Prof. univ. dr. ing. Titi PARASCHIV

STATISTICAL PACKAGE FOR SOCIAL SCIENCES – SPSS TEORIE ȘI APLICAȚII

EDITURA UNIVERSITĂȚII „TITU MAIORESCU" • EDITURA HAMANGIU

București, 2023

CAPITOLUL 2

STATISTICI DESCRIPTIVE ȘI INFERENȚIALE

2.1 STATISTICI DESCRIPTIVE

2.1.1 Descrierea variabilelor. Diagrame și Tabele

Diagramele și tabelele evidențiază trăsăturile datelor. Analizele de date le utilizează pentru a permite interpretarea distribuțiilor de variabile.

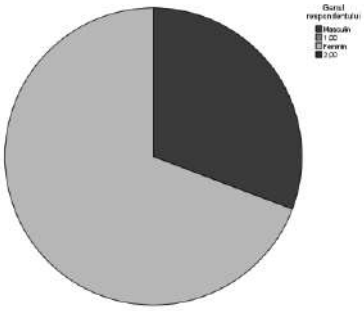
Tabelele de frecvențe numără aparițiile diferitelor valori ale variabilelor.

Se realizează cu comanda:

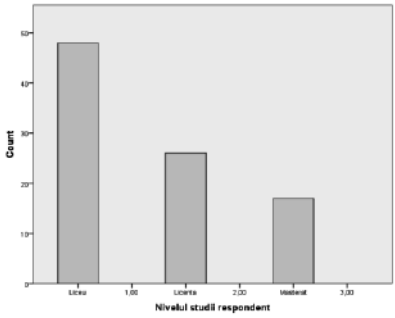
Analyse – Tables – Custom tables...

	Genul respondentului	
	Masculin	Feminin
	Mean	Mean
Nivel_anxietate respondent	28,82	33,49

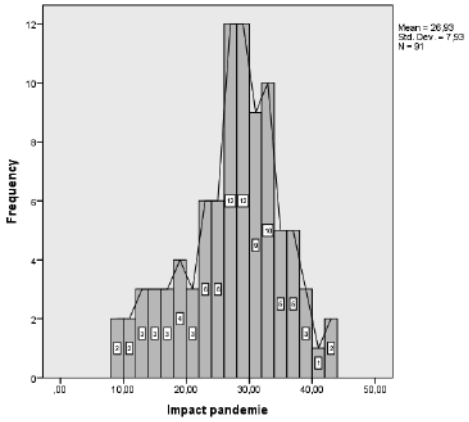
Diagramele circulare sunt utilizate atunci când variabila deține un număr mic de valori distincte. Se utilizează la conferințe și prelegeri.



Diagramele cu bare se utilizează pentru un număr mai mare de valori ale variabilelor.



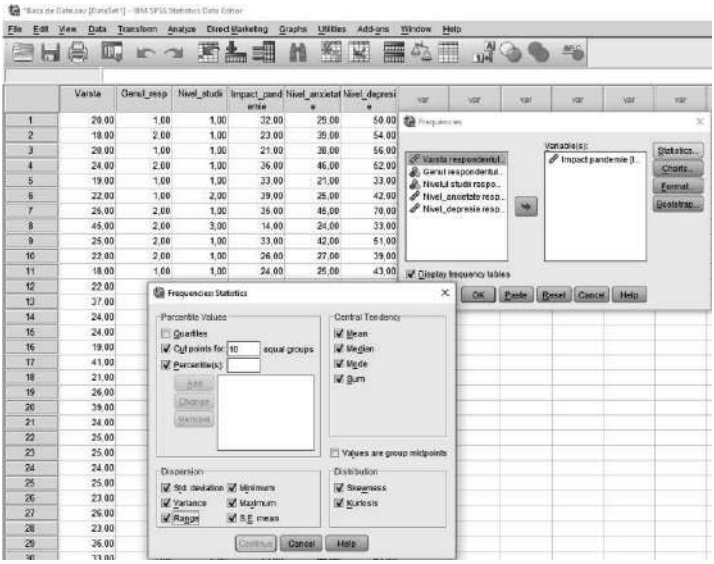
Histogramele sunt similare diagramelor cu bare, dar se utilizează mai ales pentru scorul numerice decât pentru categorii.



