Vlad BOCĂNEȚ

Statistică și probabilități

Suport de curs

Editura UTPRESS
Cluj Napoca, 2023
ISBN 978-606-737-673-9

Vlad BOCĂNEȚ

Statistică și probabilități

Suport de curs

UTPRESS
Cluj-Napoca, 2023
ISBN 978-606-737-673-9

Editura UTPRESS
Str. Observatorului nr. 34
400775 Cluj-Napoca
Tel.: 0264-401.999
e-mail: utpress@biblio.utcluj.ro
www.utcluj.ro/editura

Director: ing. Dan COLȚEA

Recenzia: Prof.dr. Ioan Cristian Chifu
Prof.dr.ing. Marius Bulgaru

Pregătire format electronic on-line: Gabriela Groza

ISBN 978-606-737-673-9

Cuprins

Introducere

Acest suport de curs este destinat studenților din anul II de la specializările Inginerie Industrială și Inginerie Economică Industrială din cadrul Facultății de Inginerie Industrială Robotică și Managementul Producției a Universității Tehnice din Cluj-Napoca. Acesta are ca scop îndrumarea studenților în asimilarea informațiilor de bază necesare educației lor inginerești.

Acest material este structurat după cursul de Statistică și Probabilități predat și este împărțit pe capitole. La finalul fiecărui capitol există o secțiune de verificare a cunoștințelor și răspunsurile corecte pentru fiecare întrebare. În primul capitol se face o introducere în noțiunile de bază legate de tipuri de date, modul în care acestea se pot folosi în analize statistice, Tendința centrală a datelor și caracterizarea împrăștierii acestora. În al doilea capitol se abordează subiectul vizualizării datelor și extragerii de informații din acestea. Se abordează atât vizualizări des întâlnite în statistică și lucrul cu date cât și alte metode precum text și tabele. Apoi, în cel de-al treilea capitol studenții fac cunoștință cu evenimentele statistice care pun bazele probabilităților. Se face o scurtă recapitulare a ce înseamnă evenimente și care sunt operațiile cu evenimente. În cel de-al patrulea capitol sunt prezentate câteva noțiuni și legi de bază în lucrul cu probabilitățile. Al cincilea capitol introduce noțiunea de variabilă și abordează variabilele aleatorii discrete și continue. În acest capitol se prezintă câteva tipuri cunoscute de variabile și câteva exemple de folosire a acestora. Apoi se introduce noțiunea de distribuție de probabilitate. Apoi studenții vor face cunoștință cu câteva dintre cele mai comune distribuții discrete: uniformă, binomială și hipergeometrică în capitolul șase și cu distribuții continue comune ca distribuția normală, Student și Chi-pătrat în capitolul al șaptelea. Aceste noțiuni despre distribuții sunt apoi folosite în rezolvarea unei probleme des întâlnite în inginerie, cea de estimare a parametrilor populației (media și dispersia) tratate în capitolul al optulea. Studenții vor învăța să folosească tabele de distribuție și să calculeze intervale de încredere cu scopul de a estima anumiți parametrii ai unei populații (medie sau dispersie). Următorul capitol prezintă aplicarea practică a statisticii în inginerie și mai concret în Controlul Statistic al Proceselor (SPC – Statistical Process Control). Unele unelte deja prezentate în capitolele anterioare sunt puse într-un cadru practic iar altele sunt nou introduse. La final, în capitolul 10 sunt tratate noțiunile de corelație și regresie liniară care sunt noțiuni foarte importante ale statisticii inferențiale și predictive.

1. Statistică descriptivă

Statistica descriptivă este o ramură a statisticii care implică colectarea, organizarea, prezentarea și analiza de date. Scopul statisticii descriptive este de a descrie caracteristicile unui set de date prin măsurători, cum ar fi măsurile de tendință centrală și măsurile de împrăștiere. Aceste măsurători sunt utilizate pentru a oferi o înțelegere mai clară a datelor și pentru a evidenția modelele și tendințele. Statistica descriptivă este utilizată în multe domenii, inclusiv în cercetarea de piață, în științele sociale, în afaceri și în științele naturale.

Dicționarul Cambridge [1] definește datele ca fiind:

„Informație, în special descrieri concrete sau numere, colectate pentru a fi examinate, analizate și utilizate pentru a ajuta la luarea deciziilor, sau informații într-o formă electronică care pot fi stocate și utilizate de către un computer.”

Datele, informațiile și cunoașterea pot fi definite în foarte multe feluri [2] și sunt considerate a fi diferite. Există și o ierarhie a datelor, informației și cunoașterii (Figura 1.1) în care datele sunt baza informației iar cunoașterea derivă din informație. Potrivit [3], [4] datele sunt simboluri, informațiile sunt date procesate pentru a fi utile și răspund la întrebări de genul „cine”, „ce’ și „când” iar cunoașterea este aplicarea datelor și informațiilor pentru a răspunde la întrebări „cum”.

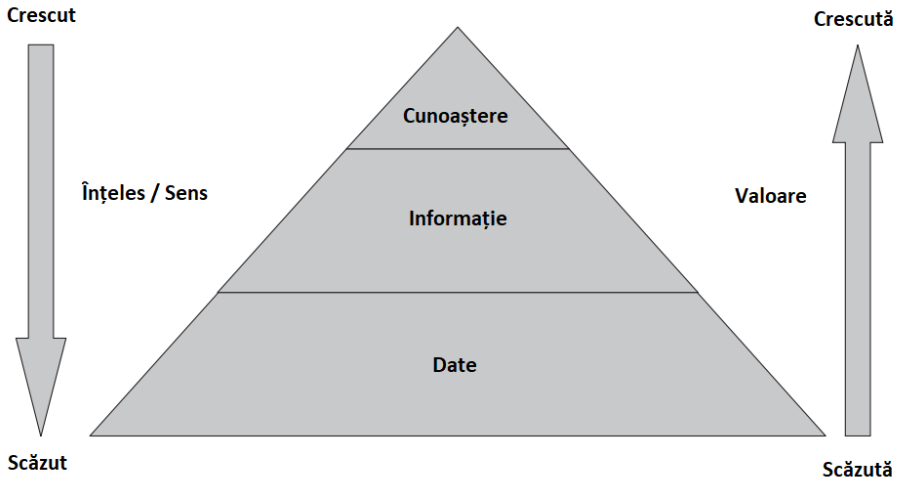


Fig. 1.1. Ierarhia Cunoaștere-Informație-Date [4], [5]

Sursele de date sunt peste tot în jurul nostru iar numărul lor crește pe măsură ce senzorii devin mai ieftini și utilizați pe scară largă. Chiar și telefoanele noastre inteligente sunt o agregare de senzori care colectează și analizează continuu date achiziționate. Companiile devin din ce în ce mai mult orientate spre culegerea și analiza datelor, strângând date atât de la clienți cât și din propriile procese de producție, cu scopul de a

colecta informații în baza cărora să poată lua decizii mai bune și de a crește performanța afacerilor acestora.

Există multe tipuri de date ce pot fi colectate. Unele sunt structurate în tabele sau alte structuri de date, dar majoritatea sunt nestructurate. Folosind noțiuni de statistică, putem „săpa" în aceste date și să ajungem la informațiile ce ne pot ajuta în procesul decizional.

Datele pot fi **numerice** sau **nenumerice**. Datele numerice sunt caracterizate prin numere și furnizează de obicei informații cantitative. Spre exemplu, masa unui obiect poate fi exprimată în kilograme, aceasta fiind o cantitate numerică. Datele nenumerice sunt de obicei calitative și arată anumite atribute ce nu pot fi exprimate numeric. Culoarea sau forma unui obiect sunt exemple de date nenumerice. Datele numerice au o încărcătură informațională mai mare și pot fi transformate (la nevoie) în date nenumerice. De exemplu, un dulap are 30 kg (informație numerică) sau zicem doar că este greu (informație calitativă). Exprimând masa dulapului în kilograme știm exact cât de greu este dulapul, pe când spunând doar că este „greu" nu ne dă informații exacte despre masa acestuia, ci doar să ne așteptăm la o masă mai mare.

În funcție de modul în care putem să caracterizăm o variabilă (ceva care își poate schimba valoarea [6]) aceasta poate fi măsurată pe diferite nivele de măsură: nominal, ordinal, interval sau rație. Fiecare nivel de măsură adaugă informații suplimentare. Nivelul de măsură este important deoarece ne indică ce unelte de analiză putem folosi pentru acea variabilă.

Dacă măsurătorile unei variabile pot fi grupate doar în categorii atunci vorbim despre nivelul **nominal** de măsurare. Un element poate să fie într-o categorie sau alta, iar aceasta este caracteristica sa. Spre exemplu, avem o cutie cu creioane colorate iar variabila pe care o urmărim este culoarea acestora. Pot exista mai multe categorii de culori: negru, albastru, roșu sau verde, iar fiecare creion are o culoare aparținând uneia dintre aceste categorii. Un creion este considerat într-o singură categorie deoarece nu poate avea în același timp două culori (e.g. albastru și verde). Când lucrăm cu variabile nominale, putem să numărăm câte elemente observate avem în fiecare categorie. Numărul de elemente se mai numește și frecvență, despre care vom discuta mai târziu în curs.

Dacă categoriile pot fi aranjate într-o ordine intuitivă, atunci spunem că datele sunt măsurate la nivel **ordinal** . Acest lucru ne dă o informație în plus despre datele noastre, și anume ordinea. Un exemplu ar fi înălțimea pomilor de Crăciun clasificându-i în una din categoriile: *scunzi*, *mijlocii* sau *înalți*. Pe lângă faptul că știm că un pom aparține unei anumite categorii, știm și poziția categoriei în ierarhia acestora (*scund* înainte de mijlociu iar *mijlociu* înainte de *înalt*).

Datele măsurate la nivelurile nominal și ordinal sunt de obicei nenumerice. Ele exprimă caracteristici **calitative** sau atributive. Cele ordinale pot primi notații numerice (spre exemplu în loc de mic, mediu, mare să notăm cu 1, 2, 3) dar trebuie să avem grijă deoarece acestea sunt doar simboluri care ne ajută să observăm ordinea și poziția în ierarhie. Nu putem face calcule cu aceste numere (adunare, scădere etc.). Pe lângă determinarea frecvenței ca pentru variabilele nominale, pentru cele ordinale putem să le atribuim ranguri și putem face o serii de analize statistice cu ranguri.

Datele numerice sunt evaluate folosind scale iar de obicei acestea au intervale egale. Dacă scala pe care o folosim la evaluarea unei variabile nu are un "zero" semnificativ, atunci vorbim despre nivelul **interval** de măsură. Zero semnificativ înseamnă că valoarea „0” este doar o altă valoare pe scala noastră, fără a simboliza lipsa unei cantități. Un exemplu des întâlnit este temperatura. Puteți măsura temperatura pe o scală cu un interval de 1 grad (Celsius de exemplu). Chiar dacă observăm valoarea "0" pe scală, aceasta nu înseamnă că, dacă atingem valoarea 0 nu avem o temperatură, ci doar că am ajuns la o altă valoare pe scală. Deoarece vorbim de variabile numerice, putem face operații aritmetice cu ele. Cantitățile măsurate pe această scală pot fi doar adunate sau scăzute, dar nu pot fi împărțite sau înmulțite. De aceea, expresia "*astăzi este de două ori mai rece decât ieri*" nu are nici un sens. Întrebați-vă, dacă astăzi sunt 0°C iar mâine este de două ori mai cald, ce temperatură vom avea mâine? În schimb putem spune că "*astăzi este cu 5 grade mai rece decât ieri*".

Dacă scala are în schimb un „zero” (o origine) semnificând absența cantității respective, vorbim despre o variabilă măsurată la nivelul de măsurare **rație**. Cantitățile măsurate cu instrumente sunt de obicei la acest nivel de măsură. Exemple de cantități: masă, greutate, lungime, voltaj, forță etc. Acesta este cel mai complet nivel la care o cantitate poate fi măsurată. Puteți face toate tipurile de operații (adunare, scădere, înmulțire, împărțire) iar zero înseamnă lipsa cantității respective. De exemplu, dacă cântăriți un sac de cartofi, puteți să-l împărțiți în jumătate sau puteți afla dublu său, iar când aveți 0 kg de cartofi, înseamnă că nu aveți cartofi. Datele numerice se mai numesc și **cantitative** deoarece exprimă cantități.

În figura 1.2 se poate vedea o imagine de ansamblu asupra diferitelor tipuri de date și a nivelelor de măsură ale acestora.

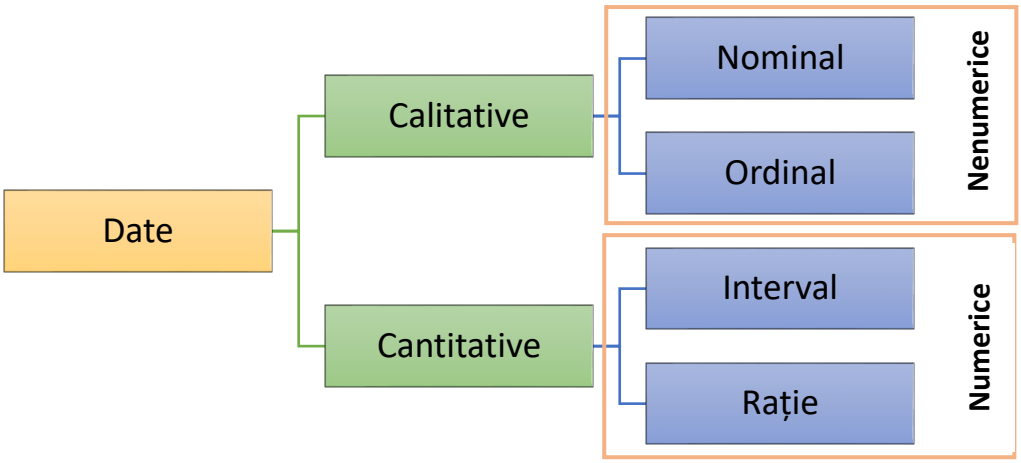


Fig. 1.2. Tipuri de date și nivelele lor de măsură

Există mai multe măsuri care pot fi utilizate pentru a descrie datele. Potrivit [7] există patru categorii:

- Măsuri de frecvență (numărul de apariții)
- Măsuri de tendință centrală (valori centrale în jurul cărora se acumulează datele)
- Măsuri de răspândire (variația datelor)
- Măsuri de poziție (poziția unei valori în cadrul setului de date căruia îi aparține)

1.1. Frecvența

Folosim termenul de **frecvență** când facem referire la cât de des apare un eveniment sau la numărul de elemente dintr-o categorie sau interval. Putem exprima frecvența prin numărul absolut de valori pe care îl observăm și spunem că determinăm **frecvența absolută**, sau să o exprimăm în procente raportându-ne la totalul observațiilor aceasta fiind **frecvența relativă**. Este o noțiune importantă care este des utilizată în statistică și analiză a datelor.

Frecvența absolută este numărul absolut de elemente dintr-o categorie sau interval sau, dacă vorbim de evenimente, este numărul de apariții ale unui eveniment. De exemplu, dacă avem 8 mere din care 3 sunt roșii și 5 sunt galbene, atunci 3 și 5 ar fi frecvențele absolute pentru cele două categorii (roșu și galben).

Frecvența relativă este considerată în raport cu numărul total de elemente (sau apariții):

$$f_i = \frac{a_i}{n}$$

unde:

f_i – frecvența relativă a intervalului sau categoriei i

a_i - frecvența absolută a intervalului sau categoriei i

n – numărul total de elemente luate în considerare (din toate categoriile cumulate)

Frecvența relativă este exprimată în fracții sau sub formă zecimală, dar cel mai frecvent o găsim exprimată în procente. Continuând exemplul de mai sus, cele 3 mere roșii ar reprezenta 3/8 sau 37.5% din numărul total de mere. Mere galbene compun restul 5/8 (sau 62.5%) de mere.

Un alt mod de a utiliza frecvențele (absolute sau relative) este de a le cumula. Aceasta se numește **frecvență cumulată**. Se poate face calculul începând cu prima categorie (crescător) sau cu ultima (descrescător). Când determinăm frecvența cumulată crescător, adunăm frecvențele din fiecare categorie începând de la prima până la categoria de interes. Frecvența cumulată crescător răspunde la întrebarea: „*Câte valori avem în primele i intervale/categorii?*”. În cazul frecvenței cumulate descrescător procedăm în mod similar, dar începem de la ultima categorie/interval în loc de prima.

Frecvențele sunt de obicei prezentate în tabelele de frecvență. Puteți vedea un exemplu în Tabelul 1.1, extins de la exemplul de mere de mai sus.

Tabel 1.1 – Tabel de frecvențe (absolută, relativă, cumulată crescător și descrescător)

Categorii	Frecvență absolută	Frecvența relativă	Frecvență absolută cumulată crescător	Frecvență absolută cumulată descrescător
Mere roșii	11	22%	11	50
Mere galbene	19	38%	30	39
Mere verzi	8	16%	38	20
Mere roșii-galbene	12	24%	50	12
Nr. total. de mere	50	100%		

1.2. Tendința centrală

Tendința centrală denotă „tendința datelor cantitative de a se grupa în jurul unei valori centrale” [8]. Această tendință a dat naștere *teoriei tendinței centrale*. Valoarea centrală în jurul căreia se grupează datele se numește o măsură a tendinței centrale. Cele mai frecvente sunt:

- **media aritmetică** (media)
- **mediană**
- **modala**

Alte măsuri utile sunt:

- **media pătratică**,
- **media geometrică**
- **medie armonică**

Mai departe le vom analiza pe fiecare în mai mare detaliu.

Media aritmetică este una dintre cele mai des folosite măsuri de tendință centrală. Pentru a calcula aceasta se face suma tuturor valorilor și se împarte la numărul de valori. Acesta are mai multe notații precum μ pentru media populației și $\bar{x}$ pentru media eșantionului, sau mai general, cu M_x.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

unde x_i sunt valorile pentru care calculăm media aritmetică, iar n este numărul de valori.

Mediana este valoarea care ocupă locul central într-un set ordonat de valori și împarte șirul în două șiruri de lungimi egale:

$$M_e = x_{(n+1)/2}$$

unde n este numărul de valori din șir.

Fiind dat următorul set de valori:

1, 1, 2, 4, 5, 7, 9

atunci 4 este valoarea mediană (avem trei valori înainte de aceasta și trei valori după ea). Dacă șirul are un număr par de valori, atunci mediana este media aritmetică a valorilor care ocupă pozițiile centrale a șirului ordonat:

$$M_e = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Față de media aritmetică, mediana are avantajul că nu este la fel de sensibilă la valori aberante (valori foarte diferite de restul șirului). Spre exemplu, pentru șirul:

1, 2, 3, 4, 500

Media aritmetică este egală cu 102 în timp ce mediana este 3, o valoare mult mai reprezentativă ca valoare centrală decât media aritmetică.

