

FACULTATEA DE PSIHOLOGIE ȘI ȘTIINȚELE EDUCAȚIEI

DISCIPLINA: *Informatică aplicată în psihologie* (disciplină obligatorie, de specialitate)

PROGRAMUL DE STUDIU: *PSIHOLOGIE*

CICLUL DE STUDII: I – diplomă de licență, anul I

INFORMATICĂ APLICATĂ ÎN PSIHOLOGIE

- suport de curs -

An universitar 2011-2012

CE ESTE STATISTICA...

Statistica se ocupă cu stringerea, descrierea și analizarea datelor în vederea extragerii unor concluzii pe baza acestora. În esența ei, statistica operează cu numere care descriu realitatea. Termenul de „statistică” provine din latinescul medieval „status”, care semnifica „stare politică”. În anul 1770, la Londra, baronul Bielfeld publică lucrarea „*The Elements of Universal Erudition*” în care există un capitol de „statistică”, definită ca fiind: „*Știința care ne învață care este organizarea politică a tuturor statelor moderne ale lumii*”.

Utilizarea sistematică a statisticii în psihologie datează de la începutul anilor '50. La vremea respectivă, în care calculatoarele se aflau încă în era electromecanică, cea mai mare problemă era efortul de calcul. Din acest motiv, a învăța statistică însemna atunci mai ales învățarea unor formule și a unor proceduri care să simplifice pe cât posibil efortul de calcul. Apariția calculatoarelor și a programelor de prelucrări statistice, au făcut posibilă trecerea la faza în care accentul se pune pe înțelegerea raționamentelor statistice.

STATISTICA ÎN ACTIVITATEA A PSIHOLOGILOR

Iată câteva argumente în sprijinul ideii că utilizarea statisticii face parte integrantă din activitatea curentă a unui psiholog:

- Una dintre cele mai obișnuite posturi ale psihologului este de a utiliza teste în evaluarea unor caracteristici psihice. Ei bine, statistica este direct și total implicată în acest aspect profesional. În faza de elaborare a testului alegerea itemilor și evaluarea consistenței interne (stabilirea calității de instrument de măsurare psihică), se bazează pe proceduri statistice. În faza de utilizare se utilizează proceduri statistice pentru fixarea etaloanelor de raportare a scorurilor individuale. Este evident că psihologul are nevoie de statistică, cu atât mai mult, în situațiile în care dorește să elaboreze el însuși un test psihologic.
- Selecția psihologică este un domeniu de aplicabilitate larg răspândit și care se bazează puternic pe aportul statisticii. Cu ajutorul ei se pune în evidență calitatea prognostică (predictivă) a unuia sau mai multor teste luate împreună (baterie), se fixează pragul de respingere și se constituie scorurile individuale pe baza cărora se ia decizia de selecție.
- Chiar și atunci când se dedică unei activități cu orientare puternic individuală, cum este psihoterapia, psihologul nu se poate dispensa de statistică. Ea îi este necesară pentru evaluarea eficienței unei noi metode terapeutice, comparativ cu o altă metodă, sau pentru identificarea anumitor condiții care pot influența eficiența ședințelor de terapie (ora zilei, similaritatea de sex dintre pacient și terapeut, etc.).
- De asemenea integrarea în mediul științific profesional obligă psihologul să utilizeze metodele statistice în elaborarea studiilor proprii sau în înțelegerea studiilor din literatura de specialitate.

Nu trebuie să uităm însă că nici statistica și nici metodologia de investigare psihologică (teste, dispozitive computerizate, etc.), oricât de sofisticate ar fi, nu dau psihologiei, prin ele însele, un caracter de știință.

DIFICULTĂȚI ȘI RISCURI ÎN ÎNSUȘIREA METODELOR STATISTICE

Se știe că statistica n-a fost niciodată disciplina preferată a studenților în psihologie. Trebuie să acceptăm adevărul că majoritatea celor care vor să înțeleagă psihicul uman nu manifestă o simpatie deosebită pentru numere, formule și algoritmi de calcul. Însă un curs introductiv de statistică aplicată în psihologie este pe deplin accesibil oricărui absolvent de liceu, chiar și celor care nu au urmat o secție de tip „real”.

Ce dificultăți pot să apară?

- abundența de concepte noi, cu semnificații uneori dificil de înțeles în mod intuitiv;
- prezența unor concepte cunoscute din limbajul cotidian dar care au semnificații diferite în domeniul statisticii;
- dificultatea înțelegerii raționamentelor de tip probabilistic.

Să reținem totuși că utilizarea în necunoștință de cauză a unor prelucrări statistice, fără a ști dacă sunt legitime sau nu și fără a le înțelege semnificația, poate conduce la grave confuzii. Acest tip de atitudine este favorizat în ultima perioadă și de utilizarea programelor de calculator, care permit proceduri statistice diverse și care întrețin iluzia că ar putea fi aplicate în afara unei profunde cunoașteri a rostului și semnificației lor.

Există și cazuri în care utilizarea statisticii se face cu scopul de a apărea mai doct. Se face astfel apel la cât mai multe date, la reprezentări grafice redundante sau la proceduri rar utilizate, chiar dacă până la urmă se ajunge la aceleași concluzii.

Este bine să ne însușim faptul că statistica în psihologie este, întotdeauna, un mijloc și nu un scop. Ea este un simplu instrument în atingerea unor obiective.

MĂSURAREA ÎN PSIHOLOGIE

A măsura înseamnă a atribui numere sau simboluri unui aspect al realității obiective sau subiective, în funcție de anumite aspecte cantitative sau calitative care le caracterizează. În acest mod relația dintre numere sau simboluri ajunge să reflecte relația dintre caracteristicile cărora le-au fost atribuite. Modul în care sunt atribuite numere sau simboluri pentru a măsura ceva, se numește „scală de măsurare”.

Este important să înțelegem procesul de măsurare deoarece:

- vom cunoaște tipurile de transformări la care putem spune în mod legitim valorile rezultate prin măsurare. De exemplu, dacă am măsurat distanța în centimetri, știm că o putem transforma în *inch* prin aplicarea unei reguli, fără a altera semnificația valorilor.
- Vom ști să alegem procedurile statistice adecvate datelor numerice și scopurilor pe care ni le propunem. De exemplu, nu vom putea alege proceduri de tip „metric” (cantitativ) atunci când variabila dependentă este de tip „non-metric” (calitativ).

Statistica operează cu valori, numerice sau de altă natură, care rezultă dintr-un proces de măsurare. Dar numerele, deși au aceeași formă, nu sunt asemănătoare unele cu altele. Ele pot avea diferite semnificații sau proprietăți în funcție de tipul de măsurare din care rezultă. În funcție de cantitatea de informație pe care o reprezintă valorile, ca rezultat al procesului de măsurare, putem distinge mai multe tipuri de scale de măsurare:

SCALE NOMINALE

O măsurare pe scală nominală înseamnă a clasifica obiectele în diverse categorii. În acest caz, o valoare nu este cu nimic mai mare sau mică decât altă valoare. De exemplu, „valoarea” atribuită genului poate fi codificată cu „M” sau „F”, ori, la fel de bine cu „2” sau „1”. În acest caz, respectivele „valori” nu sunt decât simboluri ale unei anumite calități pe care o ia caracteristică de gen a unei persoane. Altfel spus, într-un asemenea caz „2” nu înseamnă că este „mai mult” sau „mai bun” decât „1” ci doar faptul că este „diferit” de acesta. Putem remarca că ambele codificări de mai sus sunt arbitrare, în locul lor putând utiliza orice alte simboluri, pe o bază de convenție. Variabilele măsurate pe scale de tip nominal pun în evidență diferențe calitative între valori și nu cantitative.

Exemple de variabile exprimate pe scale nominale: bolile psihice (depresie, nevroză, etc.), tipurile temperamentale (sanguin, coleric, flegmatic, melancolic) etc..

Valorile măsurate pe o scală de tip nominal au un caracter calitativ și nu suportă operații numerice, altele decât cele de însumarea.

SCALE ORDINALE

Valorile plasate pe o scală de tip ordinal au o anumită semnificație cantitativă. O anumită valoare este “mai mare” sau “mai bună” decât alta, aflată sub ea. Implicit, ea poate fi “mai mică” sau mai “puțin bună” decât altă valoare, aflată deasupra ei. Dacă o anumită persoană este mai preferată decât alta, și atribuim primei valoarea 1 iar celei de-a doua valoarea 2, atunci cele două valori se exprimă pe o scală de tip ordinal, care indică doar ordinea preferinței și nu măsura intensității acestei preferințe.

Asadar numerele de tip ordinal ne spun dacă o valoare este mai mare sau mai mică decât alta, dacă o anumită calitate este prezentă într-o măsură mai mare sau mai mică, fără a putea preciza care este „diferența de cantitate” a caracteristicii măsurate. Ca urmare, valorile de tip ordinal au, ca și cele de tip nominal, o semnificație calitativă și nu una cantitativă.

SCALE DE INTERVAL

O variabilă măsurată pe o scală de interval ne oferă informații nu doar despre ordinea de mărime ci și despre „dimensiunea” exactă a caracteristicii măsurate. Valorile de acest tip au un caracter cantitativ, exprimat numeric, iar intervalele dintre ele sunt egale.

Exemplu: scala Celsius. Dacă într-o zi se măsoară 5 grade iar în ziua următoare 10 grade, se poate spune cu precizie că a doua zi a fost cu 5 grade mai cald;

Ceea ce este caracteristic valorilor măsurate pe scală de interval este absența unei valori 0 absolute. Cu alte cuvinte, valorile de acest tip nu ne permit evaluări de genul: „O temperatură de 10 grade este de două ori mai mare decât una de 5 grade”. Aceasta deoarece temperaturile măsurate pe scala Celsius nu au o valoare 0 absolută (dacă acceptăm că nici un om viu nu are inteligență nulă).

Posibilitatea măsurării pe scale de interval în psihologie face adesea obiectul unor controverse. Aceasta pentru că cele mai multe dintre variabilele psihologice sunt expresia unor evaluări subiective, fapt care face greu de demonstrat egalitatea intervalelor dintre două valori consecutive.

Cu toate acestea, multe dintre măsurătorile studiilor psihologice sunt asimilate scalei de tip interval.

SCALE DE RAPORT

Valorile exprimate pe o scală de raport au (pe lângă egalitatea intervalelor) și o valoare 0 absolut la care se raportează (nu este posibilă nici o valoare mai mică de 0). Astfel este permisă aprecierea raportului dintre două valori.