Pentru a realiza o diagramă pentru date categoriale se utilizează comanda:

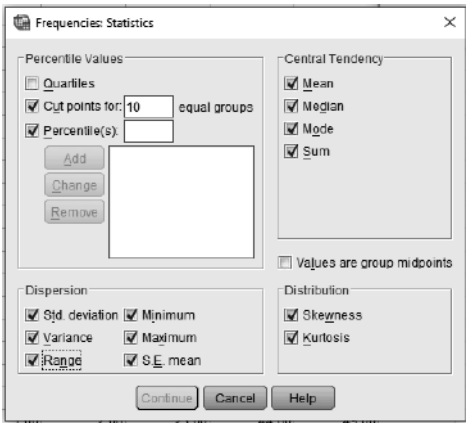
a. **Graphs – Chart Builder...** apoi se alege tipul de chart și tipul de variabilă categorială.
b. **Graphs – Legacy Dialogs...** apoi se alege tipul de chart și tipul de variabilă scor sau categorială.

2.1.2 Descrierea numerică a variabilelor



Tehnicile utilizate pentru descrierea numerică a variabilelor, sunt tehnici univariate, ce generează un indice numeric pentru a descrie o caracteristică a datelor.

Cu excepția modulului care poate fi utilizat pentru orice tip de date, toate tehnicile sunt destinate datelor sub forma unor scoruri numerice.



Media reprezintă media aritmetică a unui set de scoruri ce se obține prin însumarea valorilor și împărțirea rezultatului la numărul de valori.

Modulul este scorul cu cea mai mare frecvență. Un set de scoruri poate dispune de mai mult de un modul în cazul în care 2 sau mai multe scoruri apar cu frecvențe egale. Modulul este valoarea scorului care apare cel mai frecvent.

Mediana este scorul din centrul distribuției dacă scorurile sunt ordonate după mărime, de la cea mai mică la cea mai mare. Media reprezintă suma unui număr de scoruri împărțită la numărul de scoruri. **Mediana** se obține prin ordonarea scorurilor în ordine crescătoare, scorul care separă prima jumătate a cazurilor de a 2-a jumătate este mediană. În statistică discutăm despre indicatorii de măsură a tendinței centrale: media, mediană și modul sunt toate măsuri ale tendinței centrale, toate reprezintă măsuri măsura scorului tipic dintr-o serie de scoruri.

Dispersia scorurilor este altă caracteristică.

Dispersie (Variance)

Dispersia unei liste de valori este pătratul abaterii standard, adică media pătratelor abaterilor numerelor de la media lor. Dispersia unei variabile aleatoare X, notată Var (X), este valoarea aşteptată a diferenţei pătrate dintre variabilă valoarea ei aşteptată: Var (X) = Exp ((X - E

$(X))^2$). Dispersia unei variabile aleatoare este pătratul erorii standard (SE) a variabilei.

Dispersie de sondaj (Sample Variance)

Dispersia de sondaj s^2 este un estimator al dispersiei populaţiei, bazat pe un eşantion aleatoriu. Ca statistică, măsoară gradul de împrăştiere a eşantionului în jurul mediei de sondaj. Presupunând că există n elemente în eşantion, cu valorile $\{x_1, x_2,..., x_n\}$, având media $M = (x_1 + x_2 +... + x_n)/n$, atunci $s^2 = [(x_1 - M)^2 + (x_2 - M)^2 +... + (x_n - M)^2]/(n-1)$. Se observă că este pătratul abaterii standard de sondaj, s. Dispersia de sondaj este un estimator nedeplasat al dispersiei populaţiei.

Diferența dintre cel mai mare scor și cel mai mic este cunoscută sub numele de **amplitudine**.

Varianța este cea mai importantă măsură a dispersiei scorurilor, se bazează pe abaterea medie pătratică. Diferența dintre fiecare scor și media tuturor scorurilor deținute, de fapt, reprezintă abaterea medie pătratică față de medie.