Modala este valoarea (sau categoria) cu frecvența cea mai mare. Cel mai simplu mod de a determina modala este grafic. Dacă am reprezenta grafic frecvențele de mere din exemplul anterior, vom obține graficul din Figura 1.3. Categoria cu cea mai mare frecvență, conține modala. În acest caz, modala se află în categoria de mere galbene.

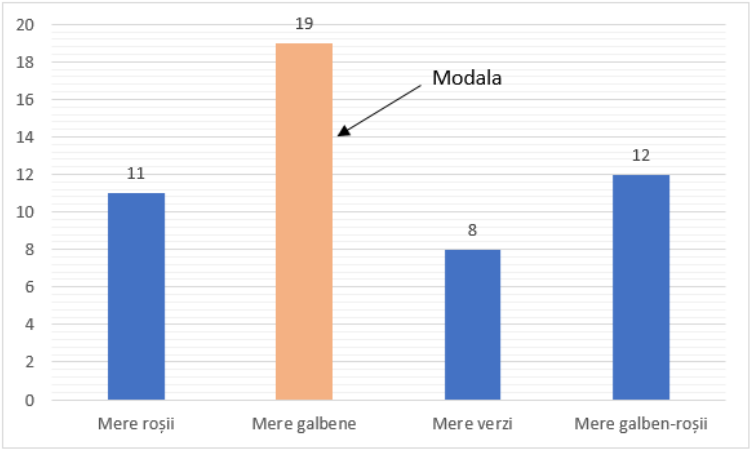


Fig. 1.3. Determinarea modalei prin metoda grafică

În unele cazuri, alte măsuri ar fi mai adecvate pentru a transmite tendința centrală. Cele mai frecvente sunt mediile pătratică, geometrică și armonică.

Media pătratică este rădăcina pătrată a mediei valorilor ridicate la pătrat. Aceasta înseamnă că trebuie mai întâi să ridicăm la pătrat fiecare valoare, să determinăm media și să extragem rădăcina pătrată. Formula arată astfel:

$$M_p = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{n}}$$

unde x_i sunt valorile pentru care este calculată media, iar n este numărul de valori.

Media geometrică poate fi calculată prin înmulțirea tuturor valorilor și extragerea radicalilor de grad n:

$$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

unde x_i sunt valorile pentru care este calculată media, iar n este numărul de valori.

Media geometrică este utilizată, de exemplu, în finanțe pentru a determina rata medie de creștere (sau rata de creștere anuală compusă). Dacă doriți să calculați rata de creștere medie a unei investiții pe parcursul a mai mulți ani, puteți utiliza media geometrică a ratei de creștere anuale pentru a obține o valoare centrală a ratei de creștere. Aceasta se folosește în locul mediei aritmetice care în acest caz nu ar fi folosită corect. Ea mai poate fi utilizată pentru a calcula rata de rentabilitate a unei investiții pe o perioadă mai lungă de timp. De exemplu, dacă doriți să evaluați performanța unui fond de investiții pe parcursul a mai multor ani, puteți utiliza media geometrică a ratelor de rentabilitate anuale pentru a obține o imagine mai precisă a performanței sale pe termen lung. Un alt exemplu este calculul ratei medie de creștere a populației. Cu ajutorul mediei geometrice se poate evalua tendințele demografice pe termen lung sau

în previziunea creșterii populației viitoare. În general, media geometrică este utilă atunci când se dorește o măsură centrală a unui set de date care conține numere pozitive și când contează dacă valori mai mari au un impact mai mare decât cele mai mici asupra mediei.

Media armonică este calculată împărțind numărul de valori cu suma inverselor valorilor:

$$M_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

unde x_i sunt valorile pentru care este calculată media, iar n este numărul de valori.

Media armonică este utilă atunci când se dorește o măsură centrală a unui set de date care conține numere pozitive și când contează dacă valorile mai mici au un impact mai mare asupra mediei. Media armonică poate fi utilizată pentru a calcula viteza medie a unui obiect în mișcare. De exemplu, dacă un obiect se deplasează cu o viteză de 40 km/h timp de o oră, apoi se oprește și se deplasează cu o viteză de 20 km/h timp de o oră, viteza medie a obiectului în cele două ore este media armonică a celor două viteze. Ea mai poate fi utilizată pentru a calcula rata medie de rentabilitate a unui portofoliu de investiții pe parcursul unui an. De exemplu, dacă un portofoliu de investiții a obținut o rentabilitate de 10% în primul trimestru, 5% în al doilea trimestru și 15% în al treilea trimestru, rata medie de rentabilitate a portofoliului pentru întregul an este media armonică a ratelor de rentabilitate trimestriale. Un alt exemplu este pentru calculul mediei notelor la teste sau examene. Se poate evalua performanța generală a unui elev, luând în considerare faptul că notele mai mici pot avea un impact mai mare asupra mediei decât cele mai mari.

1.3. Variaţia sau împrăştierea datelor

Datele pe care le colectăm au o variație naturală, ceea ce înseamnă că acestea sunt diferite unele de altele. Măsurile de tendință centrală nu pot să surprindă această variabilitate a datelor. De exemplu, putem avea un grup de studenți care au aceeași medie pentru două materii diferite (Tabelul 1.2). Cu toate acestea, valorile pentru cele două materii nu sunt răspândite în același mod, ceea ce poate fi observat. Prin urmare, este necesar să utilizăm măsuri de împrăștiere sau variație pentru a obține o înțelegere mai precisă a datelor.

Tabelul 1.2. Un grup de studenți având aceeași medie la 2 materii diferite

	Materia 1	**Materia 2**
Student 1	5	6
Student 2	9	7
Student 3	4	7
Student 4	8	8
Student 5	9	7
Medie	**7**	**7**

Cele mai frecvent folosite măsuri de împrăștiere sunt:

- Amplitudinea
- Intervalul intercuartilic
- Dispersia
- Abaterea standard

Amplitudinea este diferența dintre valoarea maximă și cea minimă:

$$R = x_{max} - x_{min}$$

Avantajul principal al amplitudinii este că este ușor de calculat. Se poate folosi spre exemplu, când se dorește a se determina diferența dintre cea mai mică și cea mai mare valoare. De exemplu, când analizați variabilitatea temperaturilor maxime și minime într-o zonă. Un alt exemplu ar fi pentru a compara variabilitatea a două seturi de date, cum ar fi variația salariilor în două companii diferite. Amplitudinea se mai poate folosi când doriți să identificați posibile puncte extreme într-un set de date. De exemplu, când analizați distribuția vârstelor angajaților unei companii și doriți să identificați dacă există angajați foarte tineri sau foarte în vârstă.

Dezavantajul este că ia în considerare doar două valori și acest lucru o face sensibilă la valori extreme (de asemenea, cunoscute și ca *valori aberante*).

Intervalul intercuartilic (IQR) este o măsură de împrăștiere care măsoară diferența dintre a treia cuartilă (75%) și prima cuartilă (25%) ale unui set de date, adică diferența dintre valoarea medianei superioare și valoarea medianei inferioare. Cele trei cuartile sunt valorile care împart un set de date în patru părți egale (sferturi). Acestea sunt:

- Prima cuartilă (Q1): este valoarea care împarte setul de date în două părți, astfel încât 25% din date sunt mai mici decât această valoare și 75% sunt mai mari decât aceasta.

- Mediană sau a doua cuartilă (Q2): este valoarea care împarte setul de date în două părți egale, astfel încât 50% din date sunt mai mici decât această valoare și 50% sunt mai mari decât aceasta.
- A treia cuartilă (Q3): este valoarea care împarte setul de date în două părți, astfel încât 75% din date sunt mai mici decât această valoare și 25% sunt mai mari decât aceasta.

Formula de calcul a intervalului intercuartilic este:

$$IQR = Q_3 - Q_1$$

Intervalul intercuartilic se poate folosi când dorim să evaluăm variabilitatea unui set de date și să eliminăm influența punctelor extreme. În unele aplicații este de preferat a se folosi IQR în locul amplitudinii, deoarece elimină variația din punctele extreme. Pe lângă aplicații în care se dorește compararea variabilității a două seturi de date, se mai poate folosi la identificarea și eliminarea punctelor extreme dintr-un set de date. De exemplu, dacă analizați distribuția vârstelor angajaților unei companii și doriți să identificați angajații care sunt mult mai în vârstă decât media sau mult mai tineri decât media.

În general, intervalul intercvartilic este o măsură mai robustă de împrăștiere (variație) decât amplitudinea și poate fi util în situații în care există puncte extreme sau diferențe semnificative între mediile a două seturi de date.

Dispersia este o altă măsură a împrăștierii datelor. Spre deosebire de amplitudine și de intervalul intercuartilic, aceasta folosește în formula de calcul toate valorile din setul de date analizat.

Pentru a calcula dispersia, trebuie mai întâi să determinăm diferența dintre fiecare punct față de medie. Apoi, ridicăm la pătrat aceste diferențe și facem media lor. Formula pentru dispersia populației, notată cu σ^2, arată astfel:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

unde x_i sunt valorile pentru care este calculată media, μ este media populației și n este numărul de valori.

În cazul dispersiei eșantionului, notată cu s^2, formula este corectată cu un factor de corecție care ține cont de faptul că dispersia este calculată doar pe un subset al populației:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

unde x_i sunt valorile pentru care este calculată media, $\bar{x}$ este media eșantionului și n este numărul de valori.

Dispersia poate fi folosită în diverse aplicații precum în economie, unde poate fi utilizată pentru a evalua volatilitatea prețurilor pe piețele financiare. Dacă o acțiune are o dispersie mare, acest lucru sugerează că prețul se schimbă adesea, ceea ce poate fi riscant pentru investitori. Ea poate fi utilizată și în medicină pentru a evalua variația parametrilor fiziologici sau de laborator, cum ar fi tensiunea arterială sau nivelurile de glucoză din sânge. Această informație poate fi utilă pentru a evalua eficacitatea unui tratament sau pentru a identifica factorii de risc pentru anumite afecțiuni. În inginerie dispersia poate fi folosită pentru a evalua calitatea producției într-un proces de fabricație. O dispersie mare poate indica probleme cu procesul de producție, cum ar fi materialele defectuoase sau o linie de producție inadecvată. În psihologie ea poate fi folosită pentru a evalua diferențele dintre indivizi în ceea ce privește trăsăturile sau comportamentele lor, cum ar fi pentru a evalua gradul de diversitate în răspunsurile la un chestionar de personalitate. Un alt exemplu poate fi pentru a evalua performanța studenților într-o clasă sau școală. O dispersie mare poate indica diferențe semnificative în performanța studenților sau probleme în procesul de predare.

Un dezavantaj al dispersiei este că unitatea de măsură în care este exprimată este pătratul unității de măsură a datelor pentru care se calculează dispersia. Putem rezolva ușor această problemă scoțând rădăcina pătrată a dispersie. Acest nou indicator se numește abaterea standard.

Abaterea standard este pur și simplu rădăcina pătrată a dispersiei. Aceasta înseamnă că avem următoarele două formule:

Pentru abaterea standard a populației:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Și pentru abaterea standard a eșantionului:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

unde x_i sunt valorile pentru care este calculată media, μ este media populației, $\bar{x}$ este media eșantionului și n este numărul de valori.

2. Modalități de prezentare și sinteză a datelor

Există mai multe modalități de prezentare a datelor cum ar fi folosind text, tabele și metode vizuale. Textul este cea mai simplă metodă și poate fi folosit pentru a prezenta datele sub formă de propoziții sau fraze. Folosind indicatori statistici putem capta și prezenta sintetic anumite caracteristici ale datelor, cum ar fi tendința centrală sau împrăștierea datelor. Tabelele implică organizarea datelor într-o structură tabelară sau într-o matrice, iar acestea pot fi utilizate pentru a evidenția diferențele și asemănările dintre diferitele seturi de date. Metodele vizuale includ graficele și diagramele, care pot fi utilizate pentru a oferi o reprezentare vizuală a datelor. Acestea pot fi utile pentru a evidenția tendințele, modelele și distribuțiile datelor și pot oferi o imagine mai clară a datelor. Fiecare metodă are propriile avantaje și dezavantaje, iar alegerea unei metode depinde de tipul de date și de obiectivul prezentării lor.

2.1. Text și indicatori statistici

Textul și indicatorii statistici sunt două modalități importante prin care informațiile pot fi comunicate în mod eficient. Textul poate fi utilizat pentru a explica contextul datelor și a oferi o perspectivă generală asupra problemei abordate. Indicatorii statistici, pot oferi o imagine clară asupra distribuției datelor și a variației lor. Prin utilizarea acestor două abordări împreună, putem furniza cititorului o imagine completă a datelor și a concluziilor importante care pot fi trase din ele.

Indicatorii statistici sunt valori care ne oferă informații valoroase despre datele pe care le analizăm. Putem folosi măsurile de tendință centrală și împrăștiere pentru a afla informații de ansamblu despre datele noastre. În tabelul 2.1 sunt prezentați sumar indicatorii statistici reprezentativi în determinarea tendinței centrale, împrăștierii datelor și caracterizării formei distribuțiilor datelor.

Tabel 2.1. Indicatori statistici reprezentativi

Măsura	**Indicator**	**Formula**
Tendința centrală	Media aritmetică	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$
	Mediana	Valoarea din centrul șirului
	Modala	Frecvența maximă
	Media pătratică	$M_p = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{n}}$

	Media geometrică	$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$
	Medie armonică	$M_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
	Valoare centrală	$x_c = \frac{Max - Min}{2}$
Împrăștiere	Min	Cea mai mică valoare
	Max	Cea mai mare valoare
	Gama	$R = x_{max} - x_{min}$
	Dispersia	$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$
	Abatere STD	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$
Forma distribuției	Asimetria	$g_1 = \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3}$
	Aplatizarea	$k = \frac{\sum_{i=1}^{n}(x_i - \mu)^4}{(\sum_{i=1}^{n}(x_i - \mu)^2)^2}$

Există totuși limitări la cantitatea pe care o putem transmite eficient prin intermediul textului. Când depășim un anumit nivel de complexitate, prezentarea datelor sub formă de text face anevoioasă înțelegerea și transmiterea mesajului dorit. De aceea, pentru a transmite mai multe informații eficient folosim tabele și vizualizări.

2.2. Tabele

Tabelele sunt un mijloc important și frecvent utilizat pentru a prezenta datele în mod organizat și ușor de citit. Acestea pot fi utilizate pentru a evidenția diferențele și asemănările dintre diferitele seturi de date, precum și pentru a oferi o imagine de ansamblu a datelor. Ele pot fi utile în identificarea tendințelor și modelelor prezente în date și pot fi o resursă valoroasă pentru analiza și interpretarea acestora.

Crearea unui tabel începe cu selectarea setului de date relevant pentru analiză. Datele ar trebui să fie organizate într-un mod coerent, astfel încât informațiile să poată fi ușor de citit și interpretat. Informațiile sunt organizate în coloane și rânduri, iar numele

coloanelor trebuie alese încât să fie semnificative pentru a putea identifica cu ușurință datele din fiecare coloană. De obicei, variabilele (caracteristicile) pe care le măsurăm se află pe coloane, iar observațiile (măsurătorile) sunt pe rânduri. În exemplul din tabelul 2.2., jucăria 1 este rotundă, are o lungime de 23 mm și culoarea roșie. Acestea sunt toate caracteristicile unui singur obiect. Pe de altă parte, observăm că toate jucăriile au doar două forme posibile: rotunde și pătrate.

Ne putem uita fie la o caracteristică a tuturor elementelor din set (într-o coloană) sau la toate caracteristicile unui singur element (într-un rând) în funcție de informațiile pe care vrem să le folosim.

Tabel 2.2 Exemplu de tabel

	Forma	**Lungime (mm)**	**Culoarea**
Jucăria 1	Rotund	23	Roșu
Jucăria 2	Rotund	27	Verde
Jucăria 3	Pătrat	19	Albastru

Formatul tabelelor este important, deoarece poate influența modul în care datele sunt percepute și interpretate. Se recomandă utilizarea culorilor și a formatărilor pentru a atrage atenția asupra informațiilor importante. De asemenea, ar trebui evitată supraaglomerarea tabelelor cu prea multe informații. Informațiile ar trebui să fie prezentate într-un mod clar și concis, astfel încât utilizatorii să poată identifica rapid informațiile relevante.

Interpretarea unui tabel implică examinarea datelor și identificarea tendințelor și șabloanelor. De exemplu, o persoană poate examina un tabel pentru a identifica variația datelor în timp sau diferențele între două grupuri de date. De asemenea, tabelele pot fi utilizate pentru a identifica frecvențele sau proporțiile diferitelor valori dintr-un set de date.

Când avem puține informații (cum ar fi exemplul de mai sus) acestea pot fi prezentate ușor într-un tabel. Dar când avem tabele mari cu zeci sau sute de variabile și mii de rânduri, poate fi aproape imposibil să obțineți informații doar uitându-ne la tabel. De aceea este important să se ia în considerare avantajele și dezavantajele utilizării tabelelor în funcție de nevoile specifice ale utilizatorului și de scopul analizei. Uneori, alte forme de prezentare a datelor, cum ar fi graficele sau diagramele, pot fi mai eficiente decât tabelele în prezentarea datelor.

2.3. Vizualizarea datelor

Vizualizarea datelor este o metodă puternică de comunicare a informațiilor, care poate fi utilizată pentru a sublinia tendințele și șabloanele importante. Graficele și diagramele pot fi create pentru a ilustra relațiile și distribuțiile din datele colectate, precum și pentru a evidenția variația și diferențele dintre grupurile de date. Acestea pot fi prezentate într-o varietate de formate, cum ar fi diagramele cu bare, diagramele cu linii, graficele cu puncte și diagramele circulare, în funcție de tipul de date și de informațiile pe care dorim să le evidențiem. Prin utilizarea vizualizării datelor într-un mod creativ și inteligent, putem furniza informații complexe într-un mod accesibil și ușor de înțeles, ceea ce poate fi extrem de valoros în procesul de comunicare a descoperirilor și concluziilor importante. Cele mai frecvent utilizate tipuri de grafice sunt:

1. ***Diagrama cu bare***: utilizată pentru a compara cantități sau frecvențe. Aceasta constă într-un set de bare verticale sau orizontale, care reprezintă valorile variabilelor.
2. ***Diagrama cu linii***: utilizată pentru a ilustra tendințele și schimbările într-o serie de date. Folosește un set de puncte conectate prin linii, care reprezintă valorile variabilelor într-o ordine cronologică sau logică.
3. ***Graficul cu puncte***: utilizat pentru a vizualiza distribuția datelor și pentru a identifica punctele aberante pentru două variabile.
4. ***Diagrama circulară***: utilizată pentru a ilustra proporțiile unui întreg. Aceasta constă într-un cerc împărțit în secțiuni, unde fiecare secțiune reprezintă o proporție a întregului.
5. ***Histograma***: utilizată pentru a ilustra distribuția datelor continue. Aceasta constă într-un set de bare care reprezintă intervalul valorilor variabilelor.
6. ***Diagrama de tip boxplot***: folosită pentru a vizualiza și compara cu ușurință distribuții ale datelor continue.

