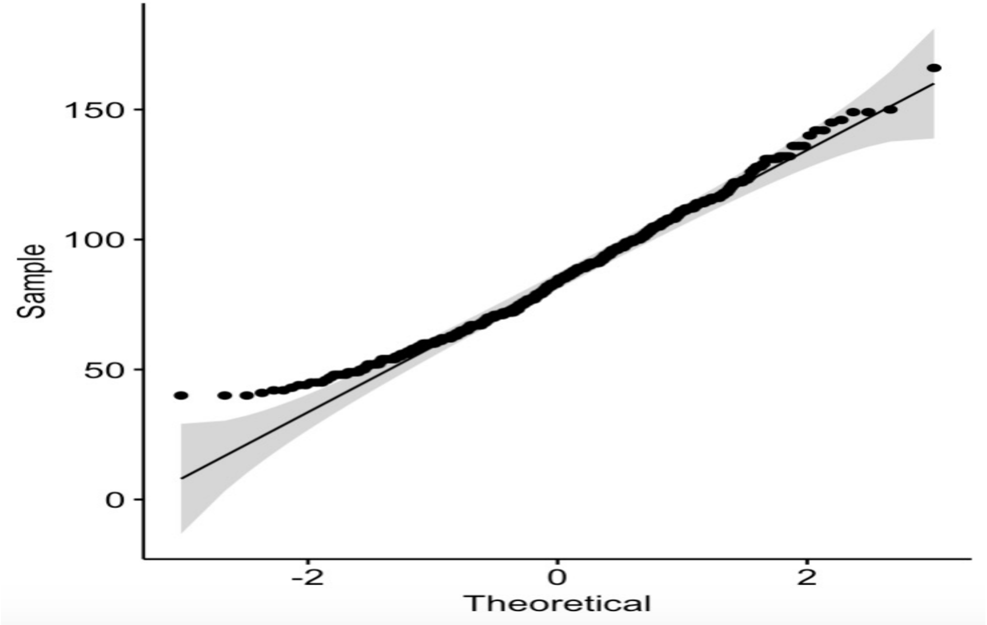
# CONDITIA DE NORMALITATE

- Distanța dintre puncta și zona gri nu este atât de mare, astfel încât putem considera că distribuția este normală.

```
install.packages("ggpubr")
```

```
library(ggpubr)
```

```
ggqqplot(bd$DERS)
```



# CONDITIA DE NORMALITATE

***3.* Aplicarea testelor pentru verificarea normalității**

- 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.
- Testul Shapiro-Wilk este mai robust decât Kolmogorov-Smirnov.

# CONDITIA DE NORMALITATE

- ***W*** – reprezintă valoarea calculată a testului și nu o vom folosi în stabilirea îndeplinirii condiției de normalitate.
- ***p*** **-** reprezintă probabilitatea ca ipoteza de nul să fie adevărată.
- ***Ipoteza de nul ($H_0$) susține că nu există o diferență semnificativă între distribuția analizată și distribuția normală.***
- Dacă ***p*** > alpha acceptăm ipoteza de nul ($H_0$) și vom afirma că distribuția îndeplinește condiția de normalitate.
- Dacă ***p*** ≤ alpha respingem ipoteza de nul și afirmăm că distribuția nu îndeplinește condiția de normalitate.

# CONDITIA DE NORMALITATE

- Alpha reprezintă nivelul de eroare acceptat de comunitatea ştiinţifică.
- Pragul alpha este egal, implicit, cu 0.05.
- Putem stabili un prag alpha mai mic decât 0,05, dar niciodată mai mare de 0, 05.

# CONDITIA DE NORMALITATE

- În acest exemplu **p** este egal cu 0.0000699.
- In concluzie $p < 0.05 \rightarrow$ ipoteza de nult este respinsă.
- Acest rezultat ne indică faptul că distribuția nu îndeplinește condiția de normalitate.

```r
# test of normality
shapiro.test(bd$DERS)
```

```
	Shapiro-Wilk normality test

data:  bd$DERS
W = 0.98164, p-value = 6.999e-05
```

# CONDITIA DE NORMALITATE


- Pragul $\alpha$ este setat implicit la 0.05.
- Dacă dorim, alpha poate avea o valoare mai mică (de exemplu, 0,01), dar niciodată mai mare de 0,05.

# CONDITIA DE OMOGENITATE A VARIANTELOR

- 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.

- Omogenitatea variantelor poate fi verificată cu ajutorul testului **Levene**.
- Testul Levene solicită pachetul **car**.