Exemplu: scala de temperaturi Kelvin (0 Kelvin este temperatura minimă absolută).

În psihologie puține sunt variabilele acceptate ca fiind măsurate pe scala de raport, deoarece sunt puține situațiile în care avem de a face cu caracteristici ce pot lua valoarea 0 absolut.

La fel ca și valorile măsurate pe scale de interval, valorile măsurate pe scală de raport suportă toate transformările matematice posibile. Din acest motiv, în practică, valorile măsurate pe scală de interval sau de raport sunt considerate similare (variabile numerice), fiind prelucrate prin același gen de proceduri statistice.

Dacă luăm în considerare proprietățile numerice și tipul de transformări suportate de fiecare scală de măsurare, atunci ordinea crescătoare a acestora este *nominal-ordinal-interval-raport*. Din acest punct de vedere se poate chiar spune că scalele de măsurare se plasează pe o scală ordinală.

Scale de măsurare și variabile

Din perspectiva măsurării, variabila se referă la o *caracteristică* supusă măsurării, în timp ce scala se referă la *modalitatea* de măsurare. Uneori aceeași variabilă (caracteristică) poate fi măsurată pe oricare dintre tipurile de scală. De exemplu, timpul de reacție poate fi exprimat pe o scală nominală („corespunzător”, „necorespunzător”), pe o scală ordinală („mic”, „mare”, „mediu”, „foarte mare”) sau pe o scală de interval/raport (în unități de timp). Uneori se folosește expresia „variabilă nominală”, („ordinală” sau „de interval”). Fără a fi greșită, atunci când folosim o astfel de exprimare trebuia să ne gândim că ea semnifică de fapt „variabilă măsurată pe o scală nominală...”, scala de măsurare și variabila fiind noțiuni diferite!

Să reținem, de asemenea, faptul că valorile măsurate pe o scală de nivel superior (cantitativ), pot fi convertite în valori măsurate pe scale calitative. Niciodată, însă, nu vom putea transforma valori calitative în valori cantitative. Atunci când există posibilitatea de a alege, se va prefera întotdeauna măsurarea pe o scală cantitativă (interval/raport).

VARIABLE, POPULATIE, ESANTION

VARIABLE DEPENDENTE I VARIABLE INDEPENDENTE

Studiul statistic își propune evidențierea legăturilor dintre diverse caracteristici ale realității adică între diverse variabile. Există astfel variabile ale căror valori sunt dependente pentru că variază în funcție de valorile altei sau altor variabile, care sunt denumite, din acest motiv, independente. Identificarea lor corectă în cadrul unui studiu statistic este esențială pentru fundamentarea procedurilor statistice.

Variabila dependentă face obiectul măsurării cu scopul de a fi supusă unor concluzii. Prin opoziție, variabila independentă este utilizată ca variabilă de influență, ale căror efecte posibile asupra variabilei dependente urmează să fie puse în evidență. Termenii „dependent”, „independent” se utilizează în mod obișnuit în legătură cu cercetarea experimentală. În acest context există variabile „manipulate” adică „independente” de reacțiile, intențiile, conduitele sau trăirile subiecților investigați (acestea fiind variabile „dependente”). În raport cu analiza statistică, definirea variabilelor ca dependente și independente nu este condiționată de măsurarea lor în condiții de experiment.

Să reținem faptul că nu există variabile care sunt „dependente” sau „independente” prin natura lor. Caracteristica de a fi de un tip sau de altul provine din rolul care le este atribuit de către cercetător într-un anumit context de cercetare.

VARIABLE CONTINUE I VARIABLE DISCRETE

Se numește „continuă” o variabilă de tip numeric (cantitativ, de tip interval/raport) care are un număr teoretic infinit de niveluri ale valorilor măsurate. Acest tip de variabilă poate lua, în principiu, orice valoare, permițând utilizarea zecimalelor.

Asadar variabilele continue, masurate pe scala de interval/raport, sint valori in numar infinit care permit inserarea unei valori intermediare intre oricare alte doua valori.

Ex. - înălțimea unei persoane

- temperatura
- vîrsta

Variabila „discretă” este cea care prezintă un număr finit al valorilor pe care le poate lua. Deci *variabilele discrete*, măsurate pe scale nominale și ordinale, sunt valori izolate care formează mulțimi finite.

Ex. - numărul de membrii a unei familii

- numărul de studenți la o oră de curs
- numărul angajaților de la o firmă

POPULAȚIE ÎN EȘANTION

Ținta oricărei cercetări științifice este identificarea unor adevăruri cu un anumit grad de generalitate. Din punct de vedere statistic „generalul” este reprezentat de totalitatea valorilor care descriu o anumită caracteristică, și este numit „populație”. Din păcate însă, investigarea tuturor „indivizilor” care compun o anumită populație nu este aproape niciodată posibilă. Ca urmare, în practica cercetării științifice se supun cercetării psihologice loturi mai restrânse, extrase din ansamblul colectivității vizate, ai căror parametri descriptivi (medie, variabilitate) sunt extrapolați, în anumite condiții și cu ajutorul unor proceduri specializate, la populația din care fac parte.

Cercetarea statistică face apel la următoarele definiții:

Populație - totalitatea elementelor care constituie obiectivul de interes al unei investigații. Prin „elemente” înțelegem cel mai adesea „persoane” (sau „subiecți”, cu un termen uzual în cercetarea psihologică). Deci prin „populație” trebuie să înțelegem extinderea maximă posibilă, sub aspectul volumului, a respectivelor elemente. Extinderea menționată este, la rândul ei, definită prin obiectivul de cercetare, ceea ce înseamnă că are o dimensiune subiectivă. De exemplu, într-un studiu cu privire la efectul oboselii asupra performanței cognitive, pot fi vizate diferite categorii de „populații”: a studenților, a elevilor etc.

Eșantion - reprezintă elementele selecționate pentru a fi efectiv studiate. Ideea pe care se bazează cercetările bazate pe eșantioane este aceea că se pot face aprecieri asupra unei întregi populații, în anumite condiții, doar pe baza caracteristicilor măsurate pe o parte a acesteia.

Exemplu: într-un studiu care vizează influența inteligenței asupra reușitei școlare în ciclul primar, populația este reprezentată de *toți școlarii din ciclul primar* iar eșantionul de *subiecții incluși în studiu*.

Dacă am reuși adunarea de date cu privire la întreaga populație care face obiectul cercetării, am putea trage concluzii directe cu privire la aceasta prin utilizarea indicatorilor statistici descriptivi cunoscuți (medie, dispersie, abatere standard) numiți și „parametrii populației”. Dar acest lucru nu este și, ca urmare, indicatorii statistici ai eșantionului sunt utilizați pentru a face estimări, inferențe, cu privire la parametrii populației.

Reprezentativitatea eșantionului

Ideea de bază este ca dacă avem un eșantion a cărui alegere respectă anumite condiții, extras dintr-o populație oricât de mare, rezultatele obținute pe acesta pot fi extrapolate la întreaga populație.

Calitatea unui eșantion de a permite extinderea concluziilor la întreaga populație din care a fost extras se numește reprezentativitate. Totuși nici un eșantion nu poate reprezenta perfect datele populației fapt pentru care estimările pe bază de eșantion conțin întotdeauna o doză mai mare sau mai mică de eroare. Cu cât eroarea este mai mică, cu atât concluziile obținute pe eșantion pot fi generalizate cu o certitudine mai mare asupra populației.

Modul de constituire a eșantionului este decisiv pentru nivelul de reprezentativitate. Esențială în acest caz este asigurarea condițiilor ca acesta să acopere în mod real caracteristicile populației.

Mentionăm în continuare cele mai utilizate metode de constituire a eșantioanelor (tehnicile de eșantionare nu constituie obiectivul acestui curs):

- *Eșantionare stratificată multistadială*. Populația se împarte în categorii, fiecare categorie în subcategorii ș.a.m.d., iar subiecții sunt selecționați aleator la nivelul categoriei de nivelul cel mai scăzut. Se obține astfel un eșantion care reproduce fidel structura populației.

- *Eșantionare prin clasificare unistadială.* Se identifică categorii pe un singur nivel iar subiecții se extrag aleator din fiecare categorie.
- *Eșantionare aleatoare.* Subiecții sunt extrași aleator din ansamblul populației. „Aleator”, înseamnă în acest caz utilizarea unei proceduri care asigură fiecărui subiect al populației absolut aceleași șanse de a fi inclus în eșantion. În acest scop se pot utiliza programe de calculator sau tabele de numere aleatoare.
- *Eșantionare pseudo-aleatoare.* Sunt utilizați subiecții „disponibili”. Este cazul cel mai frecvent întâlnit în practică.

Asadar, presupunând că am obținut anumite rezultate pe un eșantion aleator, raționamentul statistic ne permite să aplicăm concluziile la întreaga populație din care a fost extras acel eșantion. Este necesară însă existența unei precizări clare a populației de referință pentru că, dincolo de limitele acesteia, extrapolarea nu este permisă.

STUDII EXPERIMENTALE I STUDII CORELAȚIONALE

Luind în discuție legătura cauzală dintre variabile, trebuie să reținem ca cel mai adesea întâlnim două tipuri de cercetări: *experimentale* și *corelaționale*.

În cazul studiilor experimentale, cercetătorul nu se limitează la măsurarea variabilei independente ci o și manipulează, iar concluziile unui astfel de studiu pot fi interpretate în mod cauzal.

În cazul studiilor numite corelaționale, variabilele dependente și independente sunt măsurate în condiții care nu permit inferențe de tip cauzal. În psihologie, ponderea studiilor corelaționale este mult mai mare decât a celor experimentale, care sunt mai pretențioase și mai dificil de realizat. Aceasta nu înseamnă că studiile „corelaționale” nu sunt relevante. Un studiu „corelațional” nu presupunem doar utilizarea „coeficientului de corelație” ci utilizarea oricărui tip de test statistic care urmărește punerea în evidență a legăturii dintre variabile, fără ca rezultatele să fi fost obținute într-un context experimental.

PARAMETRIC/NEPARAMETRIC, DESCRIPTIV/INFERENTIAL

STATISTICA PARAMETRIC I STATISTICA NEPARAMETRIC

Statistica presupune utilizarea unor proceduri de calcul care urmăresc punerea în evidență a unor legături dintre variabile. Atunci când aceste proceduri se aplică unor situații în care variabilele dependente sunt de tip cantitativ (interval/raport), procedura se numește „parametrică”. De cealaltă parte, procedurile aplicate în cazul în care variabilele dependente sunt de tip „calitativ” (nominale sau ordinale) se numesc „neparametrice”. Alegerea procedurilor statistice nu este deloc un proces simplu. El va fi mai clar înțeles poate după finalizarea cursului și dobândirea unei practici în materie. În cele ce urmează considerăm suficientă o prezentare generală a celor două categorii de proceduri statistice.