În continuare vom intra mai în detaliu pentru fiecare tip de diagramă și vom vedea cum sunt construite și când le folosim.

2.3.1. Diagrama cu bare sau coloane

Diagrama cu bare este o metodă eficientă de a prezenta datele în mod vizual. Aceasta este formată dintr-un set de bare verticale (numite și coloane) sau orizontale care reprezintă valorile variabilelor. Diagrama cu bare este folosită pentru a compara de obicei frecvențele variabilelor sau pentru a evidenția tendințele în date.

Diagrama cu bare este utilă atunci când lucrăm cu date calitative sau cantitative discrete, unde variabilele sunt reprezentate de categorii sau de numere întregi. Este recomandat să utilizăm diagrama cu bare în situațiile în care avem cel puțin patru sau cinci categorii.

Pentru a crea o diagramă cu bare, trebuie urmați următorii pași:

1. Identificați variabila pe care doriți să o reprezentați și categoriile acesteia.
2. Decideți asupra lățimii barelor. Barele trebuie să aibă aceeași lățime.
3. Desenați o axă verticală sau orizontală și marcați valorile variabilelor pe această axă.
4. Trasați barele cu înălțimea corespunzătoare pentru fiecare categorie.

Un exemplu de utilizare a diagramei cu bare ar fi reprezentarea frecvențelor de apariție a unor culori într-un set de obiecte. În acest caz, variabila ar fi "culoarea", iar categoriile ar fi culorile posibile. Diagrama cu bare arată frecvența cu care apare fiecare culoare și poate evidenția preferințele sau tendințele în alegerea culorilor.

Există mai multe tipuri de diagrame cu bare, cum ar fi diagrama cu bare simple, diagrama cu bare multiple sau diagrama cu bare împerecheate. Diagramele cu bare simple pot fi cu coloane (Figura 2.1) sau bare orizontale (Figura 2.2). Aceste tipuri pot fi utilizate în funcție de scopul și caracteristicile datelor.

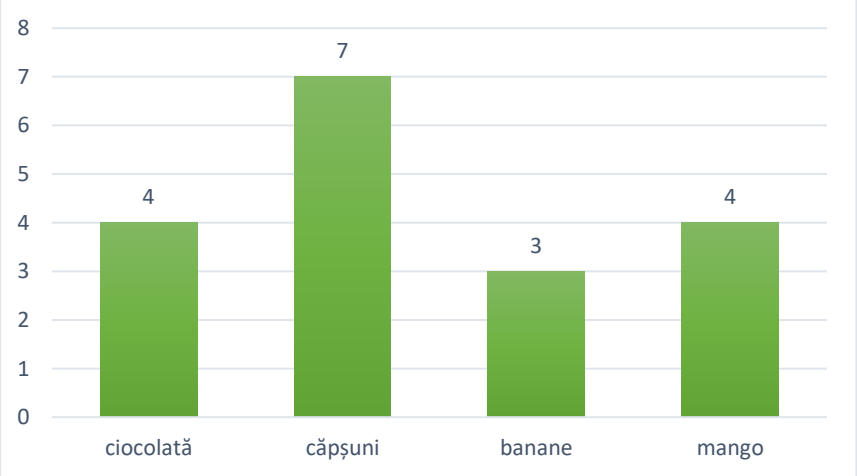


Fig. 2.1. Diagramă simplă cu coloane

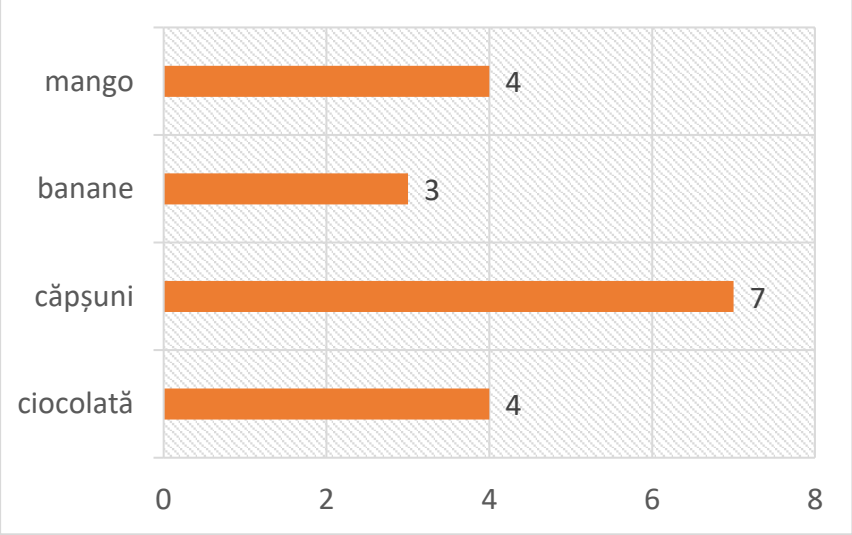


Fig. 2.2.. Diagramă simplă cu bare orizontale

Într-o diagramă putem prezenta mai multe instanțe ale aceleași variabile. Spre exemplu, să zicem că vindem înghețată în două magazine și vrem să comparăm vânzările pentru cele două magazine în funcție de aromele puse în vânzare. Putem crea o diagramă ca cea din figura 2.3 în care să comparăm vânzările pentru fiecare magazin în funcție de aroma vândută.

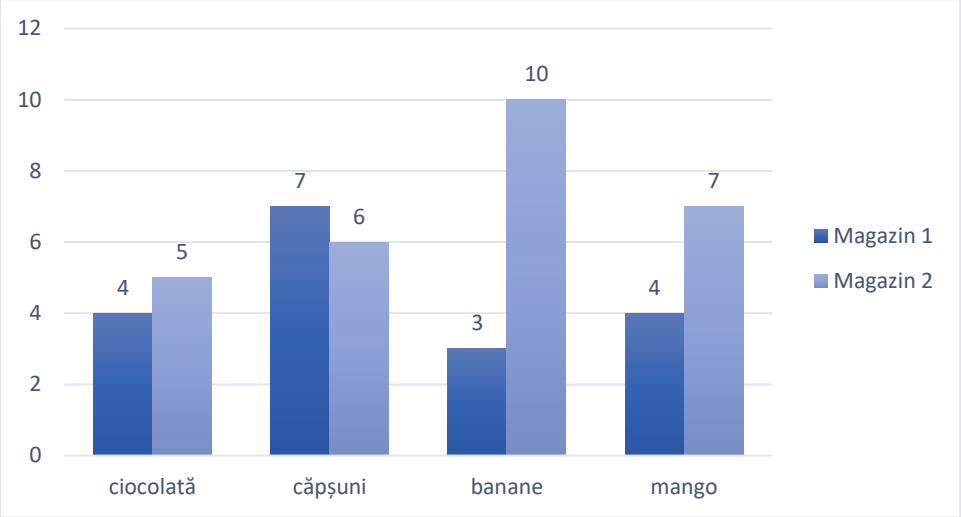


Fig. 2.3. Diagramă cu coloane cu mai multe instanțe

Aceleași informații le putem prezenta ca proporție dintr-un total sub formă de coloane stivuite (Figura 2.4). Pentru fiecare aromă vândută putem observa care este proporția vândută în magazinul 1 și respectiv în magazinul 2.

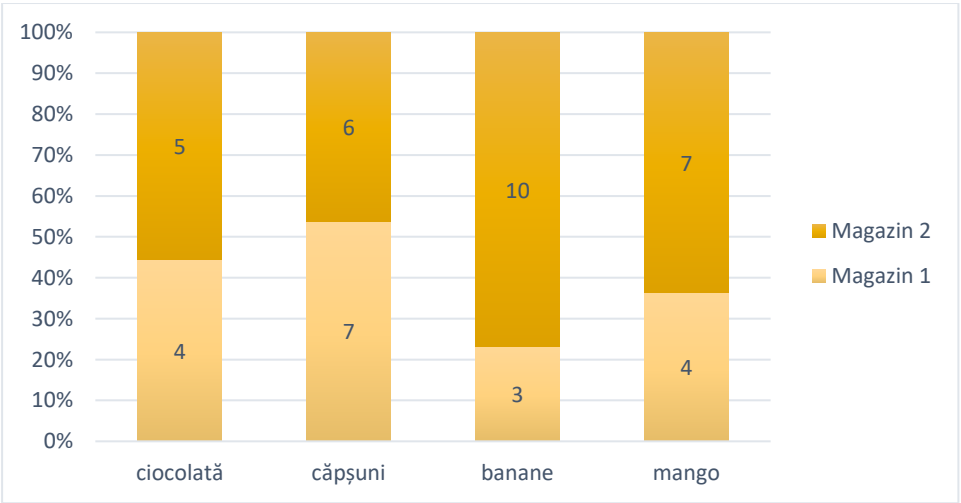


Fig. 2.4. Diagramă cu coloane stivuite

2.3.2. *Diagrama cu linii*

Diagrama cu linii este o metodă eficientă de a evidenția evoluția unei serii de date într-un anumit interval de timp.

Pentru a crea o diagramă cu linii, trebuie să avem datele pe care dorim să le prezentăm și să le organizăm într-un tabel cu două coloane: una pentru valorile timpului

(sau un index corespunzător unei cronologii) și una pentru valorile variabilei de interes. Apoi, trasăm un sistem de coordonate, unde axa orizontală reprezintă timpul și axa verticală reprezintă valorile variabilei de interes. Pe baza acestor coordonate, putem marca punctele pentru fiecare valoare și să le unim cu o linie pentru a crea diagrama.

Există multe exemple în care diagrama cu linii, cum ar fi pentru a arăta evoluția temperaturilor medii într-un anumit oraș de-a lungul anului sau pentru a urmări creșterea sau scăderea vânzărilor unei companii într-un anumit interval de timp. De asemenea, acest grafic poate fi folosit pentru a evidenția tendințele sau fluctuațiile din datele colectate într-un studiu științific. Spre exemplu, în figura 2.5 este prezentată evoluția prețului zilnic maxim în USDT al unei criptomonede (BTC) în perioada Ianuarie-Februarie 2022.

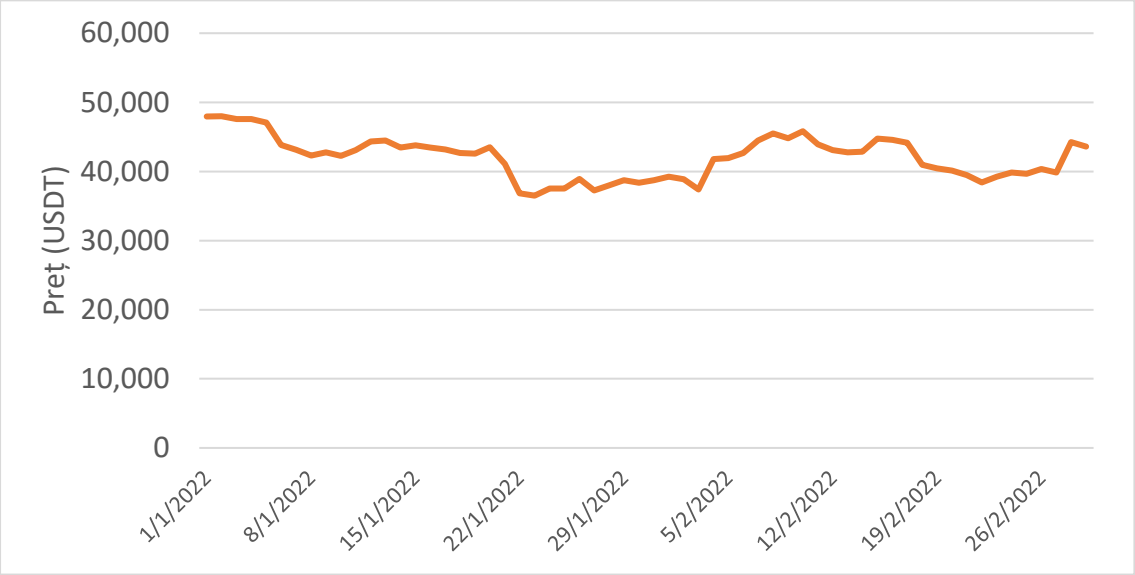


Fig. 2.5. Diagramă cu linii

În general, diagrama cu linii este o modalitate eficientă de a comunica datele complexe într-un mod accesibil și ușor de înțeles.

2.3.3. Graficul cu puncte (scatter plot)

Graficul cu puncte este o reprezentare grafică a datelor care permite vizualizarea relației dintre două variabile continue. Acest tip de grafic este utilizat în special în analiza de regresie, dar poate fi utilizat și pentru a observa distribuția și variația datelor.

Pentru a crea un grafic cu puncte, este necesar să se colecteze datele celor două variabile, să stabilim care variabilă este pe axa x și care pe axa y, și să plasăm fiecare pereche de valori pe grafic. De obicei, se utilizează un punct sau un simbol pentru fiecare pereche de valori.

Graficul cu puncte poate fi util în multe situații, cum ar fi:

- Pentru a analiza relația dintre două variabile. Dacă punctele sunt împrăștiate în mod aleatoriu pe grafic, atunci cele două variabile nu sunt corelate între ele. În cazul în care punctele formează o formă de linie, acest lucru indică o relație liniară între cele două variabile.
- Pentru a identifica valorile extreme (aberante). Punctele care sunt mult mai distanțate de celelalte pot fi observate cu ușurință pe grafic.
- Pentru a observa variația datelor. Dacă punctele sunt împrăștiate în mod uniform pe grafic, atunci distribuția datelor este omogenă. Dacă punctele sunt mai dense într-o anumită zonă, acest lucru indică o variație mai mică în acea regiune.

Spre exemplu, suntem interesați să observăm asocierea dintre înălțimea (în centimetri) și masa (în kilograme) a unui grup de persoane. După ce colectăm datele, plasăm înălțimea pe axa orizontală și masa pe axa verticală. Pentru fiecare persoană avem o pereche de valori înălțime-masă (e.g. 157 cm și 44 kg) reprezentate pe grafic sub forma unui punct (Figura 2.6). Reprezentând grafic aceste punte putem observa asocieri sau deviații. În acest caz putem observa că pe măsură ce înălțimea crește, și masa crește: de obicei persoanele mai înalte au masă mai mare. Există și excepții, cum ar fi ultima persoană (marcată pe grafic cu un pătrat mov) care deși este mai înaltă, are o masă mai mică.

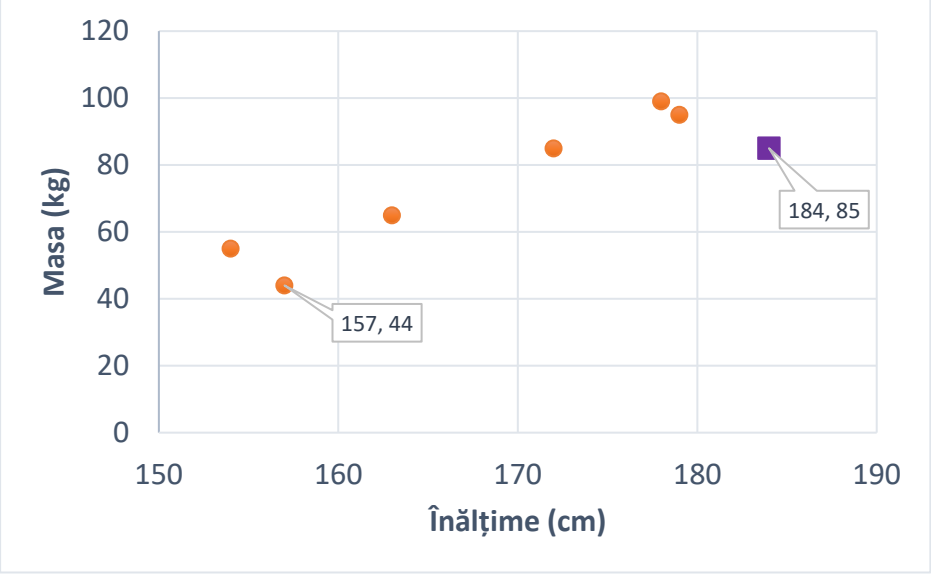


Fig. 2.6. Diagramă cu puncte

Graficul cu puncte este o metodă utilă de vizualizare a datelor și poate fi folosit în multe domenii, inclusiv în statistică, economie, biologie și altele.

2.3.4. Diagrama circulară

Diagrama circulară, cunoscută și sub denumirea de grafic cu sectoare sau grafic în proporții, este un tip de grafic utilizat în mod obișnuit pentru a arăta proporția fiecărei categorii de date într-un set de date prin reprezentarea fiecărei categorii ca sector circular.

Diagrama circulară este utilă atunci când dorim să vizualizăm proporțiile (frecvența relativă) a diferitelor categorii dintr-un set de date, mai degrabă decât numărul absolut (frecvența absolută). Acest lucru este util deoarece permite o vizualizare rapidă și ușoară a distribuției datelor și poate fi utilizat pentru a face comparații între categorii diferite.

Pentru a crea o diagramă circulară, trebuie să:

1. Calculăm proporția fiecărei categorii în setul de date.
2. Transformăm proporțiile în grade, înmulțind cu 360.
3. Desenăm un cerc și să-l împărțim în sectoare, astfel încât unghiul fiecărui sector să corespundă proporției pentru categoria respectivă.
4. Etichetați fiecare sector cu numele sau eticheta corespunzătoare categoriei.

Diagrama circulară poate fi utilizată în multe domenii diferite, inclusiv:

- În afaceri, pentru a ilustra proporția vânzărilor pe produs sau pentru a arăta cum sunt cheltuiți banii într-o companie.
- În cercetare, pentru a ilustra distribuția diferitelor răspunsuri la o întrebare sau pentru a arăta distribuția demografică a unei populații.
- În educație, pentru a ilustra proporția elevilor care învață diferite subiecte sau pentru a arăta distribuția notelor într-o clasă.

Să zicem că facem un studiu și vrem să aflăm preferințele pentru pizza a unui grup de studenți din cămin. După ce colectăm datele, le putem prezenta sub forma unei diagrame circulare, ca în figura 2.7. Din figură observăm că cea mai preferată pizza de către studenți este Marguerita cu 30% dintre răspunsuri. Deși nu vizualizăm frecvențele absolute, putem să vedem cum variază proporțiile între diferitele categorii ale unui tot.