# CONDITIA DE OMOGENITATE A VARIANTELOR

- **F** este valoarea calculată a testului.
- **Pr (p)** – indică probabilitatea ca ipoteza de nul să fie adevărată.
- Asumpția ipotezei de nul este că nu există diferențe semnificative între varianțe → varianțele sunt omogene.
- Dacă p <= alpha $H_0$ este respinsă și concluzionăm că varianțele nu sunt omogene.
- Dacă p > alpha acceptăm $H_0$ și concluzionăm că varianțele sunt omogene.

```
install.packages("car")
```

```
library(car)
```

```
# Levene test
leveneTest(bd$DERS, bd$Gender, center=mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   1   1.994 0.1587
      390
```

# CONDITIA DE OMOGENITATE A VARIANTELOR

- **p** = 0.158 → ipoteza de nul este acceptată.
- În concluzie nu există o diferență semnifcativă între împăștierea scorurilor DERS obținute de participanții de gen masculine și varianța scorurilor obținute de participanții de gen feminin → variațele sunt omogene.

```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   1   1.994 0.1587
      390
```

# CONDITIA DE OMOGENITATE A VARIANTELOR

- ***pe m*ă*sur*ă *ce volumul eșantionului crește, testul Levene poate respinge omogenitatea varianțelor, chiar dac*ă *în realitate varianțele grupurilor sunt egale.***

***SOLUTIE!!!***

- ***raportul varianțelor - presupune a calcula raportul dintre grupul cu varianța cea mai mare și grupul cu varianța cea mai mic*ă*, iar rezultatul obținut va fi comparat cu o valoare critic*ă *(de obicei aceast*ă *valoare este egal*ă *cu).***
- ***Dac*ă *raportul este mai mic sau egal cu 5 varianțele sunt omogene.***

# CONDITIA DE OMOGENITATE A VARIANTELOR

- ***Folosind funcția var() calculăm varianța DERS pentru cele două eșantioane.***
- ***Astfel, pentru bărbați $s^2$ = 173.243***
- ***Pentru femei $s^2$ = 205,066***
- ***Hartley's F Max = 205,066/173,243 -> Hartley's F Max = 1.836***

```
var(dbc5$DERS [dbc5$Genul==1])

var(dbc5$DERS [dbc5$Genul==2])
```