STATISTICA DESCRIPTIV I STATISTICA INFERENȚIAL

Statistica descriptivă. Dacă aplicăm un test de timp de reacție unui număr de 100 de persoane, putem calcula valoarea medie a timpilor de reacție, împrăștierea acestora sau, utilizând o tehnică de reprezentare grafică, modul în care se distribuie valorile prin raportare la un sistem de coordonate. Asadar, cu ajutorul statisticii descriptive ne putem face o imagine cu privire la caracteristicile unui distribuții luate în sine, fără a putea emite judecăți comparative prin raportare la populația din care face parte distribuția respectivă sau la un alt lot de valori din aceeași populație. Cu toate acestea, vom vedea că statistica descriptivă este un pas obligatoriu și esențial pentru statisticile avansate.

Statistica inferențială. Vizează metodele de verificare a ipotezelor de cercetare prin testarea ipotezelor statistice. Să presupunem că cei 50 de subiecți de mai sus sunt supuși aceluiași test de timp de reacție în condiții de noxe de mediu (zgomot excesiv, de exemplu) pentru a verifica ipoteza că zgomotul reduce promptitudinea reacțiilor. Într-un astfel de caz statistica inferențială ne pune la dispoziție metode specifice prin care să putem afirma, cu anumită probabilitate, că o eventuală diferență dintre media timpilor de reacție măsurați în cele două condiții diferă semnificativ sau nu.

Este important de reținut că alegerea statisticii parametrice sau neparametrice se face pornind de la *natura variabilei dependente*. Atunci

când aceasta este de tip cantitativ, și nu se abate de la condițiile impuse de procedura statistică pe care dorim să o aplicăm, se utilizează teste statistice parametrice. În orice alte condiții, se apelează la teste neparametrice. Această distincție este oarecum dificil de înțeles în acest stadiu de început al discuției dar este important să fie ținută minte.

Tablou recapitulativ al scalelor de măsurare și statisticile adecvate

Scale	
-------	--

CALITATIVE				NON PARAM

	NOMINALE		<p>Valorile sunt expresia <u>denumirii</u> unei caracteristici a variabilei care privește „un grup de subiecți” (de exemplu: tipul temperamental, o categorie de boli psihice, etc.).</p> <p>O anumită valoare nu ne spune nimic cu privire la celelalte valori.</p>	ETRICE
--	----------	--	--	--------

	ORDINALE	<p>Nivelurile variabilei exprimă doar ordinea unora față de celelalte. Lungimea intervalelor dintre valori este incertă, subiectivă. O anumită valoare ne spune că există valori mai mari sau mai mici decât ea, dar nu și care este dimensiunea acestei diferențe dintre valori. (exemplu: Cât de mult apreciezi să fii cu colegii în timpul liber: F. PUȚIN, PUȚIN, MULT, F. MULT). Atribuirea unei valori se face prin comparație cu alte valori, și nu prin raportare la un criteriu extern („obiectiv”)</p>	
CANTITATIVE	INTERVAL	<p>Nivelurile sunt egal distribuite, adică, o unitate într-o zonă a scalei este egală cu o unitate din orice altă zonă a scalei. Asta înseamnă că valorile variabilei definesc o anumită caracteristică prin raportare la un „etalon extern”, cel care garantează echivalența intervalelor (De exemplu: scala termometrului. Sau scorul la un test care marchează „numărul de puncte” realizate. În acest caz, „acumularea punctelor”).</p>	<p>PARAMETRICE</p> <p>(dacă sunt respectate anumite condiții)</p>

	RAPORT	La fel ca mai sus, cu specificația că nivelurile variabilei nu sunt doar egal distribuite ci există și un zero absolut, care face să existe o valoare care indică absența totală a caracteristicii.	
--	--------	---	--

Idea fundamentală rămâne aceea că, atunci când variabila dependentă implicată într-un studiu statistic este măsurată pe o scală de tip calitativ (nominal sau ordinal), se aplică una dintre procedurile statistice neparametrice. În cazul variabilelor măsurate pe scale cantitative se aplică, de regulă, statistici parametrice, fără ca acest lucru să fie posibil întotdeauna.

DISTRIBUTIA DE FRECVEN E

DISTRIBUTIA DE FRECVENTE SIMPLE

Presupunem ca in urma aplicarii unui test de memorie s-au obtinut urmatoarele scoruri: 9, 8, 6, 5, 8, 8, 7, 6, 6, 4, 4, 10, 6, 4, 10, 9, 8, 8, 7, 7, 6, 6, 6, 4, 2

Avem un sir de date irelevant. Pentru a obtine o imagine mai clara asupra acestor date putem proceda la aranjarea lor într-o ordine descrescătoare:

10, 10, 9, 9, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 6, 6, 5, 4, 4, 4, 4, 2

Privind datele astfel aranjate observam cu ușurință câteva lucruri: valoarea cea mai mare, valoarea cea mai mică si valorile care se repetă.

Primul pas – aflarea departarii intre valorile extreme: *amplitudinea*. Acesta este in fapt diferenta intre valoarea cea mai mare (X_{max}) si valoarea cea mai mica (X_{min}).

$$10(X_{max}) - 2(X_{min}) = 8(\text{amplitudinea})$$

Adaugind 1 la acesta diferenta aflam cite valori distincte avem in sirul de date.

$$X_{max} - X_{min} + 1$$

$$10 - 2 + 1 = 9$$

Asadar vom avea 9 valori intregi, distincte. Putem sistematiza in continuare materialul trecind intr-o coloana valorile diferite in ordine descrescinda (10, 9, 8,) si notam in dreptul fiecareia, de cite ori se repeta acesta in sirul initial.

Putem determina in acest fel efectivele corespunzatoare fiecarei valori, realizindu-se in acest fel o distributie de efective sau de frecventa. Acest lucru se mai poate scrie si sub forma unui tabel de frecventa cu forma:

TABELUL 1. FRECVEN E SIMPLE

Valoare	fa
10	2
9	2
8	5
7	3
6	7
5	1
4	4
3	0
2	1
Total	$\Sigma fa=25$

Am realizat astfel un protocol sistematizat, clasificat. Sintem aici la nivelul procedurilor statistice liminare adica in pragul analizei statistice, pregatind aplicarea tehnicilor ulterioare. La acest nivel datele pot fi reprezentate si sub diferite forme grafice (diagrama in dreptunghiuti, in stelute, bare etc.)

Dacă luăm în considerare seria de valori de mai sus, un tabel al frecvențelor simple (absolute) este compus din lista valorilor distincte, ordonate descrescător, la care se adaugă frecvența absolută (**fa**) a fiecărei valori (de câte ori se întâlnește în cadrul seriei fiecare valoare).

Se observă astfel că datele au un caracter mai ordonat iar coloana frecvențelor absolute scoate în evidență anumite aspecte cum ar fi, de exemplu, faptul că cea mai frecventă valoare este 6 (apare de 7 ori). Observăm că seria de valori din tabel include toate valorile posibile între valoarea cea mai mare (10) și cea mai mică (2), incluzând

și valorile care nu se întâlnesc în mod real în cadrul seriei. În cazul nostru avem valoarea 3, cu frecvența de apariție 0. Suma frecvențelor absolute (Σfa) indică totalul valorilor din cadrul seriei (25).

În practică, pe lângă frecvențele absolute se iau în considerare și alte tipuri de frecvențe (tabelul 2):

- **Frecvența cumulată (fc).** Totalul valorilor care se cumulează începând de la valoarea cea mai mare până la valoarea cea mai mică din tabel. De exemplu, în tabelul sintetic de mai jos,

avem 6 valori mai mici sau egale cu 5, 21 de valori mai mici sau egale cu 8 și, evident, 25 de valori mai mici sau egale cu 10.

- **Frecvența relativă raportată la unitate $fr(1)$.** Este raportul dintre frecvența absolută și suma frecvențelor absolute ($fa/\Sigma fa$).
 - Exemple: pentru valoarea 10: $fa/\Sigma fa=2/25=0.08$;
- **Frecvența relativă procentuală $fr(\%)$:** Exprimă procentul valorilor care se situează până la o anumită valoare din cadrul distribuției. Se calculează fie prin înmulțirea **$fr(1)$** cu 100, fie prin calcularea directă a procentului pe care îl reprezintă o anumită valoare raportat la totalul valorilor dintr-o distribuție. Suma frecvențelor relative procentuale este întotdeauna egală cu 100.
 - Exemplu: 8% dintre subiecții evaluați au realizat 10 răspunsuri corecte
- **Frecvența relativă cumulată procentuală ($frc\%$):** Exprimă procentul valorilor dintr-o distribuție care se plasează până la o anumită valoare (inclusiv aceasta).
 - Exemplu: 52% dintre studenți au obținut o notă egală sau mai mică de 6
 - Frecvența relativă procentuală cumulată se numește **rang percentil**. Astfel, despre valoarea 6 din distribuția de mai sus se poate spune că are rangul percentil 52, adică, 52% dintre valorile unei distribuții sunt între cea mai mică valoare și valoarea 6, inclusiv.
 - În mod complementar, numim **percentilă**, valoarea dintr-o distribuție care corespunde unui anumit rang percentil. În exemplul de mai sus, rangului percentil 52 îi corespunde valoarea 6, numită, de aceea, percentila 52.
 - În practică, există anumite percentile care au o importanță aparte. Acestea sunt percentilele corespunzătoare rangurilor percentile cu valorile 10, 20, 30,..., 100. Despre semnificația lor vom vorbi mai târziu în acest curs. De asemenea, se utilizează termenul de quartile pentru pecentilele care împart distribuția în patru zone egale ca număr de valori. Acestea sunt corespunzătoare rangurilor percentile de 25, 50 și 75. Cu alte cuvinte, valoarea dintr-o distribuție până la care se află 25% din valori este percentila 25, valoarea până la care se află 50% este percentila 50, iar valoarea până la care se află 75% din valori este percentila 75.

Tabelul 2. Tabloul sintetic al frecvențelor simple

Valoare	Fa	Fc	Fr (1)
10	2	25	
9	2	23	
8	5	21	
7	3	16	
6	7	13	
5	1	6	
4	4	5	
3	0	1	
2	1	1	
Total	$\Sigma fa=25$		

DISTRIBUȚIA DE FRECVENȚE GRUPATE

Prelucrarea realizată mai sus este indicată doar în cazul existenței unui număr redus de date, adică întinderea variației (amplitudinea) este redusă.