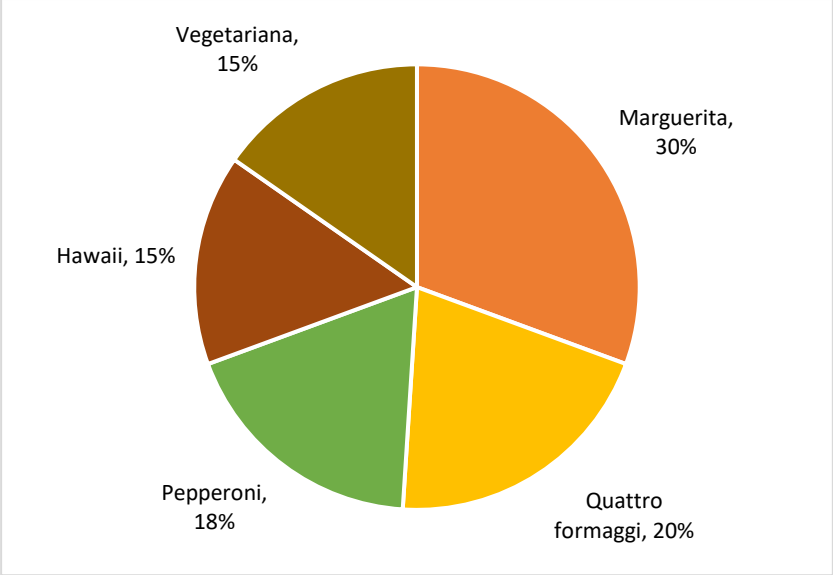


Fig. 2.7. Diagrama circulară

În general, diagrama circulară poate fi folosită pentru a ilustra proporții și distribuții într-un mod vizual și ușor de înțeles.

2.3.5. Histograma

O histogramă este o diagramă cu bare care arată frecvența cu care apar anumite valori într-un set de date continue. Cu acest tip de grafic putem vizualiza distribuția frecvenței unei variabile continue. De exemplu, putem utiliza un histogramă pentru a vedea cum sunt distribuite vârstele unei populații în intervale de vârstă.

Pentru a crea o histogramă trebuie să:

1. Determinăm numărul de intervale cel mai potrivit pentru datele pe care vrem să le vizualizăm
2. Împărțim amplitudinea la numărul de intervale
3. Calculăm capetele fiecărui interval
4. Determinăm frecvența cu care apar valorile în fiecare interval
5. Desenăm barele în funcție de frecvența determinată.

Numărul de intervale poate afecta cum arată distribuția. Există câteva metode pentru a alege numărul optim de intervale, dintre care cele mai frecvent folosite sunt:

- **Regula rădăcinii pătrate**: Numărul optim de intervale (K) este aproximativ radical din numărul de observații (n): Formula este:

$$K = \sqrt{n}$$

- **Regula lui Sturges**: Numărul optim de intervale este dat de formula:

$$K = 1 + 3.322 * log_2\,(n)$$

- **Regula lui Rice**: Numărul optim de intervale este dat de formula:

$$K = 2\sqrt[3]{n}$$

Indiferent de metoda pe care o folosim, numărul de intervale trebuie să fie cel mai apropiat număr întreg de rezultatul obținut din calcul. Nu putem avea 4.3 intervale, ci avem fie 4 sau 5 intervale.

Un exemplu de utilizare a unei histograme este prezentarea distribuției greutății unor indivizi dintr-un grup de adulți. Histograma poate arăta dacă distribuția este normală sau nu, precum și dacă există valori extreme care ar putea fi anomalii sau erori de măsurare. Despre distribuția normală vom discuta mai târziu în curs când vom aborda tipuri de distribuții.

Presupunem că avem următorul set de date pentru greutățile adulților luați în studiu:

70, 68, 73, 64, 72, 69, 76, 77, 75, 71, 68, 70, 72, 73, 74, 70, 71, 75, 73, 72

Pentru a crea histograma vom utiliza regula lui Sturges pentru numărul de intervale, vom parcurge următorii pași:

a. Determinăm numărul de observații din setul de date: n = 20
b. Calculăm numărul optim de intervale utilizând regula lui Sturges:

$$K = 1 + 3.322 * log_2\,(n) \approx\ 5.32\ \approx\ 6$$

c. Împărțim amplitudinea valorilor în 6 intervale egale:

$$A = 77 - 64 = 13$$
$$d = \frac{13}{6} = 2.166$$

d. Folosind d determinăm capetele intervalelor

Tabel 2.3. Cele 6 intervale și limitele aferente fiecărui interval

Interval	**Limite**
Interval 1	64 - 66.17
Interval 2	66.17 - 68.33
Interval 3	68.33 - 70.5
Interval 4	70.5 - 72.67
Interval 5	72.67 - 74.83
Interval 6	74.83 - 77

e. Determinăm frecvența cu care apar valorile în fiecare interval

Tabel 2.4. Tabelul de frecvență pentru cele 6 intervale

Interval	Limite	Frecvența
Interval 1	64 - 66.17	1
Interval 2	66.17 - 68.33	2
Interval 3	68.33 - 70.5	4
Interval 4	70.5 - 72.67	5
Interval 5	72.67 - 74.83	4
Interval 6	74.83 - 77	4

f. Desenăm barele în funcție de frecvența determinată (Figura 2.8)

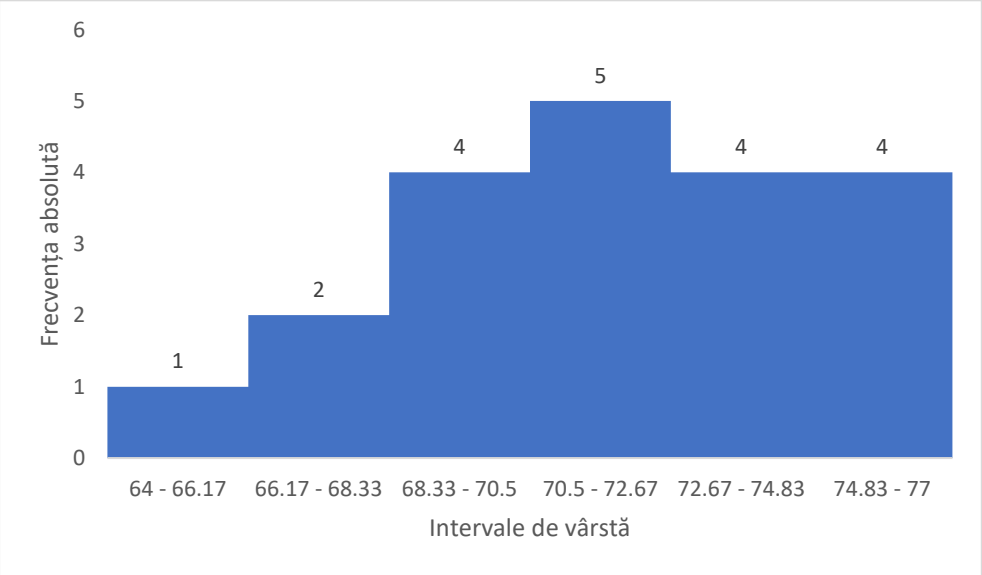


Fig. 2.8. Distribuția vârstelor în intervale de vârstă

Observăm că majoritatea valorilor se concentrează în jurul intervalului 70.50 – 72.67. Histograma ne poate ajuta să identificăm și să eliminăm valorile extreme care ar putea fi anomalii sau erori de măsurare.

2.3.6. Diagrama de tip boxplot

O diagramă de tip boxplot este un grafic care oferă o imagine de ansamblu asupra distribuției unei variabile continue și poate fi folosit pentru a identifica valorile aberante și pentru a compara distribuțiile mai multor grupuri.

Pentru a crea un boxplot, sunt necesare următoarele etape:

1. Determinarea valorilor minime, maxime și medianei (valoarea centrală a datelor) pentru variabila de interes.

2. Calcularea cuartilelor Q1 și Q3.
3. Calcularea intervalului intercuartilic (IQR), care este diferența dintre Q3 și Q1.
4. Identificarea valorilor aberante (extreme), care sunt datele care se află la o distanță mai mare de 1.5 ori IQR de la Q1 sau Q3.
5. Crearea graficului, care constă într-un dreptunghi care reprezintă intervalul intercuartilic (Q1 - Q3), o linie verticală care reprezintă mediana și segmente care se extind din dreptunghi până la cel mai mic și cel mai mare punct care nu sunt valori aberante. Valorile aberante sunt reprezentate prin puncte sau cercuri.

Diagram de tip Boxplot poate fi utilizată în mai multe cazuri, printre care:

1. ***Compararea distribuțiilor***: Este folosit pentru a compara distribuțiile a două sau mai multor variabile continue. Aceasta poate fi utilă pentru a vedea dacă există diferențe semnificative între distribuțiile și dacă acestea au aceeași formă sau nu.
2. ***Identificarea valorilor aberante***: Valorile aberante pot fi semnificative în analiza datelor, deoarece pot indica erori sau probleme în colectarea sau înregistrarea datelor.
3. ***Vizualizarea distribuțiilor asimetrice***: Oferă o imagine de ansamblu asupra medianei și a intervalului intercuartilic. Acest lucru poate fi important în identificarea diferențelor între distribuțiile simetrice și cele asimetrice.
4. ***Compararea mai multor grupuri***: Boxplot-ul poate fi folosit pentru a compara distribuțiile mai multor grupuri, fie că acestea sunt grupuri experimentale, de control sau de alt tip. Acesta poate ajuta la identificarea diferențelor semnificative între grupuri și poate fi util în procesul de interpretare a datelor.

Spre exemplu, cu ajutorul acestui tip de grafic putem compara înălțimea studenților din două clase (Figura 2.9). Din grafic observăm că studenții din clasa B sunt în general mai înalți, dar observăm că avem și o distribuție mai împrăștiată a valorilor.

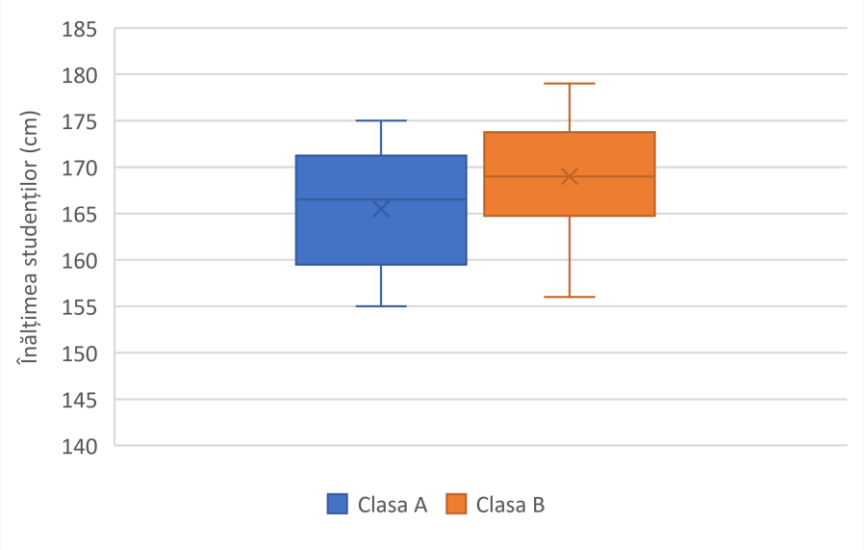


Fig. 2.9. Comparația distribuțiilor înălțimii studenților din două clase diferite

Boxplot-ul este un instrument util pentru a vizualiza distribuția unei variabile continue și poate fi utilizat în multe situații diferite pentru a compara și vizualiza distribuții a variabilelor continue.

Pe lângă tipurile de grafice prezentate în acest capitol mai există foarte multe tipuri de grafice care se pot folosi pentru vizualizarea datelor. Datele geografice pot fi vizualizate pe hărți, evoluția în timp a unor variabile pot fi vizualizate cu ajutorul unor grafice de tip „spaghetti” sau putem vizualiza diferite rezultate și probabilitățile aferente unui experiment cu ajutorul graficului arborescent. Cu toate acestea, graficele prezentate sunt cele mai comune tipuri de grafice și pun bazele altor tipuri de vizualizări. Important este să alegeți tipul de grafic corect în funcție de natura datelor pe care vreți să le reprezentați.

2.4. Verificarea cunoștințelor

Tipuri de date

1. Care sunt cele patru niveluri de măsurare a datelor?
   a) Nominal, ordinal, interval, statistic
   b) Numeric, nominal, ordinal, cardinal
   c) Nominal, ordinal, interval, rație (raport)

2. Care este cea mai mare diferență dintre datele interval și cele rație?
   a) Cele de tip interval au un zero absolut, ceea ce nu se întâmplă în cazul datelor rație.
   b) Cele de tip rație au un zero absolut, ceea ce nu se întâmplă în cazul datelor de tip interval.
   c) Nu există diferențe semnificative între datele interval și cele raționale.

3. Dacă se colectează date despre opinia oamenilor cu privire la satisfacția pentru un nou produs (pe o scală de la 1 la 5), ce nivel de măsurare ar fi aceste date?
   a) Nivel nominal
   b) Nivel ordinal
   c) Nivel interval

4. Ce nivel de măsurare ar fi datele despre starea civilă a unor subiecți?
   a) Nivel nominal
   b) Nivel ordinal
   c) Nivel interval

5. Dacă se colectează date despre temperatura unui loc în diferite momente ale zilei, ce nivel de măsurare ar fi aceste date?

6. Care este diferența dintre datele nominale și cele ordinale?

7. Ce fel de date ar fi data nașterii a unei persoane?

8. Ce nivel de măsurare ar fi datele despre nota obținută la un test?

9. Ce fel de date ar fi înălțimea unei persoane măsurată în cm?
   a) Nivel nominal
   b) Nivel ordinal
   c) Nivel interval
   d) Nivel rație

10. Ce nivel de măsurare ar fi datele despre clasamentul la un concurs sportiv?
   a) Nivel nominal
   b) Nivel ordinal
   c) Nivel interval
   d) Nivel rație

Răspunsuri corecte

1. c) Nominal, ordinal, interval, raport

2. b) Cele de tip rație au un zero absolut, ceea ce nu se întâmplă în cazul datelor de tip interval.

3. b) Nivel ordinal

4. a) Nivel nominal

5. Răspuns: Datele despre temperatura unui loc sunt de nivel interval.

6. Răspuns: Datele nominale sunt utilizate pentru a identifica sau clasifica elemente în categorii, în timp ce datele ordinale au în plus și o ordine naturală.

7. Răspuns: Data nașterii este de nivel interval deoarece anul 0 nu înseamnă lipsa timpului.

8, Răspuns: Datele despre note sunt de nivel rație.

9. d) Nivel rație

10. b) Nivel ordinal

Indicatori statistici de tendință centrală

1. Ce este tendința centrală și cum se calculează?
2. Ce este media și cum se calculează?
3. Ce este mediana și cum se calculează?
4. Ce este modala și cum se calculează?
5. Care este diferența dintre media și mediana unui set de date?
6. Care dintre următoarele nu este o măsură de tendință centrală?
   a. Media
   b. Modala
   c. Dispersia
   d. Mediana

7. Care dintre următoarele măsuri este cea mai potrivită pentru a măsura tendința centrală a unei variabile nominale?
   a. Media
   b. Modala
   c. Mediana
   d. Toate de mai sus

8. Ce se întâmplă cu media dacă se adaugă o valoare extremă mare?
   a. Media crește
   b. Media scade
   c. Media nu se modifică
   d. Media devine negativă

9. Care este mediana setului de date: 3, 7, 9, 11, 12 ?
   a. 7
   b. 9
   c. 11
   d. 10

10. Ce este modala setului de date: 2, 2, 3, 5, 5, 5, 6, 7 ?
   a. 7
   b. 5
   c. 2
   d. 6

Răspunsuri corecte

1. Tendința centrală este o măsură de poziție a valorilor unei distribuții și este utilizată pentru a reprezenta valoarea centrală a acestora. Cele mai frecvent utilizate măsuri de tendință centrală sunt media, mediana și modala.

2. Media este o măsură de tendință centrală care reprezintă suma valorilor dintr-o distribuție împărțită la numărul de valori. Media poate fi afectată de valori extreme, numite valori aberante.

3. Mediana este o măsură de tendință centrală care reprezintă valoarea din mijlocul unei distribuții ordonate crescător sau descrescător. Dacă numărul de valori este par, mediana este media dintre cele două valori centrale.

4. Modala este o măsură de tendință centrală care reprezintă valoarea care apare cel mai frecvent într-o distribuție. Dacă toate valorile apar cu aceeași frecvență, distribuția este bimodală sau multimodală.

5. Media aritmetică a unui set de date se determină împărțind suma valorilor la numărul acestora, în timp ce mediana este valoarea din mijlocul acesteia. Media este afectată de valori extreme, în timp ce mediana nu este.

6. c. Dispersia

7. b. Modala

8. a. Media crește

9. b. 9

10. b. 5, deoarece aceasta este valoarea care apare cel mai frecvent în distribuție, de trei ori.

Indicatori statistici de împrăștiere

1. Care este definiția indicatorului de împrăștiere „amplitudine”?
2. Cum se calculează intervalul intercuartilic și ce informații furnizează acesta?
3. Ce este dispersia și cum se calculează?
4. Care este relația dintre dispersie și abatere standard?
5. Când ar trebui să folosim dispersia în locul abaterii standard?
6. Care dintre următoarele afirmații este adevărată pentru intervalul intercuartilic?
   a. Reprezintă diferența dintre valoarea minimă și valoarea maximă a setului de date.
   b. Reprezintă diferența dintre a treia și prima cuartilă.
   c. Reprezintă diferența dintre medie și mediana setului de date.
   d. Reprezintă diferența dintre două valori consecutive din setul de date.
7. Care dintre următoarele afirmații este falsă referitor la abaterea standard?
   a. Reprezintă rădăcina pătrată a dispersie.
   b. Măsoară variabilitatea sau împrăștiere valorilor în raport cu media.
   c. Este exprimată în aceleași unități de măsură ca și media.
   d. Este mereu mai mică sau egală cu dispersia.
8. Dacă toate valorile dintr-un set de date sunt egale, atunci:
   a. Dispersia este zero, dar abaterea standard nu este zero.
   b. Abaterea standard este zero, dar dispersia nu este zero.
   c. Atât dispersia, cât și abaterea standard sunt zero.
   d. Dispersia și abaterea standard nu pot fi calculate în această situație.
9. Dacă adăugăm o valoare extrem de mare sau extrem de mică la un set de date, cum se va modifica abaterea standard?
   a. Va crește
   b. Va scădea
   c. Nu se va schimba
   d. Va deveni negativă
10. Care dintre afirmații este adevărată referitor la dispersie și abatere standard?
    a. Dispersia și abaterea standard sunt mereu egale.
    b. Dispersia și abaterea standard pot fi egale în anumite situații.
    c. Dispersia și abaterea standard sunt mereu diferite.
    d. Dispersia și abaterea standard sunt măsurate în aceeași unitate de măsură.