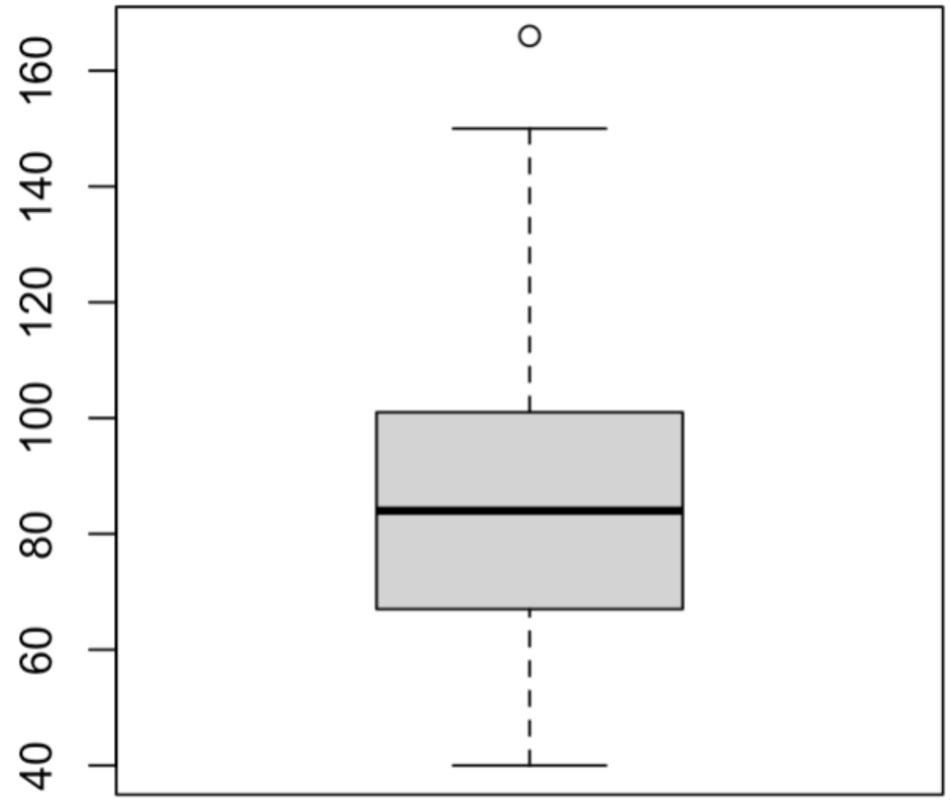
```
173.2434

205.0665
```

- ***Deoarece Hartley's F Max < 5 -> varianțele sunt omogene***

# TRATAREA VALORILOR EXTREME

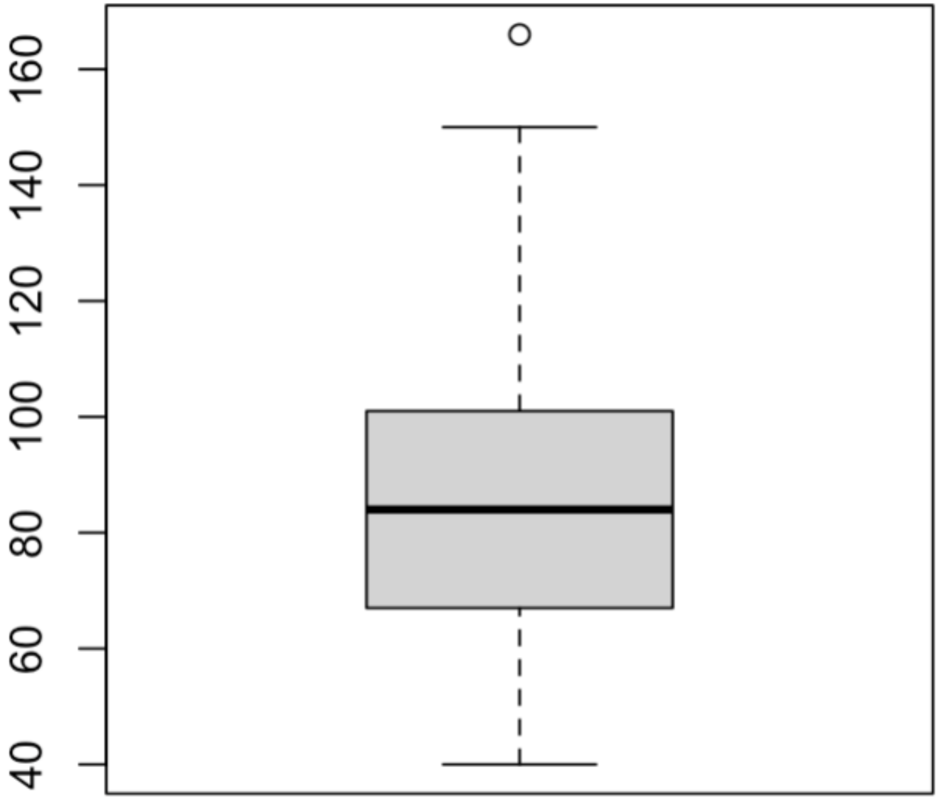
- ***Graficul Box-Plot***
- ***Scorurile z – scorurile aflate în afara intervalului -3z → 3z sunt considerate scoruri extreme***

# OUTLIERS




- Graficul Boxplot reprezintă o metodă simplă de identificare a valorilor extreme.
- Înălțimea cutiei este egală cu abaterea interquartilă (quartila 3 – quartila 1)
- Marginea de sus a cutiei este Quartila 3, în timp ce marginea de jos este Quartila 1.

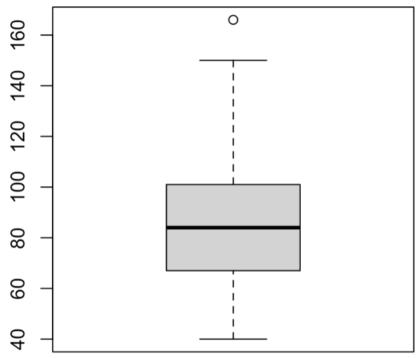
# OUTLIERS




- Limita superioară este calculată folosind formula Quartila 3 + 1,5*H
- Limita inferioară este calculată folosind formula Quartila 1– 1,5*H

# OUTLIERS

```
summary(bd$DERS)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  40.00   67.00   84.00   85.16  101.00  166.00 
```




- Q1 = 67, while Q3 = 101.
- Abaterea interquartilă (H) = 101 – 67 -> H = 34
- Limita inferioară = Q1 – 1.5*H -> lower limit = 67 – 1.5*34
- Limita inferioară = 67-51 -> lower limit = 16.

- Limita superioară – Q3 +1.5*H -> upper limit = 101+1.5*34
- Limita superioară = 101 + 54 -> upper limit = 155

- Values that are not between 16 and 155 are outliers.

# *MULȚUMESC!!!*