

Statistică Descriptivă utilizând funcțiile predefinite și pachetul de funcții DATA ANALYSIS din Microsoft Excel:

Scopul LP: dobândirea abilităților necesare unei analize statistice descriptive în Excel.

Obiective educaționale: După parcurgerea acestei, ar trebui să fii capabil să:

- calculezi statisticile descriptive asociate variabilelor cantitative utilizând funcții predefinite (AVERAGE, MEDIAN, MODE, STDEV, QUARTILE) din Excel
- calculezi statisticile descriptive asociate variabilelor cantitative utilizând opțiunea Descriptive Statistics din pachetul de funcții DATA ANALYSIS
- interpretezi statisticile descriptive

TEMA 1

Un medic de familie care are înscrisi pe liste un număr de 2542 pacienți a realizat un studiu pentru a identifica riscul cardiovaacular („sansa” de a face un infarct miocardic sau AVC într-o perioada definită de timp). Datorită resurselor limitate s-a recurs la alegerea aleatoare a unui eșantion de 100 pacienți.

Au fost incluși în eșantion prin eșantionare sistematică tot al 41-lea pacient cu vârstă între 20 și 90 ani și care a semnat acordul de participare la studiu. Următoarele date au fost culese de la fiecare pacient: vârstă (ani, fișa pacientului), sex (M/F), mediul de proveniență (Urban/Rural), greutate (kg), înălțime (cm), glicemie (mg/dl), colesterol (mg/dl), trigliceride (mg/dl) HDL colesterol (mg/dl) și risc cardiac (da/nu).

În urma studiului variabilelor menționate mai sus precum și a altor factori (ereditatea, etc.), medicul de familie a decis pentru fiecare pacient dacă are sau nu risc cardiac (variabila Risc cardiac(da/nu)). Datele au fost colectate și introduse în fișierul **BD_Lab06_AMG.xlsx**.

Cerințe

1. Creați pe desktop un folder denumit **Lab06**.
2. Salvați fișierul **BD_LP06_AMG.xlsx** (afișat pe site la acest lp) cu denumirea **Lab06NP.xlsx** unde N = inițiala numelui d-voastră și P = inițiala prenumelui d-voastră în folderul **Lab06**.
3. **Utilizând funcțiile predefinite ale Excel-ului**, calculați (în **Tabelul 1** din Excel) următoarele statistici descriptive asociate variabilei **Vârstă**: **media aritmetică, mediana, modulul, deviația standard, cuartila Q1 și cuartila Q3** – *vezi instrucțiuni*
4. Interpretați cuartila Q1 a variabilei **Vârstă**. Scrieți răspunsul în foaia de calcul Excel (la dreapta valorii obținute pentru cuartila Q1).

5. Reprezentați grafic **variabila Vârstă** folosind un grafic box-plot (cutie cu mustăți) – *vezi instrucțiuni*
6. Copiați (copy-paste) variabilele **Vârstă, Greutate, Înălțime, Glicemie, Colesterol, Trigliceride, HDL-colesterol, Risc cardiac** într-o nouă foaie de calcul a aceluiași fișier denumită **Pacienti_Risc_Cardiac**.
7. Sortați datele în raport cu variabila **Riscul cardiac** pentru că dorim să grupăm pacienții în funcție de prezența sau absența riscului cardiac. – *vezi instrucțiuni*
8. Din foaia de calcul **Pacienti_Risc_Cardiac**, ștergeți toate datele măsurate **la pacienții FARA risc cardiac**.
9. În foaia de calcul **Pacienti_Risc_Cardiac**, folosind pachetul de funcții **DATA ANALYSIS** (*vezi instrucțiuni*), calculați pentru TOATE VARIABILELE CANTITATIVE, statisticile descriptive de centralitate (medie, mediana, modul), de dispersie (varianța, deviație standard, eroare standard, amplitudine), de simetrie (asimetrie, boltire) și de localizare (minim, maxim).
10. Completați tabelul obținut la punctul 9 cu **coeficientul de variație** utilizând formula de mai jos:

$$\text{Coeficient de variație} = (\text{Deviația standard}/\text{media}) * 100$$

11. În aceeași foaie de calcul **Pacienti_Risc_Cardiac**, raspundeți la următoarele întrebări:
 - ♦ valorile colesterolului total măsurate **la pacienții cu risc cardiac**, au o **distribuție simetrică**?
 - ♦ valorile colesterolului măsurate **la pacienții cu risc cardiac**, urmează o **distribuție Gaussiană**?
 - ♦ cum se interpretează valoarea coeficientului de variație al colesterolului?

TEMA 2 - problemă recapitulativă

1. În foaia de calcul Date, realizați un grafic cu procente care să conțină repartitia pacienților după **Riscul cardiac**.
2. Realizați o histogramă cu **5 clase** (intervale de clasă) a **glicemiei** pacienților din eșantionul de studiu.
3. Utilizând **funcțiile predefinite din Excel**, calculați media, mediana și modulul **glicemiei** la pacienții fără risc cardiac.

Este bine de reținut că....

- Media, mediana, modulul se vor calcula utilizând funcțiile predefinite AVERAGE, MEDIAN respectiv MODE
- Vom spune că VARIABILA CANTITATIVĂ de interes are o **distribuție simetrică** dacă media aritmetică a valorilor variabilei de studiu este aproximativ egală cu mediana și modulul atunci vom spune că distribuția variabilei este simetrică

- Vom spune că VARIABILA CANTITATIVĂ de interes are o **distribuția normală** de probabilitate dacă media aritmetică a valorilor variabilei de studiu este aproximativ egală cu mediana și modulul și în plus coeficienții de asimetrie și boltire sunt în jurul lui 0 (intervalul $[-1; 1]$)
- Graficul potrivit pentru reprezentarea distribuției unei singure VARIABLE CALITATIVE (dihotomiale, nominale, ordinale) poate fi graficul sectorial (pie), coloane sau graficul prin bare
- Graficul potrivit pentru reprezentarea distribuției unei singure variabile CANTITATIVE este histograma sau graficul cutie-cu-mustăți
- Repartiția subiecților din eșantion în raport cu două VARIABLE CALITATIVE poate fi reprezentată grafic sub forma unui grafic prin coloane sau bare