Răspunsuri corecte

1. Amplitudinea reprezintă diferența dintre valoarea maximă și valoarea minimă a unui set de date.

2. Intervalul intercuartilic se calculează prin diferența dintre a treia și prima cuartilă și furnizează informații despre împrăștierea valorilor din mijlocul setului de date.

3. Dispersia reprezintă media pătratelor diferențelor dintre fiecare valoare din setul de date și media acestuia și se calculează prin împărțirea sumei acestor diferențe la numărul total de valori din setul de date.

4. Dispersia este pătratul abaterii standard.

5. Dispersia este preferabilă atunci când dorim să comparăm variația între două sau mai multe seturi de date care au unități de măsură diferite, în timp ce abaterea standard este preferabilă atunci când dorim să comparăm variația între două sau mai multe seturi de date care au aceeași unitate de măsură.

6. b. Reprezintă diferența dintre a treia și prima cuartilă.

7. d. Este mereu mai mică sau egală cu dispersia.

8. c. Atât dispersia, cât și abaterea standard sunt zero.

9. a) Va crește.

10. b) Dispersia și abaterea standard pot fi egale în anumite situații.

Vizualizarea datelor

1. Ce este o diagramă cu linii și când este cel mai bine să îl folosim?

2. Cum putem compara două variabile folosind o diagramă cu puncte?

3. Ce este o diagramă cu bare și când este utilă?

4. Când este util un grafic cu puncte?

5. Ce tip de grafic este cel mai bun pentru a arăta tendințele în timp?
   a. Diagrama circulară
   b. Diagrama cu linii
   c. Histograma
   d. Diagrama cu puncte

6. Care este cel mai bun mod de a vizualiza distribuția unei variabile continue?
   a. Diagrama circulară
   b. Diagrama cu linii
   c. Histograma
   d. Diagrama cu puncte

7. Ce este o diagramă circulară și când este utilă?
   a. Un grafic care arată distribuția proporțiilor unei variabile
   b. Un grafic care arată tendințele în timp
   c. Un grafic care compară două variabile
   d. Un grafic care arată relația dintre două variabile

8. Ce este un grafic cu puncte și cum poate fi folosit?
   a. Un grafic care arată distribuția unei variabile
   b. Un grafic care compară două variabile
   c. Un grafic care arată relația dintre două variabile
   d. Un grafic care arată tendințele în timp

9. Ce este o diagramă boxplot și cum poate fi folosit pentru a analiza datele?
   a. Un grafic care arată distribuția unei variabile
   b. Un grafic care compară două variabile discrete
   c. Un grafic care arată relația dintre două variabile
   d. Un grafic care arată tendințele în timp

10. Care este cea mai bună modalitate de a arăta distribuția unei variabile de tip nominal?
   a. Diagramă circulară
   b. Grafic de linie
   c. Histogramă
   d. Diagrama cu bare

Răspunsuri corecte

1. O diagramă cu linii este un grafic care folosește linii pentru a arăta evoluția cronologică de la o valoare la alta. Este utilă atunci când dorim să vedem evoluția unei variabile în timp.

2. O diagramă cu puncte este utilă pentru a arăta relația dintre două variabile continue. Putem observa dacă există o asociere între cele două variabile și dacă da, care este tipul acesteia.

3. O diagramă cu bare este utilă când dorim să comparăm cantități între diferite categorii sau grupuri. De exemplu, putem utiliza un grafic cu bare pentru a compara vânzările a două companii într-un anumit trimestru.

4. Un grafic cu puncte este util pentru a vizualiza relația dintre variabile continue.

5. b) Grafic cu linii

6. c) Histogramă

7. a) Un grafic care arată distribuția proporțiilor unei variabile

8. c) Un grafic care arată relația dintre două variabile

9. a) Un grafic care arată distribuția unei variabile

10. d) Diagrama cu bare

3. Evenimente statistice

Pentru a înțelege noțiunea de probabilitate, trebuie să definim niște termeni. Când vorbim despre probabilități, mai întâi facem un experiment statistic. **Un experiment** este o procedură care se întâmplă în anumite condiții predefinite. Într-un experiment putem avea una sau mai multe **încercări**. Fiecare încercare are un **rezultat** care este numit un **eveniment** și pe care îl putem observa. Setul evenimentelor dintr-un experiment se numește **câmp de evenimente**.

Un exemplu des întâlnit de experiment este aruncarea unei monede. Putem să o aruncăm de mai multe ori, ceea ce înseamnă că avem mai multe încercări. Rezultatele pot fi *cap* sau *pajură*, acestea fiind rezultatele posibile. Cap și pajură reprezintă întregul câmp de evenimente atunci când experimentul constă dintr-o singură aruncare a monedei.

3.1. Definirea evenimentelor

Există mai multe tipuri de evenimente cum ar fi cele simple, care sunt un singur eveniment, cum ar fi aruncarea unui zar și obținerea unui număr par, și evenimente compuse, care sunt compuse din mai multe evenimente simple, cum ar fi obținerea a două numere pare la două aruncări diferite de zar. Când aruncăm două zaruri putem obține o combinație de numere. Evenimentul în care suma numerelor este egală cu 7, este un eveniment compus.

Unele evenimente au semnificații aparte, cum ar fi **evenimentul imposibil** și **evenimentul sigur**. În cazul în care rezultatul nu poate avea loc în cadrul unui experiment, acesta este numit un eveniment imposibil. De exemplu, nu putem obține numărul 7 atunci când se aruncă un zar cu 6 laturi. Evenimentul sigur este opusul celui imposibil; este sigur să se întâmple. De exemplu, obținerea unui număr par sau impar la aruncarea unui zar, este un eveniment sigur.

Evenimente sunt notate cu majuscule și conținutul lor sunt puse în interiorul unor acolade:

$$A = \{1, 2, 3\}$$

Evenimentul A conține elementele 1, 2 și 3.

Un eveniment compus poate avea mai multe elemente sau grupuri de elemente:

$$B = \{(1, 2), (3, 4)\}$$

Evenimentul B conține două grupe de evenimente (1, 2) și (3, 4). Fiecare grup are câte două elemente fiecare.

Un mod util de reprezentare a evenimentelor este prin utilizarea diagramelor Venn. Diagramele Venn sunt forme pe care le folosim pentru a reprezenta mulțimi de evenimente. Să spunem că avem următoarele două evenimente:

$$A = \{1, 2, 3\}$$

$$B = \{4, 5, 6\}$$

Câmpul de evenimente (notat cu S) este:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Putem desena o diagramă Venn ca cea din figura 3.1.

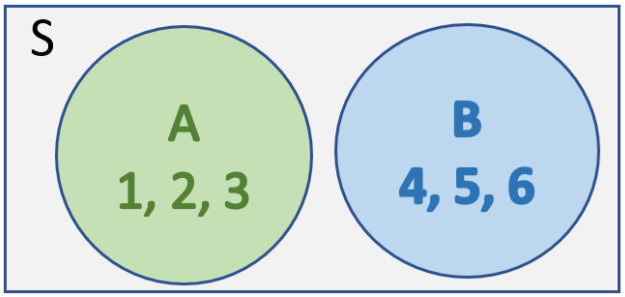


Fig. 3.1. Reprezentarea evenimentelor prin diagrame Venn

În cazul în care câmpul de evenimente este format din doar două mulțimi care nu au elemente în comun, acestea sunt numite **mulțimi disjuncte**. Evenimentele A și B din exemplul de mai sus sunt mulțimi disjuncte.

3.2. Tipuri de evenimente

Evenimente pot fi fie **compatibile** sau **incompatibile**. Evenimentele compatibile pot avea loc simultan în același experiment, în timp ce evenimentele incompatibile nu pot. Dacă aruncăm o monedă o dată, evenimentul de a obține cap și evenimentul de a obține pajură sunt incompatibile. Putem avea fie una fie alta. Pe de altă parte, dacă aruncăm o monedă de două ori, aceleași două evenimente devin compatibile. La o aruncare puteți obține cap și la a doua pajură.

Evenimente pot fi **dependente** sau **independente**. Dacă un eveniment are impact asupra altuia, atunci acestea sunt dependente. Dacă nu, sunt independente. Dacă aveți o urnă cu 5 bile negre și 5 bile albe, extragerea unei bile din urnă schimbă raportul dintre bilele albe și negre, care are impact asupra rezultatului celei de-a doua extrageri. În acest caz, a doua extragere depinde de prima. Pe de altă parte, dacă am arunca o monedă de două ori, rezultatul primei aruncări nu are efect asupra rezultatului celei de-a doua aruncări.

3.3. Operațiuni cu evenimente

Există două tipuri de operațiuni de care suntem interesați în utilizarea evenimentelor: **reuniunea** și **intersecția**.

Reuniunea (Figura 3.2) este adunarea elementelor a două sau mai multe mulțimi (sau evenimente). Este notată cu ∪ și poate fi interpretată ca adunare (+). Pentru evenimentele A și B avem:

$$A = \{1, 2\}$$

$$B = \{3, 4\}$$

$$A \cup B = \{1, 2, 3, 4\}$$

Și citim "*A sau B*" sau "*A reunit B*"

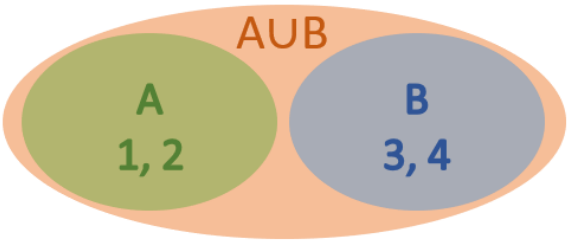


Fig. 3.2. Reuniunea a două evenimente

Intersecția (Figura 3.3) este setul de elemente comune pentru două sau mai multe mulțimi. Este notat cu ∩ și este interpretat ca înmulțire (*). Pentru două evenimente A și B, intersecția lor este:

$$A = \{1, 2, 3\}$$

$$B = \{3, 4\}$$

$$A \cap B = \{3\}$$

Și citim "*A și B*" sau "*A intersectat cu B*"

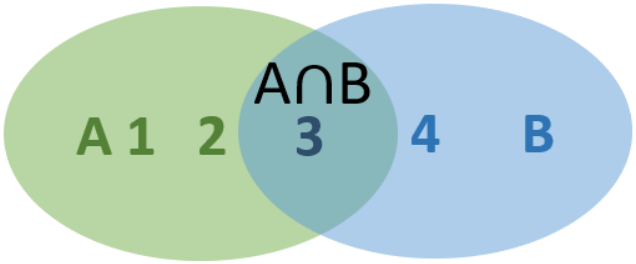


Fig. 3.3. Intersecția a două evenimente

Complementul (Figura 3.4) unui eveniment este tot ceea ce nu este în acel eveniment. Dacă avem un eveniment A, complementul său este marcat cu $\bar{A}$ și reprezintă toate elementele care nu fac parte din A.

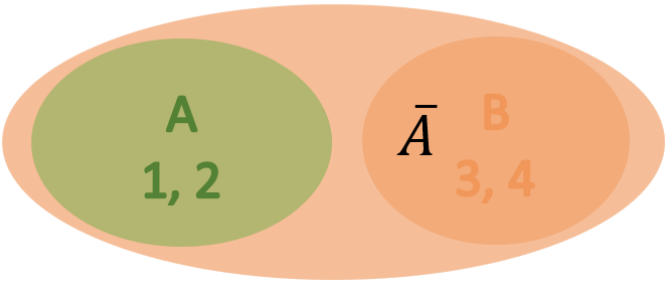


Fig. 3.4. Complementul unui eveniment

Complementul evenimentului sigur este evenimentul imposibil. Invers este de asemenea adevărat, complementul evenimentului imposibil este evenimentul sigur.

Intersecția și Reuniunea au următoarele proprietăți:

$$A \cup A = A \qquad A \cap A = A$$

$$A \cup B = B \cup A \qquad A \cap B = B \cap A$$

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cup \bar{A} = E \qquad A \cap \bar{A} = \Phi$$

$$A \cup E = E \qquad A \cup \Phi = A$$

$$E \cap \Phi = \Phi \qquad E \cup \Phi = E$$

$$(\overline{A \cap B}) = \bar{A} \cap \bar{B} \qquad (\overline{A \cup B}) = \bar{A} \cup \bar{B}$$

3.4. Verificarea cunoștințelor

1. Care dintre următoarele opțiuni descrie cel mai bine un eveniment în contextul probabilității?
   a. Un proces aleatoriu
   b. O situație imposibilă
   c. O combinație de cifre
   d. O afirmație matematică

2. Ce este un eveniment compus?
   a. Un eveniment care nu se poate întâmpla niciodată
   b. Un eveniment care constă într-un singur rezultat posibil
   c. Un eveniment care combină mai multe evenimente simple
   d. Un eveniment care apare în mod necesar

3. Care dintre următoarele reprezintă un exemplu de eveniment simplu?
   a. Obținerea a două fețe de zar diferite
   b. Extragerea unei cărți roșii dintr-un pachet de cărți
   c. Întâlnirea unei persoane cu numele "John"
   d. Atingerea unei temperaturi de -100 grade Celsius

4. Ce este câmpul de evenimente?
   a. Un set de evenimente imposibile
   b. Un set care conține toate evenimentele posibile într-un experiment statistic
   c. O zonă geografică cu proprietăți neobișnuite
   d. Un termen folosit în geografie

5. Ce este reuniunea a două evenimente?
   a. Evenimentul care apare în ambele evenimente
   b. Evenimentul care apare în unul sau în celălalt eveniment
   c. Evenimentul care nu apare în niciunul dintre cele două evenimente

6. Ce este intersecția a două evenimente?
   a. Evenimentul care apare în ambele evenimente
   b. Evenimentul care apare în unul sau în celălalt eveniment
   c. Evenimentul care nu apare în niciunul dintre cele două evenimente

7. Ce este evenimentul complementar?
   a. Evenimentul care apare cel mai frecvent într-un experiment
   b. Evenimentul care nu apare niciodată într-un experiment
   c. Evenimentul care nu apare în evenimentul original

8. Ce este o secvență de evenimente independente?
   a. Evenimente care nu se pot întâmpla în același timp
   b. Evenimente care nu se influențează unul pe celălalt
   c. Evenimente care se întâmplă întotdeauna împreună

9. Ce este o serie de evenimente dependente?
   a. Evenimente care nu se pot întâmpla în același timp
   b. Evenimente care nu se influențează unul pe celălalt
   c. Evenimente care depind de unul sau mai multe evenimente anterioare

10. Despre ce este vorba când două evenimente se exclud reciproc?
   a. Ele nu pot apărea în același timp
   b. Ele sunt independente și nu se influențează reciproc
   c. Ele au o intersecție nevidă

Răspunsuri corecte

1. a) Un proces aleatoriu

2. c) Un eveniment care combină mai multe evenimente simple

3. b) Extragerea unei cărți roșii dintr-un pachet de cărți

4. b) Un set care conține toate evenimentele posibile într-un experiment statistic

5. b) Evenimentul care apare în unul sau în celălalt eveniment

6. a) Evenimentul care apare în ambele evenimente

7. c) Evenimentul care nu apare în evenimentul original

8. b) Evenimente care nu se influențează unul pe celălalt

9. c) Evenimente care depind de unul sau mai multe evenimente anterioare

10. a) Ele nu pot apărea în același timp

4. Probabilitatea

Probabilitatea este o modalitate de cuantificare a șansei unui eveniment de a avea loc sau nu. De obicei, este exprimată numeric fie cu valori zecimale între 0 și 1, fie cu procente între 0% și 100%. Valorile extreme sunt valori teoretice și corespund evenimentului imposibil (0%) și evenimentul sigur (100%). Probabilitatea unui eveniment X se notează cu $P(X)$.

Pentru a determina probabilitatea producerii unui eveniment, trebuie să cunoaștem numărul evenimentelor favorabile și numărul total de evenimente dintr-un experiment. Probabilitatea producerii unui eveniment este numărul evenimentelor favorabile împărțit la numărul total de evenimente egal probabile:

$$P(X) = \frac{nr.evenimente\ favorabile}{nr.evenimente\ egal\ probabile}$$

Spre exemplu, atunci când se arunca un zar cu șase fețe, avem șase evenimente la fel de probabile (numerele de la 1 la 6). Dacă suntem interesați în obținerea numărului 6, atunci spunem că acesta este evenimentul nostru favorabil. Probabilitatea de a obține 6 atunci când aruncăm un zar este în consecință de 1/6.

4.1. Probabilitate condiţionată

Rezultatul unor evenimente poate depinde de rezultatul altor evenimente. Acestea sunt numite evenimente dependente. În acest caz, probabilitatea apariției unui eveniment depinde de probabilitatea apariției unui alt eveniment. Aceasta se numește **probabilitate condiționată**. Dacă avem un eveniment A care depinde de rezultatul evenimentului B, atunci vom spune: *"probabilitatea ca evenimentul A să aibă loc , **dat fiind că** a avut loc evenimentul B* " și vom scrie $P(A|B)$.

Știm că evenimentul B a avut loc și singurul loc unde avem A în B este intersecția lor. Drept urmare probabilitatea ca A să aibă loc dat fiind că a avut loc B este probabilitatea intersecții lui A cu B raportată la probabilitatea ca B să se întâmple:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Să presupunem că avem un pachet standard de 52 de cărți de joc și vrem să calculăm probabilitatea de a trage un as, presupunând că știm că cartea trasă este o carte de pică. În acest caz, evenimentul *A* este "*a trage un as*", iar evenimentul *B* este "*a trage o carte de pică*".

Există 4 ași în pachet și 13 cărți de pică. Doar unul dintre aceștiași este și carte de pică (Asul de pică). Deci, probabilitatea de a trage un as din pachet este $P(A) = \frac{4}{52}$, iar

probabilitatea de a trage o carte de pică este $P(B) = \frac{13}{52}$. Probabilitatea de a trage Asul de pică, care este și as și carte de pică, este $P(A \cap B) = \frac{1}{52}$.

Utilizând formula probabilității condiționate, avem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

Deci, probabilitatea de a trage un as, presupunând că am tras deja o carte de pică, este 1/13.

4.2. Reguli de înmulţire şi adunare

În funcție de tipul evenimentului (dependent, independent, compatibil, incompatibil) intersecția și reuniunea funcționează diferit. Tabelul 4.1 rezumă calculele pentru fiecare caz.