Când amplitudinea este relativ mare urmează să facem o distribuție mai compactă pentru a asigura o concentrare suficientă a informației.

Să presupunem că valorile de mai jos reprezintă distribuția variabilei „memorie consemne” măsurată prin aplicarea unui test specific.

Tabel 3.

Datele din tabel sunt deja ordonate crescator.

1. Se impune atunci gruparea valorilor in clase pentru relevarea regularitatii de ansamblu a datelor. Dincolo de sugestiile statisticienilor in acest sens (ex. H. Sturges – relatia empirica) alegerea numarului de clase, dupa un interval de grupare “i” este o operatiune in buna masura la latitudinea cercetatorului (experienta acestuia) dar ea trebuie sa tina cont totusi de o serie de recomandari:

- nu exista, teoretic, un numar optim de intervale insa alegerea acestora nu trebuie sa influenteze tendinta de ansamblu a fenomenului studiat;
- numarul de clase sa nu depaseasca 20, dar sa se evite frecventele mici pe clasa
- in cercetarile psihologice se alege de obicei un numar impar de clase (3, 5, 7, 9, 11) pentru o serie de avantaje (observarea simetriei distributiei, existenta unei clase centrale) ;
- pentru ca gruparea reprezinta un moment important in organizarea datelor se poate apela la mai multe variante de grupare pina se ajunge la situatia optima.

2. Precizarea intervalului de grupare “i” se face cu formula:

$$i = \frac{X_{\max} - X_{\min} + 1}{k} \text{ unde } k = \text{nr. de clase}$$

In exemplul nostru, alegind un numar de 9 clase vom avea

$$i = \frac{79 - 26 + 1}{9} = 6$$

Vom determina clasele si apoi cite valori din sirul initial intra in fiecare clasa determinind in acest fel efectivele pe fiecare interval.

		Determinarea efectivelor		
		\\		
		\\\\\\\\		
		\\\\\\\\\\\\\\\\		
		\\\\\\\\\\\\\\\\\\\\		
		\\ \\		
		\\ \\\\\\\\\\\\\\\\		
		\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\		
		\\\\\\\\\\\\\\\\\\\\		

		\\ \\ \\		
		\\ \\ \\		

Daca insumam coloana efectivelor obtinem **N** (efectul total al grupului).

Pentru ca, de multe ori, este nevoie de tabele comparative se apeleaza la exprimarea procentuala a frecventelor pe interval. Procedul de calcul este simplu – regula de trei simple: N se ia 100% si frecventa pe fiecare interval. Se obtine o noua coloana in procente care, totalizate, trebuie sa dea 100 sau ceva f. apropiat.

Incheind operatia de grupare, se obtine o **distributie statistica** care formeaza baza prelucrarii statistice a materialului.

Cind dispunem de computer si soft aferent, toate aceste operatiuni pot fi executate foarte usor.

Limite „aparente” și limite „reale” ale intervalelor de clasă

În exemplul de mai sus distribuția conține valori, exprimând, să zicem, scorul obținut la test în numere întregi. Dar dacă rezultatul ar fi

măsurat în numere zecimale? (ex. 67,7). Într-un asemenea caz limitele intervalelor din tabelul de mai sus ar fi limite „aparente” deoarece prin rotunjire, valorile zecimale cu mai mult de 5 zecimi ar fi aduse la întregul următor, în timp ce valorile cu până la de 5 zecimi s-ar rotunji în jos (acest mod de rotunjire este unul convențional). Cu alte cuvinte, 67.7 ar intra în intervalul 68-73. Astfel, limitele „reale” ale intervalelor, deși sunt cu numere zecimale, vor apare ca numere întregi (limite „aparente”).

REPREZENTĂRI GRAFICE

Reprezentările graficele sunt forme intuitive de prezentare a distribuțiilor de frecvențe („o imagine *face mai mult decât o mie de cuvinte*”). Ele sunt foarte frecvent utilizate pentru analiza și prezentarea datelor în psihologia aplicată deoarece facilitează înțelegerea semnificației datelor numerice. În prezent, programele computerizate oferă mijloace extrem de puternice și de sofisticate pentru elaborarea reprezentărilor grafice. Dar simpla utilizare a unui astfel de program nu garantează realizarea unui grafic eficient. În esență, un grafic eficient este o combinație reușită între formă și conținutul statistic pe care îl reflectă. Realizarea acestei combinații depinde de respectarea câtorva principii esențiale:

- focalizarea pe conținutul și nu pe forma graficului
- este recomandabil să fie utilizate grafice care favorizează comparații între variabile și nu doar reprezentări individuale, „statice”, ale acestora
- fiecare grafic trebuie să servească un singur scop, exprimat clar și evident
- orice grafic va fi însoțit de informații statistice și descrierile necesare pentru a fi ușor și corect înțeles
- un grafic trebuie să scoată în evidență datele și nu abilitățile tehnice de editare ale celui care l-a creat.

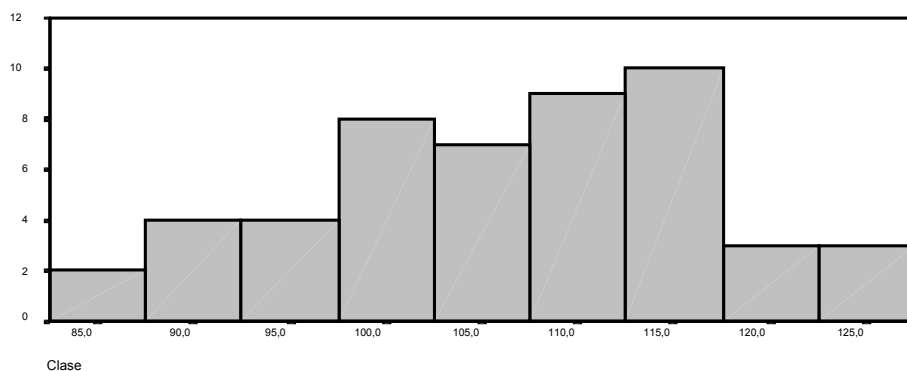
Formele de expresie grafică a datelor statistice sunt foarte numeroase. Cele mai utilizate metode în grafică: **diagrama, histograma, poligonul frecvențelor, curba frecvențelor cumulate, graficul circular și graficul de tip „stem and leaf”** („tulpină și frunze”).

DIAGRAMA

Reprezintă cel mai simplu mod de reprezentare grafică a datelor. Se utilizează atunci când dorim să reprezentăm o variabilă „discretă” (care prezintă valori întregi, de exemplu, numărul de răspunsuri corecte la un test în funcție de nivelul de instruire al subiecților).

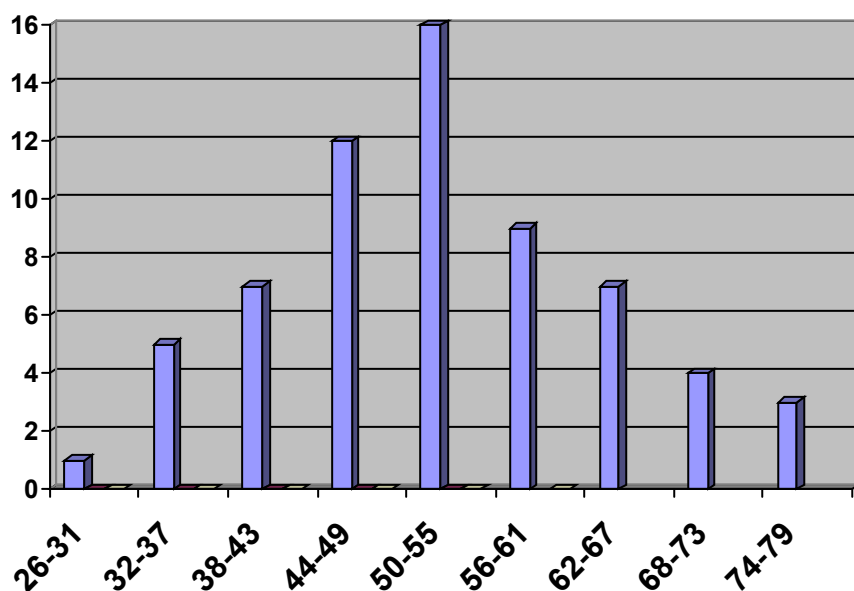
În mod obișnuit, un grafic se prezintă ca o imagine inclusă într-un sistem de axe perpendiculare:

- Axa orizontală (Ox) pe care sunt reprezentate valorile distribuției
- Axa verticală (Oy) pe care sunt reprezentate frecvențele fiecărei valori, sub forma unei bare rectangulare.



HISTOGRAMA

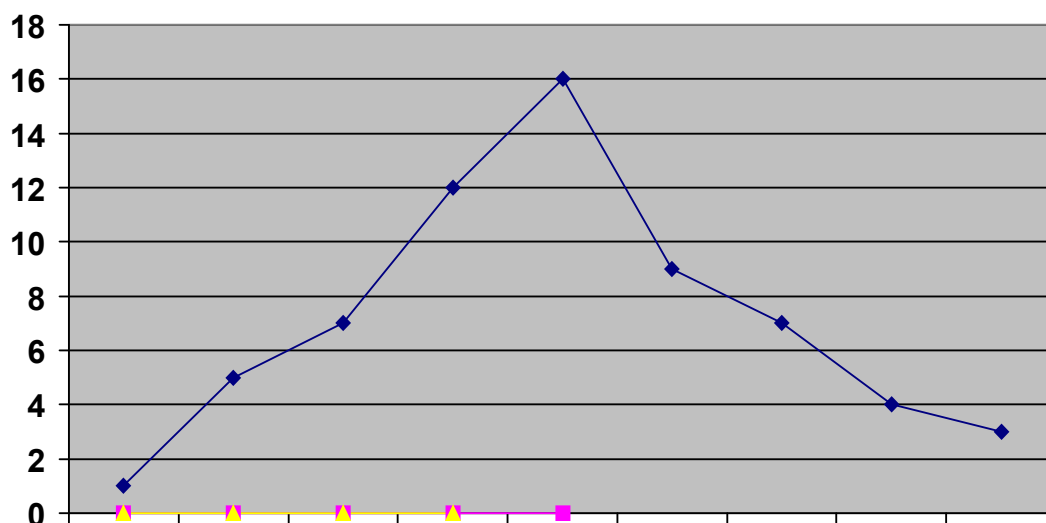
La prima vedere, histograma este asemănătoare cu graficul de tip bară. Ea este adecvată pentru situațiile când variabila pe care dorim să o reprezentăm este de tip „continuu” (adică poate lua orice valoare pe o scală numerică). Iată, de exemplu, histograma distribuției de frecvențe din tabelul 3 (realizată cu programul SPSS):



Se observă faptul că programul a realizat automat o grupare de frecvențe, afișând pe axa Ox limita minimă a intervalului ca „etichetă” a acestuia.