Percentilele indică punctele de separație pentru procentajele scorurilor. Împărțind amplitudinea în 100 de părți. **Cvartidele** reprezintă valori ale distribuției care indică punctele de separare pentru cele mai mici decât 25%, mai mici de 50% și cele mai mici de 75% dintre scoruri.

Suma reprezintă totalul scorurilor pentru o variabilă.

Indicele de asimetrie ne arată că distribuțiile de frecvență nu sunt simetrice față de medie. Acesta este un indice al asimetriei sau înclinării.

Indicele de aplatizare este indicele care ne arată cât de ascuțită sau de plată este distribuția scorurilor pentru o variabilă, comparativ cu distribuția normală.

Abaterea standard (estimare) este o evaluare a măsurii în care scorurile diferă în medie față de media scorurilor, pentru o anumită variabilă.

Varianța (estimare) este o evaluare a cantității cu care variază, în medie, scorul de față de media scorurilor pentru variabila respectivă.

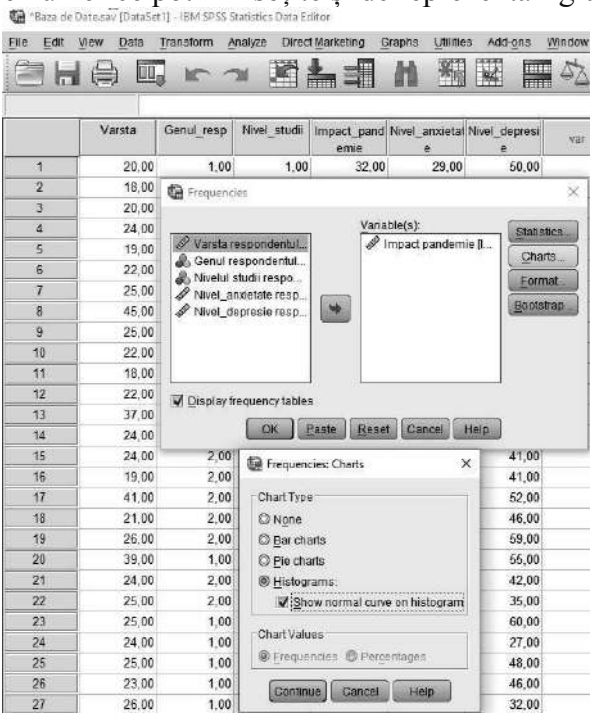
Amplitudinea este diferența numerică dintre cel mai mare și cel mai mic scor obținute pentru o variabilă.

Minim este valoarea reprezintă valoarea celui mai mic scor pentru o anumită variabilă.

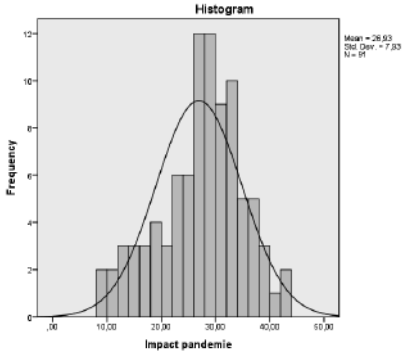
Maxim este valoarea celui mai mare scor pentru o anumită variabilă.

Eroarea standard (ES medie) reprezintă valoarea medie cu care mediile eșantioanelor extrase dintr-o populație diferă față de media populației.