Tabel 4.1. Regulile de înmulțire și adunare în funcție de tipurile de evenimente

Operațiunea	Tipul evenimentului	Calculul probabilității
Reuniune	Incompatibile	$P(A \cup B) = P(A) + P(B)$
	Compatibile și independente	$P(A \cup B) = P(A) + P(B) - P(A) * P(B)$
	Compatibile și dependente	$P(A \cup B) = P(A) + P(B) - P(A) * P(B\|A)$
Intersecție	Compatibile și independente	$P(A \cap B) = P(A) * P(B)$
	Compatibile și dependente	$P(A \cap B) = P(A) * P(B\|A)$

4.3. Legea probabilităţii totale

Legea probabilității totale (sau formula probabilității totale) este o metodă folosită pentru a calcula probabilitatea unui eveniment A, luând în considerare toate posibilele căi de realizare a acestuia. Dacă B1, B2, B3,... Bn sunt evenimente mutual exclusive și acoperă întreg câmp de evenimente S, atunci pentru orice eveniment A avem:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Aceasta înseamnă că (Figura 4.1), fiind date patru evenimente (B_1, B_2, B_3, B_4) care ocupă întregul spațiu de eșantionare și un eveniment dependent de aceste patru, probabilitatea ca evenimentul A să se producă este suma probabilităților apariției lui A în fiecare dintre cele patru evenimente.

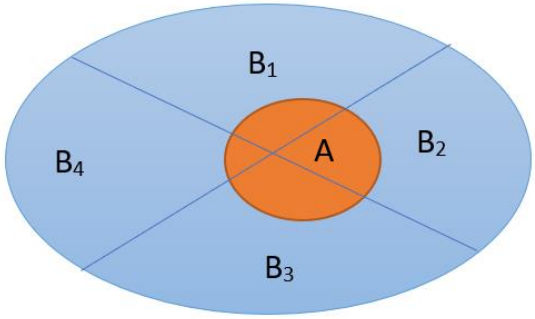


Fig. 4.1. Legea probabilității totale

Spre exemplu, dacă avem o țară cu patru regiuni distincte (B_1 - B_4) și suntem interesați de zona muntoasă (A), atunci suprafața totală cu munți este:

$$A = \sum_{i=1}^{4} A \cap B_i$$

4.4. Regula lui Bayes

Regula lui Bayes este o metodă matematică care ne permite să actualizăm probabilitatea unui eveniment, pe baza informațiilor noi care devin disponibile. Regula se bazează pe ideea că probabilitatea unui eveniment poate fi calculată diferit, în funcție de informațiile disponibile. Mai precis, ea ne permite să calculăm probabilitatea unui eveniment A, bazat pe probabilitatea inițială a evenimentului A și pe o nouă informație B care devine disponibilă.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

unde:

- P(A|B) reprezintă probabilitatea evenimentului A dat fiind informația B;

- P(B|A) reprezintă probabilitatea informației B dat fiind evenimentul A;

- P(A) reprezintă probabilitatea inițială a evenimentului A;

- P(B) reprezintă probabilitatea informației B.

Regula lui Bayes este foarte utilă în multe domenii, cum ar fi medicina, informatica, ingineria, economia și multe altele. De exemplu, în medicină, regula lui Bayes poate fi folosită pentru a calcula probabilitatea ca un pacient să aibă o anumită boală, bazat pe simptomele sale și pe istoricul medical. În inginerie, regula lui Bayes poate fi folosită pentru a calcula fiabilitatea unui sistem complex, bazat pe datele disponibile despre componentele sale individuale.

Să zicem că lucrați într-o fabrică care produce becuri. Fiecare lucrător are postul său și produce becuri care sunt apoi colectate și puse împreună într-un container.

Recipientul se duce apoi la verificarea calității. La verificarea calității, se constată că un bec este defect. Care este probabilitatea ca acesta să vină de la postul Dvs.?

Notăm cu P(D) probabilitatea ca becul să fie defect. Apoi P(X) este probabilitatea ca becul să fi venit de la postul Dvs. Să presupunem, pentru simplitate, că există doar un alt lucrător și probabilitatea ca un bec să provină de la postul său este P(Y).

Pentru că există doar doi lucrători (Dvs. și Y) suma probabilităților lor trebuie să fie egală cu 100%:

$$P(X) + P(Y) = 1$$

Dacă desenăm o diagramă Venn arată ca în figura 4.2.

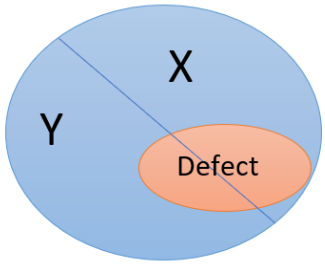


Fig. 4.2. Regula ui Bayes

În partea de sus sunt piesele făcute de X, dintre care unele sunt defecte și în partea de jos există becuri făcute de Y, dintre care unele sunt, de asemenea, defecte.

Dacă considerăm acest lucru din punct de vedere al evenimentului, putem scrie:

$$P(X \cap D) \; = \; P(D \cap X)$$

Deoarece D depinde de X, acest lucru poate fi scris ca:

$$P(X) * P(D|X) = \; P(D) * P(X|D)$$

Probabilitatea de care suntem interesați de este $P(X|D)$, și anume probabilitatea unui bec defect să provină de la postul Dvs. Aceasta o extragem din formula de mai sus:

$$P(X|D) = \frac{P(D|X) * P(X)}{P(D)}$$

În contextul Regulii lui Bayes, probabilitatea a priori (P(A)) reprezintă probabilitatea ipotezei înainte de a lua în considerare dovezile, iar probabilitatea posteriori reprezintă probabilitatea ipotezei date fiind dovezile noi (P(A|B)).

4.5. Verificarea cunoștințelor

1. Ce este probabilitatea condiționată a evenimentului A dat fiind evenimentul B?
   a. Probabilitatea evenimentului A și evenimentului B
   b. Probabilitatea evenimentului A dat fiind că evenimentul B s-a întâmplat
   c. Probabilitatea evenimentului B dat fiind că evenimentul A s-a întâmplat

2. Când sunt două evenimente independente?
   a. Evenimente care nu se pot întâmpla în același timp
   b. Evenimente care nu se influențează unul pe celălalt
   c. Evenimente care depind de unul sau mai multe evenimente anterioare

3. Ce este probabilitatea intersecției a două evenimente compatibile independente?
   a. P(A) + P(B)
   b. P(A) - P(B)
   c. P(A) * P(B)

4. Ce este legea probabilității totale?
   a. Metoda de a calcula probabilitatea unei reuniuni de evenimente
   b. Metoda de a calcula probabilitatea unei intersecții de evenimente
   c. Metoda de a calcula probabilitatea unui eveniment dat fiind alte evenimente

5. Ce este evenimentul condiționat?
   a. Evenimentul care depinde de alte evenimente
   b. Evenimentul care nu depinde de alte evenimente
   c. Evenimentul care nu poate să apară împreună cu alte evenimente

6. Ce este probabilitatea a prior?
   a. Probabilitatea unui eveniment cu informația actualizată
   b. Probabilitatea unui eveniment dat fiind alte evenimente
   c. Probabilitatea unui eveniment înainte de a avea orice informație nouă

7. Probabilitatea posterioară este:
   a. Probabilitatea unui eveniment dată anterior
   b. Probabilitatea unui eveniment după ce se iau în considerare alte informații noi
   c. Probabilitatea unui eveniment dat fiind alte evenimente

8. Un test medical este pozitiv pentru o anumită boală în 95% din cazurile în care persoana este bolnavă, dar și în 2% din cazurile în care persoana este sănătoasă. Dacă o persoană testează pozitiv, care este probabilitatea ca aceasta să fie cu adevărat bolnavă știind că 3% din persoane au această boală?

a. Probabilitatea este de aprox. 60%.
b. Probabilitatea este de aprox. 50%.
c. Probabilitatea este de aprox. 75%.
d. Probabilitatea este de aprox. 5%.

9. Într-un oraș există două companii de taxi, A și B. 80% dintre taxiurile companiei A sunt de culoare galbenă, iar 60% dintre taxiurile companiei B sunt de culoare galbenă. Dacă 60% dintre taxiurile din oraș sunt ale companiei A, care este probabilitatea ca un taxi de culoare galbenă selectat aleatoriu să fie de la compania B?

a. 0.25
b. 0.40
c. 0.50
d. 0.33

10. Dacă 25% din studenții unei facultăți au laptop-uri Apple, iar 60% dintre cei cu laptop Apple au și un iPhone, care este probabilitatea ca un student selectat la întâmplare să aibă un laptop Apple și un iPhone?

Răspunsuri corecte

1. b) Probabilitatea evenimentului A dat fiind că evenimentul B s-a întâmplat

2. b) Evenimente care nu se influențează unul pe celălalt

3. c) P(A) * P(B)

4. c) Metoda de a calcula probabilitatea unui eveniment dat fiind alte evenimente

5. a) Evenimentul care depinde de alte evenimente

6. c) Probabilitatea unui eveniment înainte de a avea orice informație nouă

7. b) Probabilitatea unui eveniment după ce se iau în considerare alte informații noi

8. a) Probabilitatea este de aprox. 60%.

Justificare: Probabilitatea ca o persoană să fie cu adevărat bolnavă dat fiind un rezultat pozitiv al testului este dată de formula Bayes, astfel încât:

$$P(Bolnav \mid Pozitiv) = \frac{P(Pozitiv \mid Bolnav) * P(Bolnav)}{P(Pozitiv \mid Bolnav) * P(Bolnav) + P(Pozitiv \mid Sănătos) * P(Sănătos)}$$

Folosind datele din enunțul problemei, se poate calcula că:

$$P(Bolnav \mid Pozitiv) = \frac{0.95 * 0.03}{0.95 * 0.03 + 0.02 * 0.97} / () = 0.595 \approx 0.60$$

9. d) 0.33. Legea probabilității totale spune că probabilitatea unui eveniment poate fi calculată ca suma probabilităților condiționate de evenimentele care îl pot influența. În acest caz, evenimentele sunt *Y = „taxi de culoare galbenă*", *A = „compania A”, B = „compania B*". Putem folosi formula legii probabilității totale:

$$P(Y) = P(Y|A) * P(A) + P(Y|B) * P(B)$$

Substituind valorile cunoscute în această formulă, obținem:

$$P(Y) = (0.8\ x\ 0.6) + (0.6\ x\ 0.4) = 0.48 + 0.24 = 0.72$$

Probabilitatea ca un taxi de culoare galbenă selectat aleatoriu să fie de la compania B este:

$$P(B|Y) = \frac{P(Y|B) * P(B)}{P(Y)} = \frac{0.6 * 0.4}{0.72} = 0.33$$

10. Pentru a găsi probabilitatea ca un student selectat la întâmplare să aibă un laptop Apple și un iPhone, trebuie să folosim teorema probabilităților condiționate. Notăm evenimentul "are laptop Apple" cu A și evenimentul "are iPhone" cu B. Avem P(A) = 0.25,

probabilitatea ca un student selectat la întâmplare să aibă un laptop Apple, și P(B|A) = 0.6, probabilitatea că un student care are un laptop Apple să aibă și un iPhone. Dorim să găsim P(A ∩ B), adică probabilitatea ca un student selectat la întâmplare să aibă atât laptop Apple, cât și iPhone. Folosind formula probabilităților condiționate, avem:

$$P(A \cap B) = P(B|A) * P(A) = 0.6 * 0.25 = 0.15$$

Prin urmare, probabilitatea ca un student selectat la întâmplare să aibă atât laptop Apple, cât și iPhone este de 0.15 sau 15%.

5. Variabile aleatorii

O **variabilă aleatorie** este o funcție matematică care atribuie o valoare numerică fiecărui posibil rezultat a unui eveniment dintr-un experiment statistic. În funcție de experiment, variabilele aleatorii pot fi discrete sau continue. Noi spunem că o variabilă aleatorie este **discretă**, atunci când se pot obține doar anumite valori. De exemplu, atunci când aruncăm un zar, putem obține doar numerele de pe zar (numerele de la unul la șase). Nu putem obține valoarea 2.7. Pe de altă parte, dacă rezultatul experimentului este un număr real, spunem că variabila aleatorie este **continuă**. De fiecare dată când măsurăm ceva, cum ar fi lungimea unui obiect, avem de obicei o valoare reală (una care poate avea un număr infinit de zecimale)

5.1. Variabile aleatorii discrete

Variabilele aleatorii discrete pot lua doar anumite valori și au o anumită probabilitate asociată acestora. Dacă experimentul nostru constă în a arunca un zar de 100 ori, de exemplu, putem determina probabilitatea de a obține un număr de pe zar. Tot ce trebuie să facem este să împărțim numărul de apariții ale unei anumite fețe la total. Dacă am aruncat zarul de 100 ori și șase a apărut de 20 de ori, atunci probabilitatea de a obține un șase cu acest zar este de 20% (20/100).

Variabilele aleatorii sunt notate cu majuscule. Putem pune toate rezultatele posibile și probabilitățile lor asociate într-un tabel. Pentru exemplul zarului putem scrie:

$$X: \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.1 & 0.15 & 0.2 & 0.1 & 0.25 & 0.2 \end{pmatrix}$$

Un lucru de observat este că probabilitatea apariției unui eveniment este frecvența sa relativă. Dacă reprezentăm grafic, vom obține diagrama din figura 5.1.

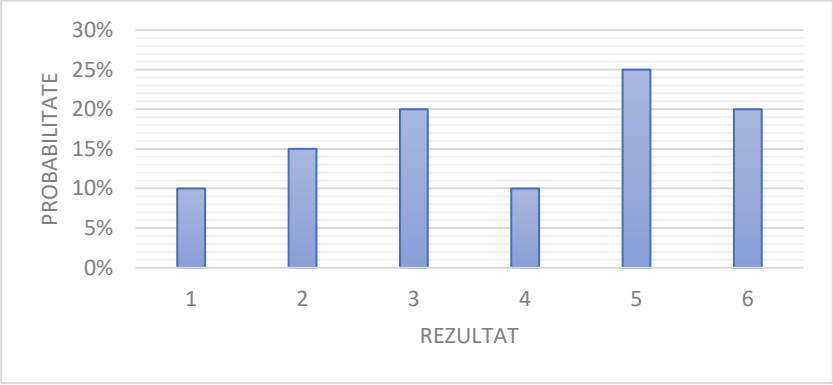


Fig. 5.1. Distribuția rezultatelor la aruncarea unui zar de 100 de ori

Acesta este modul în care probabilitățile sunt aranjate sau distribuite. Aceasta se numește o distribuție. Graficul de mai sus ne ajută să vizualizăm distribuția probabilităților pentru acest zar. Dacă este un zar corect (asupra căruia nu s-a intervenit astfel încât să favorizeze un anumit rezultat), toate părțile au aceeași probabilitate de

apariție. Dacă îl aruncăm de un număr mare de ori, probabilitățile se vor apropia toate de 1/6 iar graficul va deveni plat (toate coloanele au aproximativ aceeași înălțime). Funcția care atribuie probabilitatea fiecărei valori din X este denumită **funcție de repartiție**.

Când suntem interesați de o anumită valoare a unei variabile aleatorii (de exemplu, X = 3), utilizăm funcția de repartiție pentru a o determina, P (X = 3) = 0.2.

Probabilitățile pentru orice valoare din X trebuie să fie mai mare sau egală cu zero:

$$P(X = x) \geq 0 \ , pentru\ toate\ \ x\ din\ X$$

Suma tuturor probabilităților trebuie să fie egală cu 1:

$$\sum P(x) = 1$$

Variabilele aleatorii discrete sunt o parte importantă a teoriei probabilităților și a statisticii și sunt utilizate într-o varietate de domenii, inclusiv data science, economie și inginerie.

5.2. Variabile aleatorii continue

Variabilele aleatorii care conțin numere reale ($\mathbb{R}$) sunt numite variabile aleatorii continue, deoarece putem obține orice număr într-un anumit interval. În acest caz, nu putem desena un tabel de valori și probabilitățile lor, deoarece avem un număr infinit de valori posibile. Datorită acestui fapt, probabilitatea unei anumite valori specifice va fi 0, deoarece avem un singur număr favorabil dintr-o infinitate de numere egal posibile. Putem totuși reprezenta probabilitatea grafic. Deoarece toate valorile sunt apropiate unele de celelalte, graficul este o curbă (Figura 5.2).

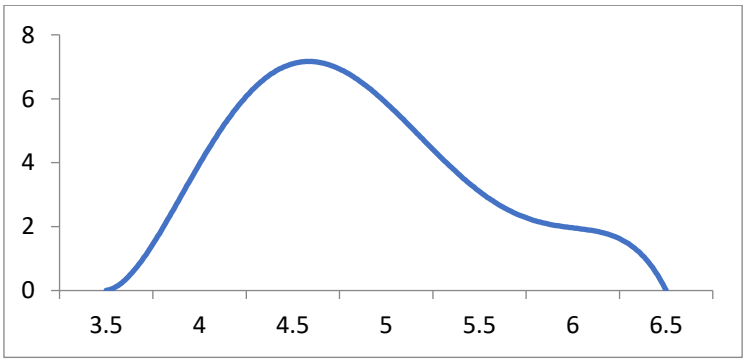


Fig. 5.2. Exemplu de curbă pentru o distribuție continuă oarecare

Suprafața de sub curba dintre două valori a și b reprezintă probabilitatea ca un eveniment x să fie între aceste două valori și poate fi determinat prin utilizarea unei integrale:

$$P(a < x < b) = \int_a^b f(x)dx$$

Să presupunem că doriți să determinați probabilitatea ca un aparat de cafea să vă dea între 95 și 100 ml de cafea. Deoarece cantitatea de cafea poate avea orice valoare reală, aceasta este o variabilă aleatorie continuă. Prin luarea unui număr de măsurători (să zicem 100), putem determina funcția, f(x), care ne dă curba și apoi putem scrie:

$$P(95 < x < 100) = \int_{95}^{100} f(x)dx$$

Prin rezolvarea integralei, putem determina probabilitatea ca *x* să ia valori între 95 și 100.

5.3. Funcția de distribuție cumulativă

Uneori trebuie să știm probabilitatea obținerii tuturor valorilor până la una anume. La un joc cu zaruri, am putea ajunge în situația în care câștigăm dacă aruncăm orice număr până la numărul 4. Pentru a determina care este probabilitatea noastră de a câștiga, trebuie să adunăm probabilitățile tuturor valorilor până la 4 (1, 2, 3 și 4). Această sumă este probabilitatea cumulată. Funcția care ne dă distribuția acestor probabilități cumulate se numește ***funcție de distribuție cumulativă***. Dacă luăm variabila aleatorie X din exemplul anterior:

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.1 & 0.15 & 0.2 & 0.1 & 0.25 & 0.2 \end{pmatrix}$$

și scriem probabilitatea cumulată pentru fiecare valoare, vom obține următorul tabel:

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.1 & 0.25 & 0.45 & 0.55 & 0.75 & 1 \end{pmatrix}$$

Dacă suntem interesați de probabilitatea cumulată până la o anumită valoare, (să zicem 3) ne uităm pentru această valoare în primul rând (3) și citiți probabilitatea asociată (0.45). Acest lucru înseamnă că probabilitatea de a obține toate valorile până la 3 este 0.45, sau 45%.