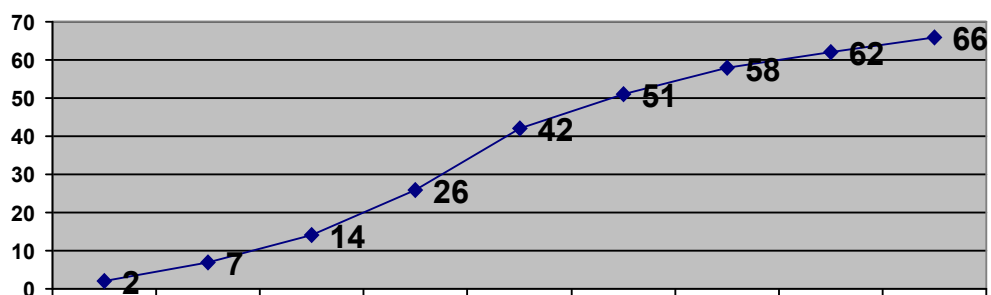
POLIGONUL DE FRECVENTE

Este o reprezentare alternativă la histogramă. Punctele centrale ale suprafețelor rectangulare care reprezintă frecvența sunt unite cu o linie care delimitează suprafața poligonului.



GRAFICUL FRECVENȚEI CUMULATE

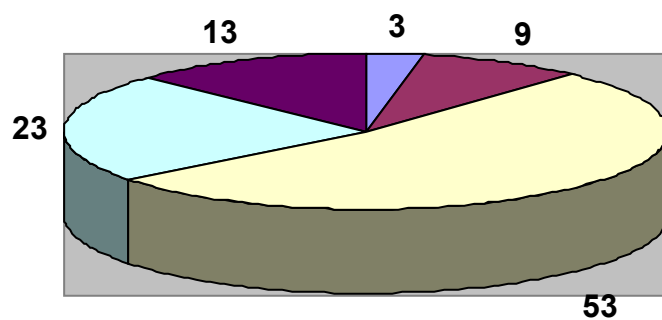
Este un grafic de tip liniar care reprezintă valorile frecvenței absolute cumulate. Pe acest grafic se vede cu ușurință câte valori se află până la o anumită valoare din distribuție (datele reprezentate sunt cele din tabelul 3).



GRAFICUL CIRCULAR

În practica psihologică se utilizează mai des diagramele circulare la construirea cărora se utilizează cercul. Acesta se împarte în sectoare proporționale cu valorile pe care le reprezintă.

Obs. De reținut că diagrama circulară se utilizează în cazurile în care datele



se pretează la prezentare în indici procentuali.

Concluzii.

Utilizarea tabelelor de frecvență și a reprezentărilor grafice aduce un important câștig în analiza datelor statistice. Este însă, important să reținem faptul că atât tabele cât și reprezentările grafice nu sunt decât începutul analizei datelor nu și sfârșitul acesteia. Cu alte cuvinte, nu vom putea trage direct concluzii pe baza lor. Ele pot fi utilizate, însă, pentru a ilustra concluzii, care devin astfel mai ușor de înțeles și de reținut. În fazele primare de analiză a datelor statistice, graficele ne ajută să ne facem o imagine generală asupra acestora, lucru util pentru alegerea procedurilor statistice. Este important să alegem tipul de grafic adecvat în raport cu natura datelor și cu ideea pe care dorim să o ilustrăm. În practică, graficele se realizează utilizând programe specializate, iar SPSS are proceduri puternice de realizare a unei largi varietăți de grafice.

INDICATORI NUMERICI AI DISTRIBUȚIILOR STATISTICE

Analiza de frecvențe este o metodă utilă pentru punerea în valoare a unor caracteristici ale distribuțiilor statistice. Cu toate acestea, ea este tributară necesității de a manipula întreaga cantitate de date, toate valorile unei distribuții (fie ele și grupate). Pentru a elimina acest neajuns sunt utilizați așa numiții *indicatori sintetici*. Aceștia sunt descriptori numerici care condensează într-o valoare unică o anumită caracteristică a unei întregi distribuții de valori. Principalele avantaje pe care le oferă sunt concentrarea semnificației și ușurința utilizării.

Tipuri de indicatori sintetici:

Trei sunt caracteristicile distribuțiilor care sunt evaluate cu ajutorul indicatorilor sintetici: tendința centrală, variabilitatea (împrăștierea) și forma distribuției. Pentru fiecare din aceste caracteristici se utilizează anumiți indicatori specifici:

- *Indicatori ai tendinței centrale:* valori tipice, reprezentative, care descriu distribuția în întregul ei;
- *Indicatori ai variabilității:* valori care descriu caracteristica de împrăștiere a distribuției;
- *Indicatori ai formei distribuției:* valori care se referă la forma curbei de reprezentare grafică a distribuției, prin comparație cu o curbă normală (oblicitate, applatizare).

INDICATORII TENDINTEI CENTRALE

Indicii prin care se determina în mod curent “tendința centrală” a rezultatelor și pe care cercetatorul trebuie să-i determine sunt: media, mediana și modul.

MEDIA ARITMETICĂ (m)

Media (m) este un descriptor al unui ansamblu de date care exprimă suma valorilor datelor numerice, împărțită la numărul acestora.

Notății uzuale:

- μ (miu), atunci când este media întregii populații de referință
- m , atunci când se calculează pentru un eșantion (cazul cel mai frecvent)

Calcularea mediei pentru o distribuție simplă de frecvențe se face prin adunarea valorilor și se împărțirea la numărul lor

$$\bar{m} = \frac{\sum x}{N}$$

in care

Σ ="suma de"

x=rezultatele, valorile

individuale

N=efectivul grupei studiate

PROPRIETĂȚILE MEDIEI ARITMETICE

- Adăugarea\scăderea unei constante la fiecare valoare a distribuției, mărește\scade media cu acea valoare
- Înmulțirea\împărțirea fiecărei valori a distribuției cu o constantă, multiplică\divide media cu acea constantă
- Suma abaterii valorilor de la medie este întotdeauna egală cu zero

ASADAR PRIN CALCULAREA MEDIEI OBTINEM O MASURA A NIVELULUI MEDIU RELATIV LA UN ESANTION, FAPT CE VA PERMITE COMPARATII INTRE GRUPE.

Combinarea mediilor provenind de la esantioane diferite

Sint cazuri cind sintem nevoiti sa aglutinam doua sau mai multe grupuri independente pentru a obtine o medie comuna. Se recomanda pentru aceasta utilizarea formulei:

$M = \frac{n_1m_1 + n_2m_2 + \dots + n_km_k}{N}$ in care

m_1, m_2, m_k – media grupelor independente

n_1, n_2, n_k – numarul subiectilor grupelor

N=nr. total de subiecti

Formula de contopire a mediilor se poate extinde la k grupe in functie de necesitati.

MEDIANA (M_E)

Este valoarea „din mijlocul” unei distribuții, adică aceea care are 50% dintre valori deasupra ei și 50% dintre valori dedesubtul ei (percentila 50).

Pentru a găsi mediana trebuie să aranjăm toate datele (valorile) după rang, de la valoarea cea mai mică la cea mai mare (sau invers). Mediana este acea valoare care **împarte sirul ordonat de valori** în două grupe egale ca număr. Altfel spus, mediana se găsește la mijlocul sirului. **Locul sau rangul** pe care-l ocupă mediana în sirul ordonat de date se determină cu ajutorul formulei: **$N+1 \div 2$** .

Atentie! Formula $N+1 \div 2$ ne indică locul mediane și nu valoarea ei.

În cazul distribuțiilor cu număr impar de valori, M_E este chiar valoarea respectivă.

În cazul distribuțiilor pare, M_E se calculează ca medie a celor două valori din mijlocul distribuției

Exemplu:

În seria de valori 5,8,3,2,5,4, ordonată crescător (2,3,4,5,5,8), $M_E=4,5$ (ca medie a valorilor 4 și 5 aflate în mijlocul unei distribuții pare). Dacă distribuția noastră ar fi avut 5 valori (fără 2, de exemplu), $M_E=5$

MODUL (M_o)

Modul este valoarea care se repetă cel mai des într-un sir de date, adică valoarea cu frecvența cea mai mare. Modul poate fi determinat prin simpla examinare a valorilor, fără a se recurge la operații de calcul.

Exemplu : în sirul de date 3, 4, 5, 6, 7, 7, 7, 8, 8, 10, 11

$M_o=7$

Pot exista cazuri extreme când un sir de date poate prezenta două valori cu frecvențe maxime egale; în acest caz vorbim de două module. Când cele două valori de frecvențe egale sunt vecine în sirul de date, modul va fi reprezentat de valoarea intermediară a celor două. De asemenea, în cazul unui sir de date în care fiecare valoare apare o singură dată (frecvență=1) nu se poate determina modul.