Caracteristicile numerice pot fi însoțite și de reprezentări grafice:

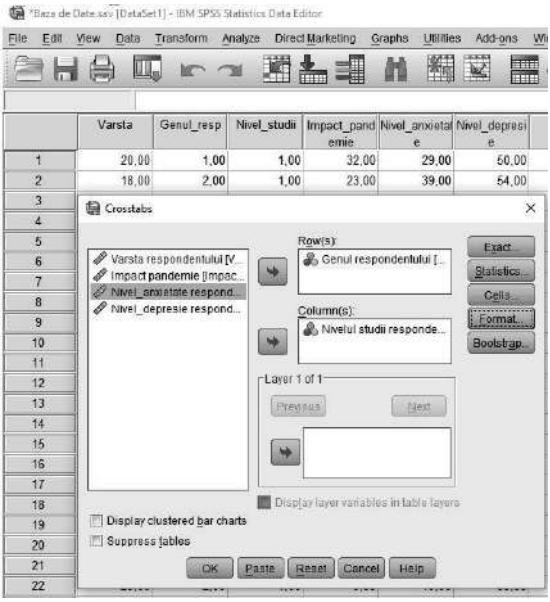


Forma graficului este în figura următoare:



Cu comanda:

Analyze – Descriptiv Statistics – Crosstabs



se realizează un tabel de forma:

Genul respondentului * Nivelul studii respondent Crosstabulation

Count

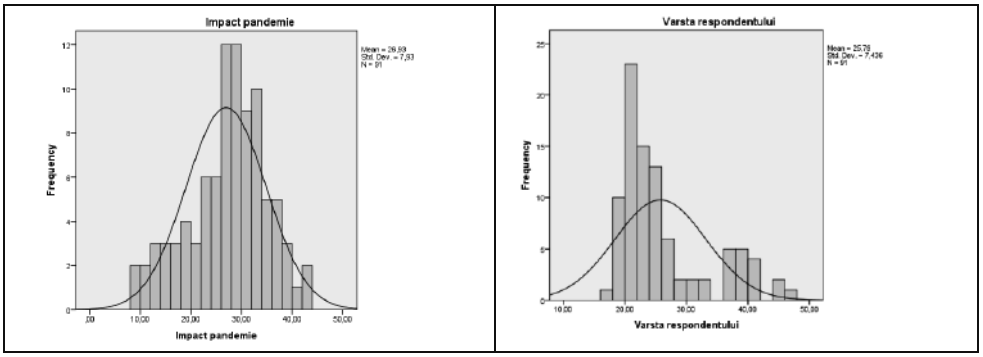
		Nivelul studii respondent			
		Liceu	Licenta	Masterat	Total
Genul respondentului	Masculin	13	10	5	28
	Feminin	35	16	12	63
Total		48	26	17	91

2.1.3. Forme ale distribuțiilor scorurilor

Este important de studiat forma distribuțiilor scorurilor pentru fiecare variabilă. Majoritatea tehnicilor statistice din statistica descriptivă și inferențială au randament și eficiență foarte bună dacă scorurile nu au valori aberante și distribuția este normală, adică o curbă de tip Gauss. Cele mai multe tehnici statistice au cea mai mare eficiență atunci când distribuțiile variabilelor implicate au o distribuție normală.

Uneori, când scorurile nu au o distribuție normală, este posibilă transformarea statistică a acestor scoruri pentru a aproxima o distribuție normală și, în general, folosim scara logaritmică, adică logaritmăm scorurile apoi le aplicăm tehnicile statistice. Logaritmul are rolul de a liniști funcțiile de distribuție. În cazul în care distribuțiile sunt profund asimetrice și a celor care conțin scoruri neobișnuit de mici sau de mari,

adică valori aberante, este nevoie să se efectueze operații de normalizare a acestor rezultate înainte de a aplica tehnicile statistice.



Testele statistice se bazează pe ipoteza că scorurile pentru variabilele supuse operațiilor de procesare statistică au o distribuție normală (de tip Gauss).

Distribuţie normală (Normal distribution)
Prin definiţie, o v.a. X are o repartiţie normală cu parametrii μ şi σ dacă densitatea sa de probabilitate este:

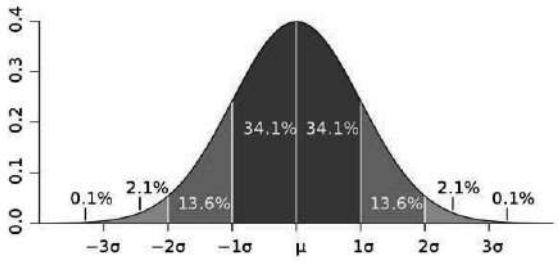
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

.