În cazul unei variabile aleatorii continue folosim integrala. Pornim de la -∞ până la valoarea dorită.

$$P(x < a) = \int_{-\infty}^{a} f(x)dx$$

Putem reprezenta grafic distribuția probabilităților ca în figura 5.3 atât pentru variabile discrete cât și pentru cele continue.

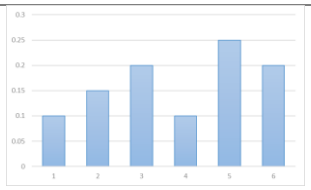
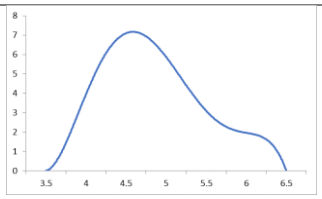
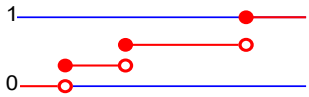
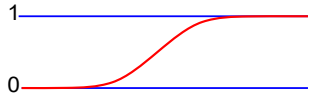
Fig. 5.3. Funcția de distribuție cumulativă pentru o funcție discretă (stânga) și continuă (dreapta) [9]

În cazul variabilei discrete aleatorie (stânga), fiecare probabilitate este asociată cu o anumită valoare și avem goluri între valorile rezultate. Suma probabilităților este 1. În cazul variabilei aleatorii continue (dreapta) avem o curbă continuă începând de la 0 și care se termină la 1.

5.4. Variabile discrete şi continue

Dacă punem cele două tipuri de variabile una lângă alta, putem vedea diferențele dintre ele (Tabel 5.1).

Tabel 5.1. Comparația între variabilele discrete și cele continue

	Variabile aleatorii discrete	**Variabile aleatorii continue**
Probabilitățile pot fi scrise într-un tabel	da	nu
Putem determina probabilitatea unei anumite valori	da	nu
Distribuția probabilităților		
Funcția de distribuție cumulativă		

Variabilele aleatorii discrete și continue pot avea o infinitate de moduri de distribuție. Cu toate acestea, există anumite distribuții care sunt mai frecvent întâlnite. Învățând despre aceste distribuții și știind când să le folosim poate fi de foarte mare ajutor în înțelegerea proceselor și luarea deciziilor.

Distribuțiile pot fi discrete sau continue în funcție de tipul de variabilă pe care îl analizăm.

5.5. Verificarea cunoștințelor

1. O variabilă aleatorie este:
   a. O variabilă care se modifică în timp.
   b. Un rezultat numeric al unui proces aleatoriu.
   c. O valoare numerică deterministă.

2. Care dintre următoarele este o variabilă aleatorie discretă?
   a. Greutatea unui măr ales la întâmplare.
   b. Numărul de elevi dintr-o clasă.
   c. Timpul necesar pentru a alerga un maraton.

3. Care este principala diferență între variabilele aleatorii discrete și cele continue?
   a. Variabilele discrete pot lua numai valori întregi, în timp ce variabilele continue pot lua orice valoare.
   b. Variabilele discrete pot lua orice valoare, în timp ce variabilele continue pot lua numai valori specifice.
   d. Nu există nicio diferență; ambele sunt tipuri de variabile aleatorii.

4. Care dintre următoarele este un exemplu de variabilă aleatorie continuă?
   a. Aruncarea unui zar.
   b. Numărul de mașini dintr-o parcare.
   c. Temperatura într-o cameră la un moment dat.

5. Dacă aruncați o monedă corectă de trei ori, variabila aleatorie care reprezintă numărul de capete este:
   a. O variabilă aleatorie continuă.
   b. O variabilă aleatorie discretă.
   d. Nici discretă, nici continuă.

6. Care dintre următoarele situații poate fi reprezentată printr-o variabilă aleatorie continuă?
   a. Numărul de e-mailuri primite într-o oră.
   b. Înălțimea elevilor dintr-o școală.
   c. Numărul de zile ploioase dintr-un an.

7. O funcție de distribuție a probabilității pentru o variabilă aleatorie discretă:

a. Arată probabilitatea ca variabila să fie continuă pe un interval.
b. Enumeră fiecare dintre rezultatele posibile și probabilitatea asociată fiecăruia dintre ele.
c. Este integrala funcției densității de probabilitate.

8. Suma tuturor probabilităților pentru toate valorile posibile ale unei variabile aleatoare discrete trebuie să fie egală:

a. 1
b. 0
c. 100

9. Un exemplu din lumea reală de variabilă aleatoare discretă este:

a. Cantitatea de lapte dintr-un pahar.
b. Numărul de manuale de pe un raft.
c. Viteza unei mașini în mișcare.

10. Probabilitatea ca o variabilă aleatoare continuă să ia o singură valoare specifică este:

a. Exact 1.
b. Exact 0.
c. Un număr pozitiv mai mare decât 0.

Răspunsuri corecte

1. b. Un rezultat numeric al unui proces aleatoriu.

2. b. Numărul de elevi dintr-o clasă.

3. a. Variabilele discrete pot lua numai valori întregi, în timp ce variabilele continue pot lua orice valoare.

4. c. Temperatura dintr-o cameră la un moment dat.

5. b. Variabilă aleatoare discretă.

6. b. Înălțimea elevilor dintr-o școală.

7. b. Enumeră fiecare dintre rezultatele posibile și probabilitatea asociată fiecăruia dintre ele.

8. a. 1

9. b. Numărul de manuale de pe un raft.

10. b. Exact 0.

6. Distribuții discrete

Funcția dată de distribuția probabilităților ne poate ajuta să obținem informații valoroase despre datele noastre și să ne ghideze în procesul decizional. Există câteva distribuții care sunt mai frecvent întâlnite în viața de zi cu zi. Acesta este motivul pentru care este bine să știm câteva lucruri de bază despre ele.

Câteva dintre cele mai des întâlnite distribuții de variabile discrete pe care le vom discuta în continuare sunt:

- Distribuția uniformă
- Distribuția binomială
- Distribuția hipergeometrică

6.1. Distribuția uniformă

Distribuția uniformă are aceeași probabilitate pentru toate rezultatele (Figura 6.1). Fiecare rezultat are probabilitatea de 1/n în cazul în care n este numărul de rezultate.

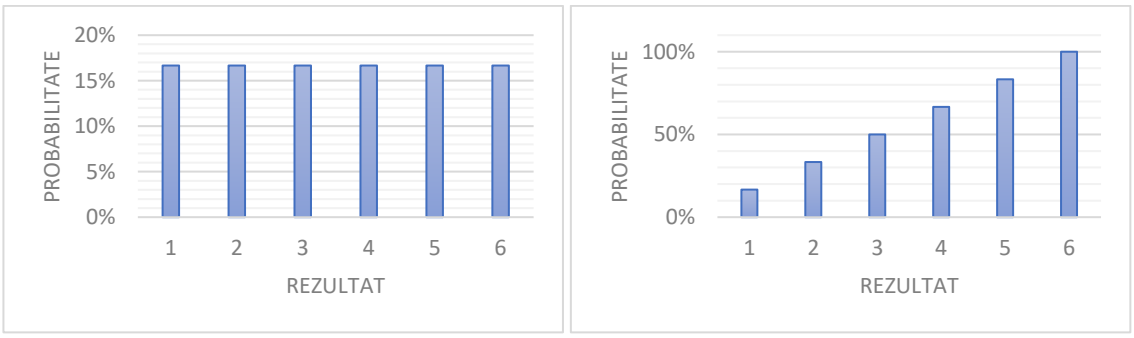


Fig. 6.1. Funcția de probabilitate și funcția cumulativă de distribuție pentru distribuția uniformă.

Un exemplu de distribuție uniformă este aruncarea unui zar. Rezultatele posibile sunt: 1, 2, 3, 4, 5, 6. Fiecare rezultat are o probabilitate de 1/6 de a se întâmpla în cazul în care zarul este unul corect (nemodificat încât să fie părtinitor).

Funcția cumulativă de distribuție pentru distribuția uniformă crește în mod constant de la o categorie la alta.

6.2. Distribuția binomială

Distribuția binomială este utilizată de fiecare dată când avem doar două rezultate posibile cu încercări repetate, iar observațiile sunt independente. Ne concentrăm doar pe un singur rezultat, numit *succes*. De exemplu, să zicem că avem o monedă cu două fețe: cap și pajură. Prin aruncarea de mai multe ori a monedei avem un experiment binomial. Ne uităm doar la un singur rezultat, să zicem că observăm numărul de capete care apar. Acesta este rezultatul „succes”.

Notații pe care o să le folosim

- p – probabilitatea de succes la o încercare
- q – probabilitatea de eșec la o încercare
- n – numărul de încercări
- k – numărul de succese la un moment dat

Se numește distribuția binomială, deoarece probabilitatea urmărește formula binomială a lui Newton:

$$(a+b)^n = C_n^0 a^n b^0 + C_n^1 a^{n-1} b^1 + \ldots + C_n^n a^0 b^n$$

De exemplu, dacă dintr-o urnă care conține două tipuri de bile (roșii și albastre) extragem de 2 ori câte o bilă, **punând de fiecare dată bila la loc**, variabila aleatorie pentru numărul de bile roșii extrase este:

$$X: \begin{pmatrix} 0 & 1 & 2 \\ q^2 & 2pq & p^2 \end{pmatrix}$$

Probabilitatea pentru fiecare rezultat este fiecare termen al descompunerii lui $(q+p)^2$.

În general, probabilitatea unui anumit rezultat k (funcția de distribuție) poate fi calculată cu formula:

$$P(X=k) = C_n^k p^k q^{n-k}$$

unde C_n^k este combinații de n luate câte k:

$$C_n^k = \frac{n!}{k!\,(n-k)!}$$

Funcția de probabilitate cumulativă are următoarea formulă:

$$P(X \le k) = \sum C_n^k p^k q^{n-k}$$

Graficele funcțiilor de distribuție și de probabilitate sunt afișate în figura de mai jos.

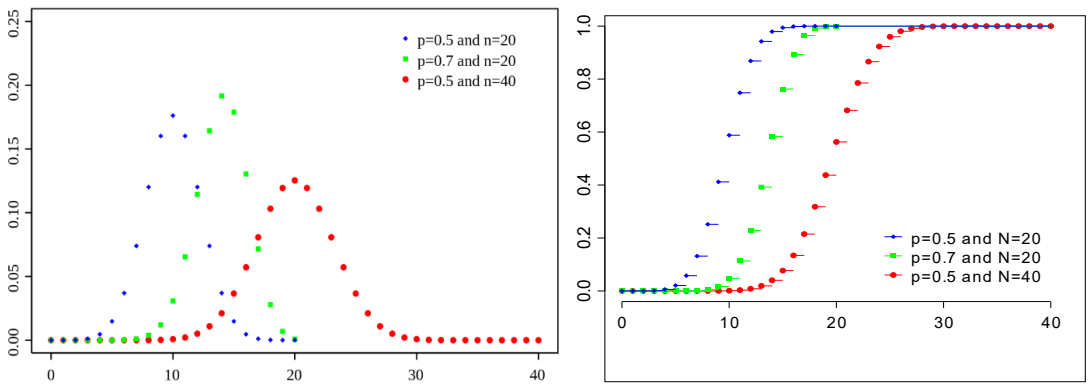


Fig. 6.2. Funcțiile de distribuție (stânga) și de probabilitate (dreapta) ale distribuției binomială [10]

Un exemplu des întâlnit în procesul de fabricație este atunci când se lucrează cu piese. Dacă împărțiți piesele în două categorii, rele și bune, atunci putem avea:

p – probabilitatea de a scoate o piesă rea

q - probabilitatea de a lua o bună piesă

Distribuția binomială este adesea folosită pentru a modela situațiile în care avem un număr fix de încercări independente, fiecare cu doar două rezultate posibile și o probabilitate constantă de succes pentru fiecare încercare. Un caz de utilizare important al distribuției binomiale este în contextul eșantionării de acceptare, care este o procedură de control al calității utilizată pentru a determina dacă un lot de piese îndeplinește anumite standarde.

În eșantionarea de acceptare, se selectează în mod aleatoriu un eșantion de piese dintr-un lot mai mare, iar fiecare piesă din eșantion este inspectată pentru defecte sau alte probleme de calitate. Condiția de acceptare se bazează apoi pe numărul de piese defecte din eșantion, care este modelat folosind distribuția binomială.

Condiția de acceptare este exprimată în mod obișnuit în funcție de doi parametri: dimensiunea eșantionului (n) și numărul maxim permis de defecte (c). Eșantionul este acceptat dacă numărul de piese defecte (X) din eșantion este mai mic sau egal cu numărul maxim permis de defecte (c). Matematic, acest lucru poate fi exprimat ca:

$$P(X \leq c) \geq A$$

unde $P(X \leq c)$ este probabilitatea cumulată de a avea cel mult *c* piese defecte în eșantion, iar *A* este nivelul de acceptare specificat, care este de obicei o valoare mică de probabilitate (de exemplu, 0.05 sau 0.01).

Cu alte cuvinte, lotul de piese este acceptat dacă probabilitatea de a observa **c** sau mai puține piese defecte în eșantion este mai mare sau egală cu nivelul de acceptare specificat. Dacă probabilitatea este mai mică decât nivelul de acceptare, atunci lotul este respins și poate fi necesară o inspecție sau o acțiune corectivă suplimentară.

Exemplu de problemă

Vrem să cumpărăm un lot de piese de la un furnizor. El declară coeficientul său de rebut (procentul de piese rele) este **p = 5%.** Extragem n = 3 piese, **de fiecare dată punând piesa la loc**. Care este probabilitatea de a obține cel mult o piesă defectă?

Cele 4 rezultate posibile sunt: 0, 1, 2, sau 3 piese rele.

Utilizând formula funcției de distribuție dată mai sus, vom obține următoarea variabilă aleatorie:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ C_3^0 q^3 p^0 & C_3^1 q^2 p^1 & C_3^2 q^1 p^2 & C_3^3 q^0 p^3 \end{pmatrix}$$

După efectuarea calculelor vom obține probabilitățile pentru fiecare rezultat posibil:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ 85.74\% & 13.54\% & 0.71\% & 0.01\% \end{pmatrix}$$

Funcția de distribuție cumulativă este:

$$X: \begin{pmatrix} 0 & \mathbf{1} & 2 & 3 \\ 85.74\% & \mathbf{99.28}\% & 99.99\% & 100\% \end{pmatrix}$$

Acest lucru înseamnă că probabilitatea de a obține cel mult 0 piese rele este de 85.74%, probabilitatea de a obține cel mult 1 piesă rea este de 99.28% și așa mai departe. Răspunsul corect în acest caz este 99.28%.

6.3. Distribuţia hipergeometrică

Distribuția hipergeometrică este foarte similară cu distribuția binomială, în sensul că avem doar două rezultate posibile și mai multe încercări. Diferența este că evenimentele depind unul de celălalt.

Notații:

- p – probabilitatea de succes la o încercare
- q – probabilitatea de eșec la o încercare
- n – numărul de obiecte
- m – numărul de încercări
- a – numărul de succese
- b – numărul de eșecuri
- k – numărul de succese la un anumit punct

Funcția de distribuție are următoarea formulă:

$$P(X = k) = \frac{C_a^k C_b^{m-k}}{C_n^m}$$

Funcția de distribuție cumulată are următoarea formulă:

$$P(X \leq k) = \frac{1}{C_n^m} \sum C_a^k C_b^{m-k}$$

$$P(X \leq k) = F(k) = \frac{1}{C_n^m} \sum_{k=0}^{k} C_a^k C_b^{m-k}$$

Graficele celor două funcții, de distribuție și cumulată, sunt afișate în figura 6.3.

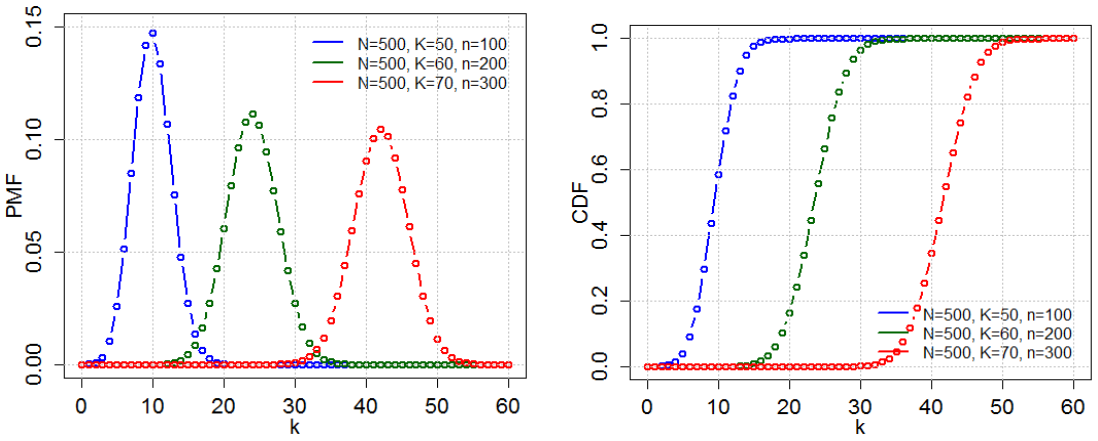


Fig. 6.3. Funcția de distribuție și funcția de distribuție cumulată pentru distribuția hipergeometrică [11]

Exemplu de problemă

Să presupunem că doriți să cumpărați un lot de n = 100 piese de la un furnizor. El spune că coeficientul său de rebut (procentul de piese rele) este p = 5%. Extragem m = 3 piese, **fără a pune piesa înapoi**. Determinați probabilitatea de a obține cel mult o piesă rea.

Cele 4 rezultate posibile sunt: 0, 1, 2, sau 3 piese rele.

Numărul de piese rele (a) este:

$$a = n * p = 100 * 5\% = 5$$

și numărul de piese bune (b):

$$b = n * q = n - a = 100 * 95\% = 95$$

Utilizând formula funcției de distribuție dată mai sus, vom obține următoarea variabilă aleatorie:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{C_5^0 * C_{95}^3}{C_{100}^3} & \frac{C_5^1 * C_{95}^2}{C_{100}^3} & \frac{C_5^2 * C_{95}^1}{C_{100}^3} & \frac{C_5^3 * C_{95}^0}{C_{100}^3} \end{pmatrix}$$

După efectuarea calculelor vom obține probabilitățile pentru fiecare rezultat posibil:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ 85.60\% & 13.81\% & 0.59\% & 0.01\% \end{pmatrix}$$

Funcția de probabilitate este:

$$X: \begin{pmatrix} 0 & \mathbf{1} & 2 & 3 \\ 85.60\% & \mathbf{99.41\%} & 99.99\% & 100\% \end{pmatrix}$$

Acest lucru înseamnă că probabilitatea de a obține cel mult 0 piese rele este de 85.60%, probabilitatea de a obține cel mult 1 piesă rea este 99.41% și așa mai departe. Deci răspunsul corect este 99.41%.