Distribuțiile pot avea un singur mod (*unimodale*), două moduri (*bimodale*) sau mai multe (*multimodale*)

Avantajele și dezavantajele indicatorilor tendinței centrale

Tabloul de mai jos prezintă, în mod sintetic avantajele și dezavantajele specifice indicatorilor tendinței centrale:

	AVANTAJE	DEZAVANTAJE
MODUL	<ul style="list-style-type: none"> - Ușor de aflat (nesemnificativ în prezent); - Poate fi utilizat pentru orice tip de scală; - Este singurul indicator pentru scale nominale;- - Corespunde unui scor real al distribuției; 	<ul style="list-style-type: none"> - În general, nesigur, mai ales în cazul eșantioanelor mici, când se poate modifica dramatic la o modificare minoră a unei valori; - Poate fi greșit interpretat. Se identifică total cu un scor anume, fără a spune nimic despre celelalte valori; - Nu poate fi utilizat în statistici inferențiale;
MEDIANA	<ul style="list-style-type: none"> - Poate fi utilizată pe scale ordinale și de 	<ul style="list-style-type: none"> - Poate să nu corespundă unei valori reale (N par);

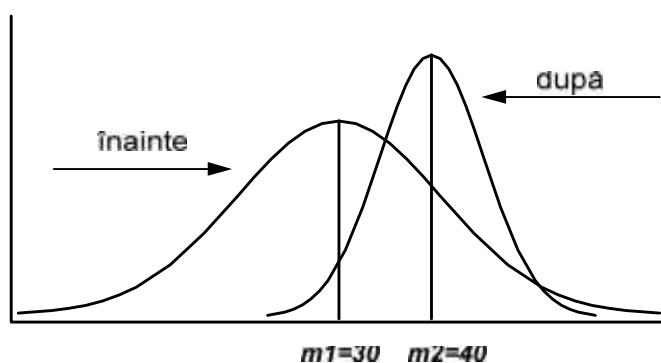
	<p>interval\raport;</p> <ul style="list-style-type: none"> - Poate fi utilizată și pe distribuții de frecvență cu clase deschise sau scoruri nedeterminate la marginile distribuției; 	<ul style="list-style-type: none"> - Nu reflectă valorile distribuției (un scor extrem se poate modifica, fără a afecta Me); - Este mai puțin sigură în extrapolarea de la eșantion la populație; - Greu de utilizat în statistici avansate
MEDIA	<ul style="list-style-type: none"> - Reflectă valorile întregii distribuții; - Are multe proprietăți statistice dezirabile; - Adecvată pentru utilizare în statistici avansate; 	<ul style="list-style-type: none"> - De obicei nu corespunde unei valori reale; - Nu este tocmai adecvată pentru scale ordinale; - Conduce la interpretări greșite pe distribuții asimetrice - Poate fi puternic afectată de scorurile extreme;

INDICATORI SINTETICI AI ÎMPRĂȘTIERII

Indicatorii tendinței centrale caracterizează un singur aspect al distribuției statistice. Două distribuții statistice pot avea aceeași medie dar ele pot fi foarte diferite sub aspectul variabilității, al omogenității.

Indicatorii împrăștierii se referă la caracteristica de variabilitate, care descrie diferențele existente între valori. În cazul tendinței centrale este scoasă în evidență caracteristica valorilor unei distribuții de a se „asemăna” unele cu altele, „asemănare” surprinsă de indicatorii tendinței centrale. În cazul împrăștierii, se urmărește descrierea tendinței valorilor de a se deosebi una de alta, de a se „sustrage” unei tendințe centrale prin îndepărtarea de aceasta. De exemplu, o distribuție de tipul 6,5,6,5,4,5 este, evident, mult mai omogenă (mai puțin variabilă) decât o distribuție de genul 2,3,4,5,8,9. Cu cât valorile diferă mai mult una de alta, cu atât variabilitatea distribuției este mai mare.

Să ne imaginăm următoarea situație: Un psiholog școlar vrea să vadă efectul unei metode de creștere a încrederii în sine pe un lot de elevi. În acest scop, el evaluează încrederea în sine înainte și după o serie de ședințe specifice. Distribuția valorilor este reprezentată în imaginea de mai jos:



Așa cum se observă, valorile încrederii măsurate înainte de intervenție au o medie de 30 și o împrăștiere (neomogenitate) mai mare, în

timp ce valorile de după intervenție prezintă o medie de 40 și o împrăștiere mai mică, (sunt mai omogene). Acest fapt sugerează că tratamentul psihoterapeutic a avut efect. Imaginea scoate în evidență și faptul că în distribuțiile mai omogene media este mai reprezentativă decât în distribuțiile mai puțin omogene.

Pentru evaluarea împrăștierii distribuțiilor statistice se utilizează mai mulți indicatori.

1. AMPLITUDINEA (R DE LA *RANGE*)

Este dată de diferența dintre valoarea **maximă** și valoarea **minimă** a unei distribuții

$$R = X_{\max} - X_{\min}$$

Utilitatea ei este dată de faptul că ne indică în mod absolut plaja de valori între care se întinde distribuția.

Principalul dezavantaj constă în faptul că poate fi influențată de o singură valoare aflată la extremitatea distribuției.

2. ABATEREA INTERQUARTILĂ (R_{SQ}):

Semnifică distanța unui scor „tipic” față de amplitudinea întregii distribuții și se calculează ca media diferenței dintre quartila 3 și quartila 1.

$$R_{SQ} = \frac{Q_3 - Q_1}{2}$$

Într-o distribuție perfect simetrică $R_{SQ} = Q_2 = Me$

R_{SQ} nu este afectată de valorile aberante fiind considerată, din acest motiv, un indicator „robust” al împrăștierii

Acest tip de indicator ilustrează împrăștierea prin distanța dintre două puncte ale unei distribuții. Unul dintre avantajele lor este acela al

ușurinței de calcul. Pe de altă parte, tocmai pentru că iau în seamă doar două dintre valorile distribuției, sunt vulnerabili și nesiguri. Utilitatea lor este în general limitată dar sunt singurii care pot fi folosiți atunci când indicatorii de care vom vorbi în continuare, nu pot fi calculați.

Spre deosebire de primii doi indicatori (elementari), următorii indicatori (sintetici) surprind împrăștierea unei distribuții prin luarea în considerare a abaterii fiecărei valori de la un anumit indicator al tendinței centrale. Cel mai uzual indicator de referință pentru împrăștiere este media. Aceasta pentru că, așa cum ne amintim, media are avantajul de a fi o „concentrare” a tuturor valorilor unei distribuții.

3. ABATEREA MEDIE

4. DISPERSIA (VARIANȚA)

Notății uzuale:

s^2 (când se calculează pentru eșantion)

σ^2 (când se calculează pentru întreaga populație)

Pentru a elimina inconvenientul abaterilor de la medie de a avea, se operează ridicarea la pătrat a abaterilor valorilor individuale.

Dacă însumăm abaterile ridicate la pătrat (pătratic) și le împărțim la numărul valorilor, obținem *dispersia* (numită și *varianță* sau *abatere medie pătratică*)

$$s^2 = \frac{\sum (X_i - m)^2}{N}$$

Totuși, din cauza ridicării la pătrat, dispersia nu reprezintă o valoare foarte bună a împrăștierii (de ex., poate fi mai mare decât amplitudinea distribuției). Soluția acestui neajuns o constituie:

5. ABATEREA STANDARD

Notații uzuale:

s (pentru eșantioane)

σ (pentru populație)

SD (*Standard Deviation*, în standardul APA)

ab.std.

Abaterea standard se obține prin extragerea radicalului din expresia abaterii medii pătratice (dispersiei).

Formula de calcul:
$$s = \sqrt{\frac{\sum (X_i - m)^2}{N}}$$

Operațiile succesive efectuate mai sus, ridicarea la pătrat și extragerea radicalului, nu trebuie văzute ca operații artificiale, „gratuite”. Aceste operații nu se referă la valorile distribuției ci la abaterile de la medie, ceea ce conduce la rezultate diferite care exprimă, într-o altă formă, aceeași caracteristică de împrăștiere a valorilor originale.

SEMNIFICATIA ABATERII STANDARD

Abaterea standard poate fi luata ca unitate de masura pe abscisa unei curbe de frecventa simetrice, acesta mai ales pentru ca ea se exprima in aceleasi unitati ca si cotele brute cu care se lucreaza. Astfel se poate pleaca de la punctul de pe abscisa care corespunde mediei esantionului in studiu si se adauga , respectiv se iau succesiv in dreapta si din stinga mediei – sens crescator si respectiv descrescator - unitati de abateri standard. Se obtin astfel punctele din dreapta: 1s, 2s, 3s, sens crescator al cotelor brute, si din stinga -1s, -2s, -3s sens descrescator al cotelor brute.

(vezi fig. din anexa - curba cu sase repere si procente pe intervale)

Deci vom avea:

-3s, -2s, -1s, m, 1s, 2s, 3s

Intr-o distributie simetrica exista aproximativ 3s sub medie si 3s deasupra mediei ceea ce se scrie +3s. Astfel amplitudinea sau intinderea variatiei este de aproximativ 6s iar abaterea standard este a sasea parte din amplitudine.

Apar astfel arii a unor suprafețe de sub curba normală ce pot servi la estimarea unor probabilități, exprimate și sub formă procentuală:

- între $-s$ și $+s$ se afla aproximativ 68% din totalul valorilor
- între $-2s$ și $+2s$ se afla aproximativ 95% din tot. valorilor
- între $-3s$ și $+3s$ se afla cam 99,95% din total adică aproape toate valorile

Să reținem deci ca:

- s măsura distanța la care se afla o cota oarecare în raport cu m
- există o interpretare procentuală a celor 6 puncte de reper
- abaterea standard se determină doar în cazul distribuțiilor normale, simetrice.

COEFICIENTUL DE VARIAȚIE

Abaterea medie și abaterea standard se exprimă în unitățile de măsură ale variabilei de referință. De exemplu, pentru o distribuție de timpi de reacție, exprimați în sutimi de secundă, $s=2.14$ înseamnă că împrăștierea standard este de 2.14 sutimi de secundă.

Dacă același eșantion face și un test de coordonare a mișcărilor, evaluat în număr de „ieșiri din traseu” a căror abatere standard este $s=20.94$, nu putem compara omogenitatea celor două serii de valori. Adică, nu putem spune dacă eșantionul este mai omogen sau mai puțin omogen din perspectiva uneia dintre cele două performanțe.

Dintre soluțiile posibile pentru eliminarea acestui neajuns, cea mai des utilizată este coeficientul de variație (variabilitate), notat cu cv (sau v), propus de Pearson. Se calculează ca raport între abaterea standard și medie. Poate fi exprimat și procentual conform formulei de mai jos:

$$cv = \frac{s}{m} * 100$$

Valoarea acestui coeficient exprimă un raport procentual dintre abaterea standard și medie. Cu cât este mai mare, cu atât putem spune că media este mai puțin „reprezentativă” pentru distribuția respectivă, dată fiind ponderea ridicată a împrăștierii. Utilizarea coeficientului de variație este limitată la valorile măsurate pe scale de raport, cu origine naturală 0. În

cazul a două variabile a căror origine este diferită una de alta, diferențele dintre valori (abaterea standard) rămân aceleași dar media se schimbă, fapt care face ca raportul exprimat în formulă să fie modificat iar comparația a doi coeficienți de variație, irelevantă.

Utilitatea coeficientului de variație vine de la faptul că valoarea sa mai este legată de unitatea de măsură. Diferența dintre două valori cv poate fi interpretată ca diferență de împrăștiere a celor două variabile, chiar dacă măsoară lucruri diferite.