Se demonstrează că μ şi σ^2 este media, respectiv dispersia, v.a. X. Conform definiţiei funcţiei de repartiţie,

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(u)du$$

şi se poate demonstra că pentru orice ***a*** ≤ ***b***, probabilitatea ca ***a*** < (X-m)/s< ***b*** este P(***a*** < (X-m)/s < ***b***) = aria de sub curba normală standard delimitată de x = ***a*** şi x = ***b***, formulă care permite calcularea probabilităţilor asociate cu repartiţia normală doar cunoscând probabilităţile asociate repartiţiei normale standard. Notaţia uzuală este $X \sim N(\mu,\sigma^2)$. Pentru distribuţia normală standard se obţine $X \sim N(0,1)$.



2.1.4. Abaterea standard. Unitatea de măsură standard în statistică

Abaterea standard a unei variabile aleatoare reprezintă o măsură a dispersiei valorilor acesteia în jurul uneia considerate mijlocii. Se mai numește și **abatere** medie pătratică sau deviație standard.

Deviația standard, este o măsură a răspândirii unei serii sau a distanței față de standard. În 1893, Karl Pearson a inventat noțiunea de abatere standard, care este, fără îndoială, cea mai utilizată măsură, în studiile de cercetare.

Este rădăcina pătrată a mediei pătratelor abaterilor de la media lor. Cu alte cuvinte, pentru un set de date dat, abaterea standard este rădăcină-media-pătrată-deviație, de la media aritmetică. Pentru întreaga populație, este indicată prin litera greacă „sigma (σ)", iar pentru un eșantion, este reprezentată prin litera latină „s".

Deviația standard este o măsură care cuantifică gradul de dispersie al setului de observații. Cu cât punctele de date sunt mai departe de valoarea medie, cu atât este mai mare abaterea în setul de date, reprezentând că punctele de date sunt împrăștiate pe o gamă mai largă de valori și invers.

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$$

Definiția Standard Error – eroarea standard

S-ar putea să fi observat că eșantioane diferite, cu dimensiuni identice, extrase din aceeași populație, vor da valori diferite ale statisticii luate în considerare, adică media eșantionului. Eroarea standard (SE) furnizează abaterea standard în diferite valori ale mediei eșantionului. Este utilizat pentru a face o comparație între mediile eșantionului din cadrul populațiilor.

Pe scurt, eroarea standard a unei statistici nu este altceva decât abaterea standard a distribuției sale de eșantionare. Are un rol important în testarea ipotezelor statistice și a estimării intervalului. Oferă o idee despre exactitatea și fiabilitatea estimării. Cu cât eroarea standard este mai mică, cu atât este mai mare uniformitatea distribuției teoretice și invers. Eroare standard pentru media eșantionului = $\boldsymbol{\sigma/\sqrt{n}}$
Unde, σ este deviația standard a populației.

Diferențele între abaterea standard și eroarea standard

- **Deviația standard** este indicatorul care evaluează valoarea variației în setul de observații. **Standard Error (eroarea standard)** măsoară acuratețea unei estimări, adică este măsura variabilității distribuției teoretice a unei statistici.
- **Deviația standard** este o statistică descriptivă, în timp ce **eroarea standard** este o statistică inferențială.
- **Deviația standard** măsoară cât de departe sunt valorile individuale de valoarea medie. **Eroarea standard** măsoară cât de apropiată este media eșantionului de media populației.
- **Abaterea standard** este distribuția observațiilor cu referire la curba normală. Față de aceasta, **eroarea standard** este distribuția unei estimări cu referire la curba normală.
- **Deviația standard** este definită ca rădăcina pătrată a variației. În schimb, **eroarea standard** este descrisă ca abaterea standard împărțită la rădăcina pătrată a dimensiunii eșantionului.
- Când dimensiunea eșantionului este crescută, aceasta oferă o măsură mai particulară a abaterii standard. Spre deosebire de abaterea standard atunci când dimensiunea eșantionului crește, eroarea standard tinde să scadă.