Sunt propuse anumite limite de interpretare a acestui indicator, astfel:

- dacă $cv < 15\%$, împrăștierea este mică și, deci, media este reprezentativă
- dacă cv este cuprins între 15% - 30% , împrăștierea este mijlocie și media este suficient de reprezentativă
- dacă cv este mai mare de 30% , împrăștierea este mare și media are o reprezentativitate redusă

Calcularea coeficientului de variație a unei distribuții, înainte de integrarea ei în proceduri statistice inferențiale, este o metodă utilă de verificare a măsurii în care media, pe care se bazează de cele mai multe ori procedurile inferențiale, este legitimă.

Alegerea indicatorului împrăștierii

- Abaterea standard este cea mai utilizată pentru scale de măsurare interval/raport.
- Amplitudinea este un indicator nesigur și care nici nu poate fi calculat în cazul scalelor nominale
- Pe distribuții asimetrice, se alege abaterea interquartilă (semi-interquartilă).

VALORI EXTREME SI FORME ALE UNEI DISTRIBUTIEI

Intilnim adeseori intr-o distributie valorile extreme, neobișnuit de mari sau de mici față de celelalte valori. Efectul lor influenteaza valorile tendinței centrale, în primul rând media.

Se pune astfel problema modului lor de tratare:

Se va urmări mai întâi natura unor astfel de valori:

- erori de înregistrare;
- erori de măsurare;
- rezultate influențate de inadvertențe ale condițiilor experimentale;
- eșantionul a fost extras dintr-o populație asimetrică;

și se va proceda la eliminare (dacă sunt erori necorectabile) sau la corectare (dacă este posibil);

Analiza valorilor extreme reprezintă unul dintre obiectivele principale ale fazelor preliminare de analiză a datelor. Prezența lor este de natură să aibă efecte majore asupra rezultatelor fapt care trebuie luat în considerare la alegerea procedurilor statistice inferențiale.

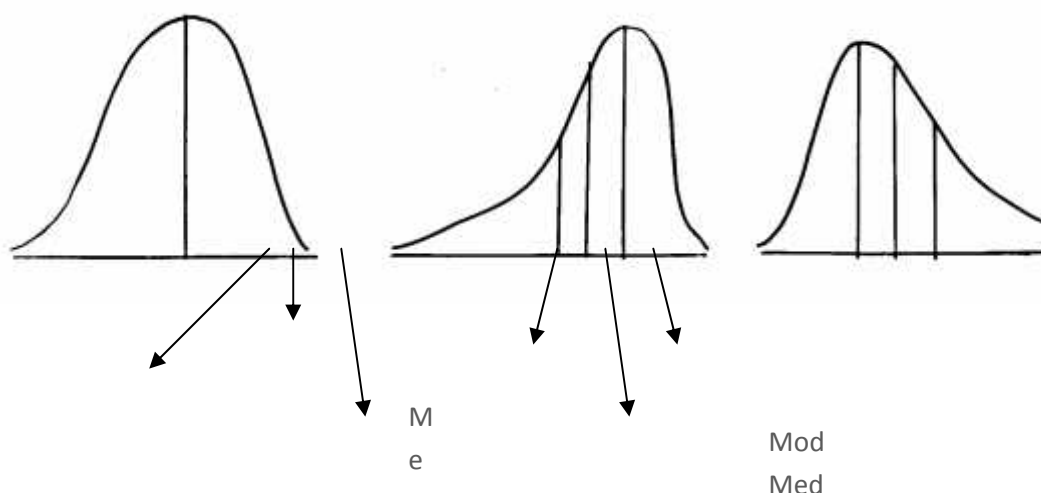
INDICATORI AI FORMEI DISTRIBUTIEI

Expresia grafică a distribuțiilor poate fi descrisă sub două aspecte esențiale: *simetria și boltirea*.

O distribuție este *simetrică* atunci când valorile acesteia se împart în mod egal de o parte și de alta a valorilor tendinței centrale. Se numesc asimetrice (*skewed*) distribuțiile ale căror valori se concentrează fie în zona valorilor mici (spre stânga) fie în zona valorilor mari (spre dreapta)

Distribuție: simetrică
asimetrică pozitiv (stanga)

asimetrică negativ (dreapta)



Figurile de mai sus arată cum se plasează cei trei indicatori ai tendinței centrale în funcție de simetria distribuției:

- În cazul distribuțiilor (perfect) simetrice, **Mo**, **Me** și **m** se plasează pe aceeași valoare
- În cazul distribuțiilor asimetrice cei trei indicatori au poziții diferite
- Mediana se plasează întotdeauna între mod și medie. Din acest motiv, mediana este cea mai reprezentativă valoare pentru distribuțiile asimetrice
- Media este afectată de valorile extreme, cu atât mai mult cu acestea sunt mai puternic deviate. Ca urmare, în cazul distribuțiilor puternic asimetrice, media nu este un indicator adecvat al tendinței centrale.

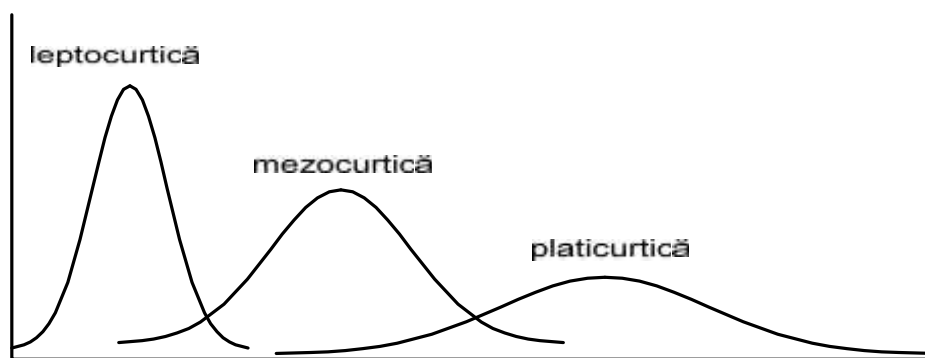
Descrierea numerică a caracteristicii de simetrie/asimetrie se face cu ajutorul unui indicator statistic specific, numit indicator de „simetrie” sau de „oblicitate” (*skewness*, în limba engleză). El este ușor de obținut cu ajutorul programelor specializate.

VARIAȚIA ACESTUI INDICATOR

Pentru o curbă absolut simetrică, indicele de oblicitate (*skewness*) are valoarea 0 (zero). Curbele asimetric pozitive au indici de oblicitate cu valori pozitive iar cele asimetric negative sunt semnalate de indici cu valori negative. Ca reper general de apreciere, recomandat de cei mai mulți autori, un indice de oblicitate a cărui valoare depășește +1/-1 semnalează o asimetrie pronunțată a distribuției.

Caracteristica de boltire (*kurtosis*, în terminologia engleză) indică gradul de extindere pe verticală a curbelor de distribuție. În termeni generali, sub aspectul boltirii, curbele pot fi de trei categorii:

- leptokurtice, cu majoritatea valorilor distribuite în zona mediei (au o formă „înaltă” și „subțire”)
- mezokurtice, cu o prezență „moderată” a valorilor în zona mediei
- platikurtice, cu valori medii relativ puține și o formă aplatizată



Deși imaginea de mai sus ilustrează boltirea pe curbe simetrice, sigur că o curbă poate fi în același timp și asimetrică și boltită excesiv.

VARIAȚIA ACESTUI INDICATOR

Indicatorul numeric al boltirii (*kurtosis*) are o plajă de variație în jurul valorii zero (care înseamnă boltire medie, „normală”, *mezocurtică*). Indicele de boltire pozitivă indică o curbă „înaltă” (*leptocurtică*), iar indicele de boltire negativă, o curbă „aplatizată” (*platicurtică*). La fel ca și în cazul indicelui de oblicitate (*skewness*), cu cât acesta este mai îndepărtat de valorile $+1/-1$, avem de a face cu distribuții cu abatere accentuată de la boltirea „normală”.

Calcularea indicatorilor de simetrie și de boltire reprezintă modalități importante de apreciere a caracteristicilor unei distribuții. Aceștia trebuie luați în considerare ori de câte ori utilizarea procedurilor statistice inferențiale reclamă anumite caracteristici ale distribuțiilor.

COEFICIENTUL DE CORELAȚIE LINIARĂ (PEARSON)

În practica cercetării pot exista situații în care avem două măsurări efectuate cu instrumente diferite. De exemplu, ne putem întreba dacă „există o legătură între timpul de reacție și extraversiune ca trăsătură de personalitate?”. În aceste cazuri avem două variabile dependente cu valori perechi și nici o variabilă independentă. Pentru astfel de situații problema

care se pune este existența unei relații de variație reciprocă a acelor două variabile. Procedeu statistic utilizat este *coeficientul de corelație*.

Termenul de corelație, înainte de a fi un concept statistic este un cuvânt uzual în limbajul cotidian. În esență, el exprimă o legătură între anumite aspecte ale realității așa cum este ea reflectată în plan observației directe. La nivel statistic, corelația exprimă o legătură cantitativă sistematică între valorile a două variabile perechi, măsurate pe subiecți aparținând aceluiași eșantion de cercetare.

Să presupunem că un grup de elevi au efectuat un test de atenție concentrată și unul altul de atenție distributivă. Dacă pe măsură ce performanța la unul dintre teste crește concomitent cu performanța la celălalt test, vom avea o corelație pozitivă. Dacă, dimpotrivă, creșterea performanței la un test este asociată cu scăderea performanței la celălalt test, ne aflăm în fața unei corelații negative. Este evident că există și posibilitatea ca variația performanței la unul din teste să nu aibă nici o legătură cu variația performanței la al doilea test.

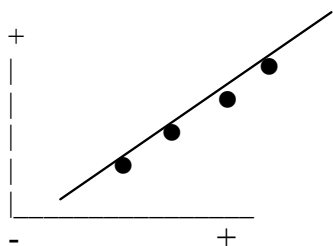
Intensitatea legăturii dintre cele două variabile se exprimă prin *coeficientul de corelație liniară*, notat cu simbolul r . Introdus de K. Pearson, el mai este cunoscut și sub numele de coeficientul de corelație Pearson, sau al „moment-produsului”, după expresia uneia din formulele de calcul.

Reprezentarea grafică a corelației

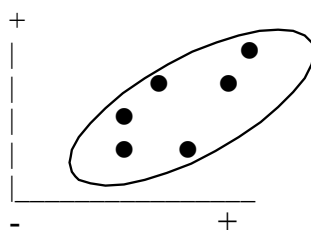
Plasarea valorilor celor două variabile pe un grafic produce o imagine intuitivă a relației dintre valori. Acest tip de grafic se numește *scatterplot*.

În cazul unei corelații pozitive, reprezentările *scatterplot* pot arăta astfel:

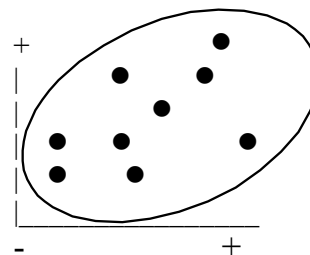
Relație directă – Corelație pozitivă
 $r = 1.00$



$r = .80$



$r = .20$



Tendința este aceea ca valorilor mari de pe axa orizontală să le corespundă valori mari pe axa verticală. În cazul unei corelații pozitive perfecte ($r=+1$), punctele de intersecție ale perechilor de valori se plasează pe o linie. Cu cât corelația este mai mică, cu atât norul de puncte este mai larg dar forma elipsei indică relația pozitivă dintre cele două variabile.

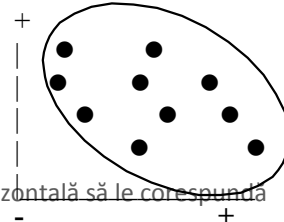
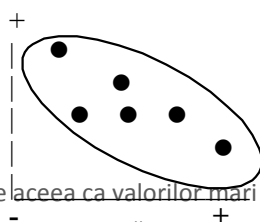
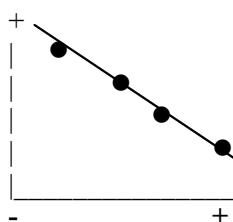
În imaginea de mai jos avem reprezentări *scatterplot* caracteristice pentru corelații liniare negative.

Relație indirectă- Corelație negativă

$r = -1.00$

$r = -.80$

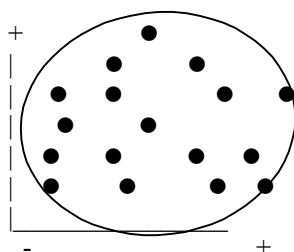
$r = -.20$



Tendința este aceea ca valorilor mari de pe axa orizontală să le corespundă valori mici pe axa verticală. Ca urmare, atât linia corelației negative perfecte ($r=-1$) cât și diagonala mare a elipsei norului de puncte al corelației imperfecte se orientează din stânga sus spre dreapta jos a sistemului de coordonate.

Atunci când corelația dintre cele două variabile este inexistentă, norul punctelor de intersecție are o formă circulară, care nu conturează nici o tendință ($r=0$).

Nici o relație – Nu există corelație $r = 0.00$



Semnificația coeficientului de corelație

Indiferent cât de mare este r calculat, nu putem avea încredere în acesta atâta timp cât nu știm în ce măsură este diferit de un r care ar rezulta prin jocul întâmplării.

Pentru ușurarea evaluării semnificației, a fost creat un tabel special cu praguri de semnificație ale coeficientului de corelație r și care poate fi folosit fără a mai fi necesară utilizarea formulei. Practic, se caută în tabel care este nivelul lui r pentru numărul gradelor de libertate ($df=N-2$) și pragul α ales în prealabil. Dacă valoarea tabelară este cel puțin egală cu valoarea calculată a lui r , atunci ipoteza de nul se respinge, coeficientul de corelație fiind considerat semnificativ.

Interpretarea coeficientului de corelație

Așa cum am spus deja, avem o corelație perfectă atunci când r este egal cu $+1$ sau -1 . O valoare obținută de $(+0.80)$ este apropiată de $+1$ ceea ce ne sugerează că între cele două tipuri de variabile există o legătură. Sigur ca $+0.80$ este mai puțin decât $+1$ dar și mai mult decât, să zicem, $+0.30$. O asemenea interpretare, deși corectă, nu poate fi satisfăcătoare. Se simte necesitatea de a avea un criteriu de valorizare a cuantificării numerice a corelației.

De-a lungul timpului au fost propuse diverse astfel de scale de valorizare, prin atribuirea unor calificative coeficienților de corelație, în funcție de mărimea lor. Această problemă comportă multe discuții iar soluțiile oferite de diferiți autori sunt deseori diferite. Ca regulă generală, toți autorii sunt de acord că valorile sub 0,1 ale coeficienților de corelație trebuie să fie considerate „neglijabile”, chiar și atunci când ating pragul de semnificație statistică.

Hopkins [sugerează](#) interpretarea valorilor coeficienților de corelație după cum urmează:

0.0-0.1	Foarte mic, neglijabil,
0.1-0.3	Mic, minor
0.3-0.5	Moderat, mediu
0.5-0.7	Mare, ridicat, major
0.7-0.9	Foarte mare, foarte ridicat
0.9-	Aproape perfect

Înainte oricărui calificativ însă, prima condiție pentru a lua în considerare existența unei corelații între două variabile rămâne atingerea pragului de semnificație (α). Dacă valoarea lui r corespunde unui nivel α mai mare de 0.05, sau decât alt prag legitim decis de cercetător, existența unei corelații este de luat în seamă, indiferent de mărimea coeficientului Pearson. Aceasta, deoarece nu avem temei pentru a accepta că se îndepărtează suficient de o valoare care ar fi putut decurge din jocul hazardului. În cele din urmă, ce trebuie să luăm în considerare, semnificația sau intensitatea asocierii? Desigur, răspunsul este unul relativ. Dacă finalitatea studiului este aceea de a lua decizii, ca în cazul selecției de personal, de exemplu, se vor căuta valori cât mai mari ale coeficientului de corelație (r). Dar, dacă obiectivul este preponderent teoretic, de a pune în evidență relații „ascunse” între variabile, atunci indiferent de mărimea lor, coeficienții de corelație vor fi luați în considerare (dar numai dacă sunt mai mari de 0.1).

Corelație și cauzalitate

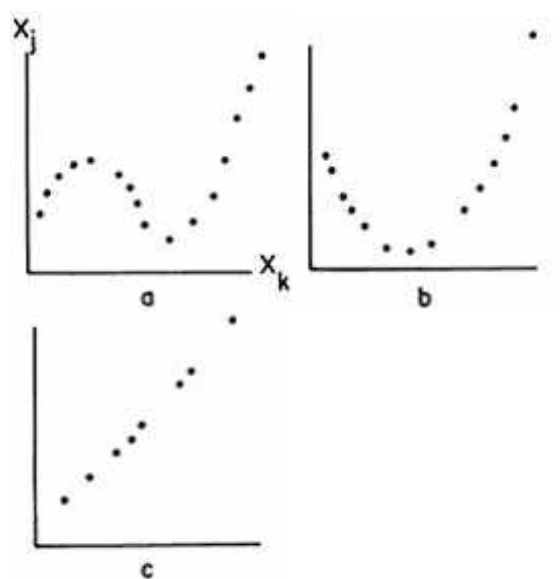
Coeficientul de corelație ne oferă informații despre modul în care variază valorile a două variabile una în raport cu cealaltă. Ca urmare, nu i se poate atribui o semnificație de cauzalitate între variabile decât atunci când cele două variabile au fost măsurate într-un context care probează cauzalitatea. Iar acest lucru se petrece numai în situații de experiment.

Coeficientul de determinare

Valorile lui r trebuie considerate pe o scală ordinală. Cu alte cuvinte, nu este permis să afirmăm că un coeficient de corelație de 0.40 este de două ori mai mare decât un altul de 0.20. Dacă dorim să comparăm în mod direct doi coeficienți de corelație trebuie să ridicăm valorile lui r la pătrat (r^2) obținând astfel ceea ce se numește *coeficient de determinare* (r squared). Pentru exemplificare, $0.85^2 = 0.72$. Dacă citim în procente rezultatul obținut, putem spune că 72% din variația (împrăștierea) uneia dintre cele două variabile este concomitentă cu variația celeilalte variabile. Sau, pentru a fi și mai corecți, cele două variabile au în comun 72% din variația care le caracterizează.

Caracterul liniar al corelației Pearson

Trebuie să reținem că ceea ce exprimă r este nivelul *corelației liniare*, adică măsura în care linia care unește valorile perechi este *rectilinie*. Aceasta este o formă de aproximare a legăturii dintre variabile. În realitate, uneori, corelația dintre două variabile are o formă care se abate de la modelul rectiliniu (este o curbă). Dacă privim imaginile de mai jos, vom vedea câteva tipuri posibile de curbe de corelație. Figurile a și b exprimă corelații perfecte dar care se supun unui model curbiliniu, în timp ce figura c reprezintă o corelație perfectă dar rectilinie.



Calcularea corelației Pearson pentru variabilele reprezentate în figurile a și b, ar conduce la valori mici ale acesteia, în ciuda asocierii grafice evidente a valorilor lor. Din fericire, astfel de situații sunt rare în realitate, modelul corelației liniare fiind adecvat pentru un mare număr de relații dintre variabilele naturale, incluzându-le și pe cele psihologice. Atunci când se raportează un coeficient de corelație fără a se preciza caracterul liniar sau curbiliniu, vom considera că acesta se referă la corelația liniară. Oricum, graficul *scatterplot* oferă informații suplimentare semnificative și, din acest motiv, este recomandabilă analizarea acestuia de fiecare dată când utilizăm testul de corelație Pearson.

Condiții pentru calcularea coeficientului de corelație Pearson

Pentru a putea utiliza în mod legitim calculul de corelație, eșantionul trebuie să fie aleator iar cele două variabile (ambele măsurate pe scale de interval/raport) trebuie să aibă o distribuție care să nu se abată grav de la distribuția normală. Această condiție este cu atât mai importantă cu cât eșantionul este mai mic.

Utilizarea coeficientului de corelație

Analiza de corelație este una dintre cele mai uzuale proceduri statistice în cercetarea psihologică. Printre utilizările cele mai comune menționăm analiza *consistenței* și *validității* testelor psihologice.

Consistența se referă la gradul în care un instrument de evaluare se concentrează asupra unei anumite realități psihice. *Validitatea*, se referă la faptul dacă ceea ce presupune că măsoară un instrument psihologic este măsurat cu adevărat (de exemplu, o scală de autoritarism măsoară cu adevărat autoritarismul).

Din cele prezentate, rezultă că putem utiliza coeficientul atunci când avem serii perechi de distribuții. Coeficientul poate fi aplicat atât pentru variabile măsurate cu aceeași unitate de măsură cât și pentru variabile exprimate în unități de măsură diferite. Aceasta deoarece formula de calcul ia în considerare expresia standardizată a valorilor (scorurile z).