6.4. Verificarea cunoștințelor

1. Ce este distribuția uniformă discretă?
   a. O distribuție de probabilitate în care fiecare eveniment are aceeași probabilitate de a se întâmpla
   b. O distribuție de probabilitate în care evenimentele sunt independente
   c. O distribuție de probabilitate care modelează probabilitatea unui succes într-un număr finit de încercări independente
2. Ce este distribuția binomială?
   a. O distribuție de probabilitate care modelează numărul de succese într-un proces continuu
   b. O distribuție de probabilitate care modelează probabilitatea unui succes într-un număr finit de încercări independente
   c. O distribuție de probabilitate care modelează probabilitatea de a avea un anumit număr de succese într-un număr fix de încercări independente
3. Ce este distribuția hipergeometrică?
   a. O distribuție de probabilitate care modelează probabilitatea de a avea un anumit număr de succese într-un număr fix de încercări independente
   b. O distribuție de probabilitate care modelează probabilitatea de a avea un anumit număr de succese într-un eșantion extras dintr-o populație finită și specificată.
   c. O distribuție de probabilitate care modelează numărul de succese într-un proces continuu
4. Ce este funcția de probabilitate a distribuției uniforme?
   a. f(x) = 1/n, unde n este numărul de evenimente posibile
   b. f(x) = x/n, unde n este numărul de evenimente posibile
   c. f(x) = n, unde n este numărul de evenimente posibile
5. Care este diferența între distribuția binomială și cea hipergeometrică?
   a. Distribuția binomială modelează probabilitatea de a avea un anumit număr de succese într-un număr fix de încercări independente, în timp ce distribuția hipergeometrică modelează probabilitatea de a avea un anumit număr de succese într-un eșantion extras dintr-o populație finită și specificată.
   b. Distribuția binomială modelează probabilitatea de a avea un anumit număr de succese într-un eșantion extras dintr-o populație finită și specificată, în timp ce distribuția hipergeometrică modelează probabilitatea de a avea un anumit număr de succese într-un număr fix de încercări independente.
   c. Distribuțiile binomială și hipergeometrică sunt la fel și modelează aceleași scenarii.

Răspunsuri corecte:

1. a) O distribuție de probabilitate în care fiecare eveniment are aceeași probabilitate de a se întâmpla

2. c) O distribuție de probabilitate care modelează probabilitatea de a avea un anumit număr de succese într-un număr fix de încercări independente

3. b) O distribuție de probabilitate care modelează probabilitatea de a avea un anumit număr de succese într-un eșantion extras dintr-o populație finită și specificată.

4. a) f(x) = 1/n, unde n este numărul de evenimente posibile

5. a) Distribuția binomială modelează probabilitatea de a avea un anumit număr de succese într-un număr fix de încercări independente, în timp ce distribuția hipergeometrică modelează probabilitatea de a avea un anumit număr de succese într-un eșantion extras dintr-o populație finită și specificată.

7. Distribuții continue

Distribuțiile continue descriu distribuirea probabilităților unei variabile aleatorii continue, care poate lua orice valoare dintr-un domeniu continuu de numere reale. Aceste distribuții sunt diferite de distribuțiile discrete, care se aplică variabilelor aleatorii discrete. Printre cele mai cunoscute distribuții continue se numără:

- distribuția Uniformă (continuă)
- distribuția Normală
- distribuția Student
- distribuția Chi-pătrat

Distribuția uniformă se referă la variabilele aleatorii care au o distribuție uniformă a probabilității între două limite, în timp ce distribuția normală, cunoscută și sub numele de distribuție Gauss, este una dintre cele mai importante distribuții continue, fiind utilizată într-o gamă largă de domenii, de la științele sociale până la finanțe și inginerie. Distribuția Student este utilizată în principal în analiza statistică când avem eșantioane de mici dimensiuni, în timp ce distribuția Chi-pătrat este folosită în multe domenii, precum în testarea ipotezelor statistice sau analiza datelor experimentale.

7.1. Distribuţia uniformă

Distribuția uniformă continuă este una dintre cele mai simple distribuții continue, dar totuși foarte importante, fiind utilizată în multe domenii, de la științele sociale până la științele naturii și tehnologie. Această distribuție descrie variabile aleatorii care au o distribuție uniformă a probabilității între două limite, așadar, fiecare valoare din interval are aceeași probabilitate de a fi aleasă. Aceasta înseamnă că probabilitatea de a obține o anumită valoare din interval este proporțională cu lungimea intervalului și nu depinde de nicio altă caracteristică a distribuției.

Distribuția uniformă are aceeași probabilitate pentru intervale diferite între niște valori limită (*a* și *b*). În afara lui *a* și *b* probabilitatea este 0. Funcția de distribuție și cea cumulativă sunt prezentate în figura 7.1. și au următoarele formule:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & for\ x < a\ \ or\ x > b \end{cases}$$

$$F(x) = \begin{cases} 0 & for\ x < a \\ \dfrac{x-a}{b-a} & for\ a \le x \le b \\ 1 & for\ x > b \end{cases}$$

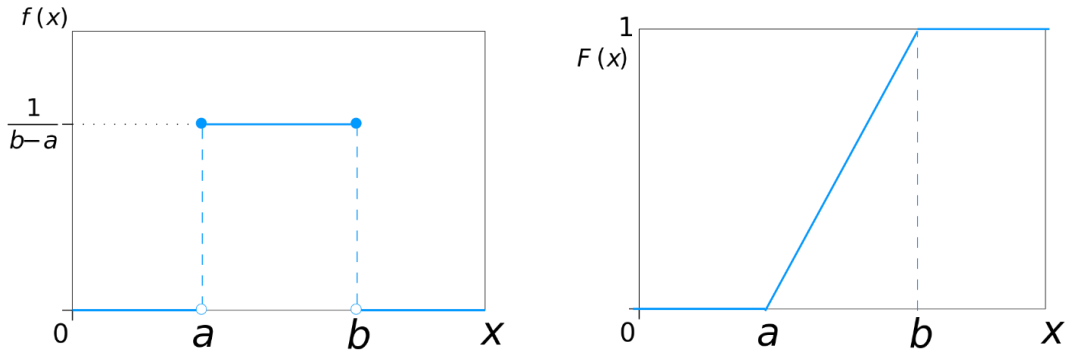


Fig. 7.1 Funcția de distribuție (stânga) și cumulativă (dreapta) pentru distribuția uniformă continuă [12]

7.2. Distribuţia normală

Distribuția normală, cunoscută și sub numele de distribuția Gauss, este una dintre cele mai importante și mai frecvent utilizate distribuții continue. Ea se caracterizează prin curba simetrică în jurul valorii sale medii, care reprezintă centrul distribuției, și prin abaterea standard, care măsoară cât de multă variabilitate există în datele distribuției. Această distribuție este utilizată într-o gamă largă de domenii, de la științele sociale și economice până la științele naturii și tehnologie. Ea este folosită pentru a modela datele care sunt aproximativ simetrice și pentru a face estimări de probabilitate sau inferență statistică. Deoarece multe fenomene naturale și sociale urmează această distribuție, distribuția normală are o importanță fundamentală în analiza datelor și în luarea deciziilor bazate pe date într-o varietate de domenii.

Funcția de distribuție are formula:

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

unde σ^2este dispersia populației și μ media populație.

Funcția cumulativă este descrisă de ecuația:

$$F(x) = \int_{-\infty}^{x} \frac{e^{-\frac{-x^2}{2}}}{\sqrt{2\pi}}$$

Funcția de distribuție are o formă de clopot și de aceea distribuția normală este denumită și distribuția clopot. Funcția de distribuție și cea cumulativă arată ca în figura 7.2.

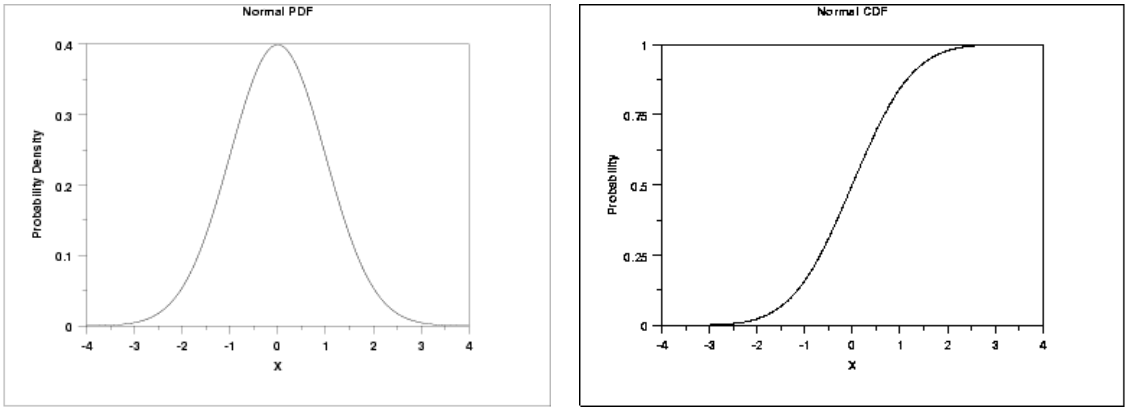


Fig. 7.2. Funcția de distribuție (stânga) și de probabilitate (dreapta) pentru distribuția normală [13]

Distribuția normală are doi parametri importanți: dispersia (σ^2) și media (μ). Prin modificarea mediei, curba se deplasează la stânga sau la dreapta pe axa orizontală. Prin modificarea dispersiei curba devine fie mai subțire și mai înaltă sau mai lată și mai aplatizată (Figura 7.3.). Puteți vedea ce efect au acești doi parametri asupra formei curbei pe mai multe site-uri web care au o curbă de distribuție normală interactivă, cum ar fi aceasta: http://onlinestatbook.com/2/Calculators/normal_dist.html

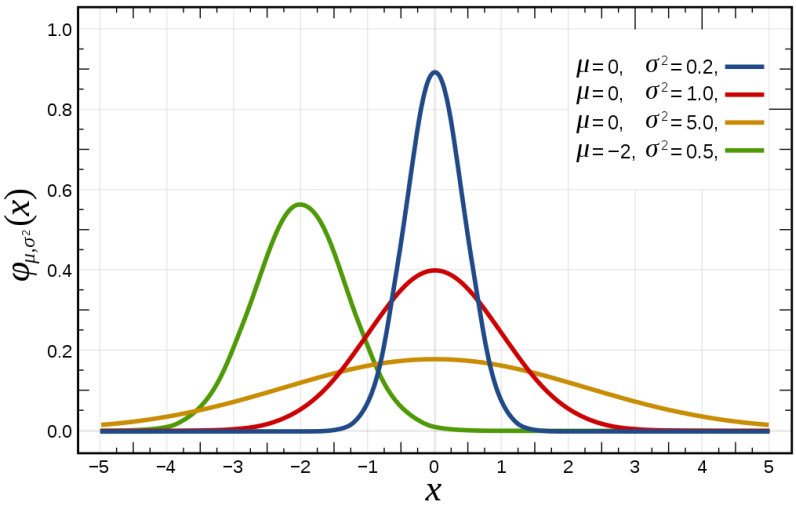


Fig. 7.3 Exemple de distribuții normale cu diferiți parametri [14]

Distribuția este simetrică față de medie și cozile merg asimptotic spre (plus și minus) infinit, fără a atinge vreodată axa. Ca și cu orice alte distribuții continue de probabilitate, aria de sub curbă este egală 1.

Există o distribuție normală mai specială, care este numită Distribuția Normală Standard. Ea are o medie de 0 și o abatere standard de 1. Este foarte utilă în determinarea poziției anumitor valori pe orice altă distribuție normală.

Aria de sub curba normală respectă regula 68 -95-99.7, unde aproximativ 68% din arie este în intervalul de o abatere standard față de medie (una de fiecare parte), aproximativ 95% din arie este la două abateri standard față de medie (de fiecare parte) și aproximativ 99.7% din arie la trei abateri de la medie.

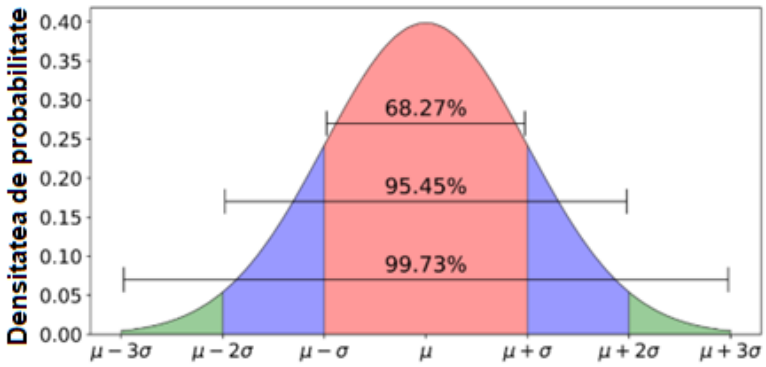


Fig. 7.4. Regula 68 -95-99.7 [15]

7.3. Distribuția Student

Distribuția Student este o distribuție de probabilitate continuă utilizată în special pentru a estima intervalul de încredere al unei medii a populației, atunci când dispersia acesteia este necunoscută și este estimată prin abaterea standard a eșantionului. Această distribuție este numită după William Gosset, care a lucrat la fabrica de bere Guinness și a dezvoltat această distribuție pentru a analiza calitatea berii. Deoarece el lucra la o fabrică de bere și nu era permis să-și publice munca sub numele său real, a folosit pseudonimul "Student". În prezent, distribuția Student este utilizată în multe domenii, precum în cercetarea de piață, medicină, științe sociale și inginerie. Este important să înțelegem această distribuție pentru a putea interpreta și analiza datele colectate prin intermediul experimentelor și sondajelor.

Distribuția este, de asemenea, simetrică față de medie, dar are cozi mai groase, ceea ce înseamnă că valorile care sunt mai departe de medie au probabilități mai mari decât echivalentul lor pe distribuția normală (Figura 7.5).

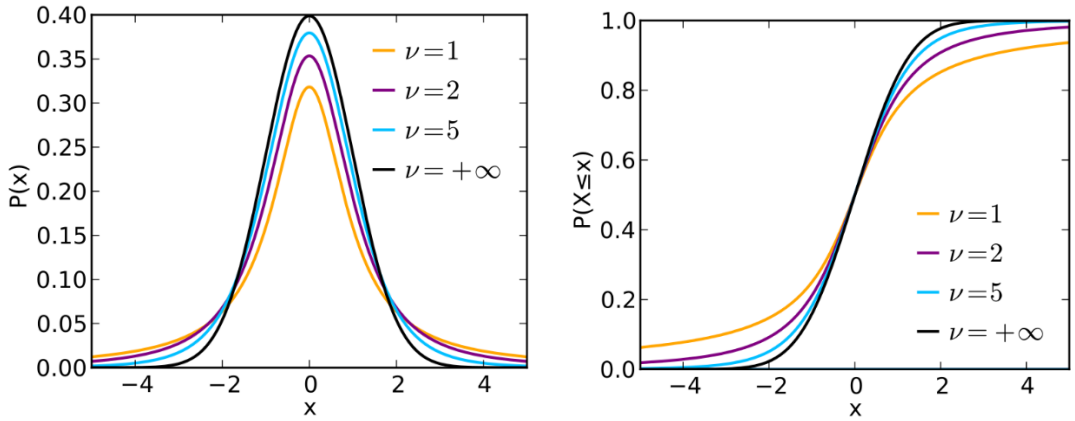


Fig. 7.5. Funcția de distribuție (stânga) și cumulativă (dreapta) pentru distribuția Student [16]

Un parametru important este numărul de grade de libertate (υ):

$$\upsilon = n - 1$$

unde *n* este numărul de valori din eșantion.

În figura 7.5 este vizibil modul în care gradele de libertate afectează forma distribuției. Pentru un număr infinit de grade de libertate, distribuția Student devine distribuția normală.

7.4. Distribuția Chi-pătrat

Distribuția Chi-pătrat (χ^2) este o distribuție de probabilitate continuă care este utilizată în domenii precum testarea ipotezelor statistice, analiza datelor experimentale sau în construirea intervalului de încredere pentru dispersia unei populații. Această distribuție derivă din distribuția normală standard și este definită de numărul de grade de libertate. Distribuția Chi-pătrat se caracterizează prin dispersia sa, fiind influențată de numărul de grade de libertate. În general, cu cât numărul de grade de libertate este mai mare, cu atât distribuția se va apropia mai mult de o distribuție normală. Prin urmare, distribuția Chi-pătrat este o distribuție importantă în multe aplicații statistice.

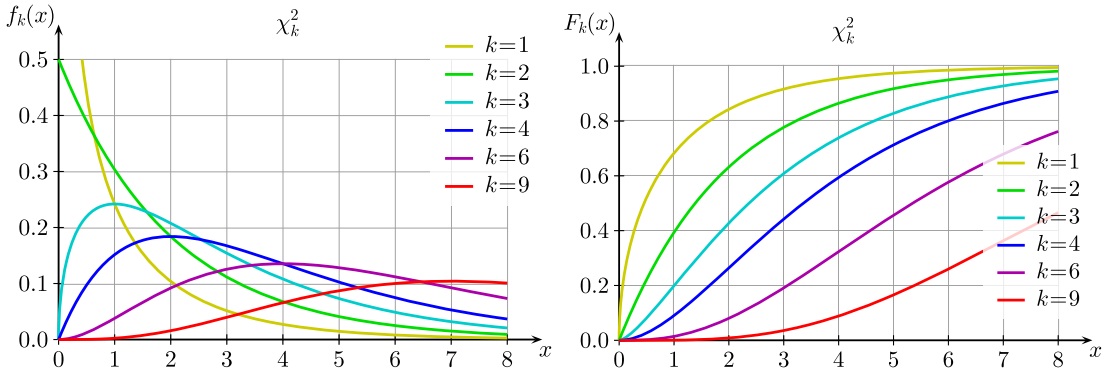


Fig. 7.6. Funcția de distribuție (stânga) și cumulativă (dreapta) pentru distribuția Chi-pătrat [17]

Distribuția Chi-pătrat este diferită de celelalte două distribuții continue discutate anterior prin faptul că nu este simetrică și nu are valori negative (Figura 7.6).

7.5. Verificarea cunoștințelor

1. Distribuția uniformă atribuie:
   a. Probabilități diferite tuturor rezultatelor.
   b. Aceeași probabilitate pentru toate rezultatele.
   c. Probabilități crescânde pentru rezultate consecutive.

2. Distribuția normală se caracterizează prin:

   a. O curbă cu singur vârf pe medie și simetrie în jurul mediei.
   b. O curbă cu două vârfuri la extreme.
   c. Probabilități distribuite uniform.

3. Care dintre următoarele afirmații este adevărată pentru distribuția normală?

   a. Este înclinată spre dreapta.
   b. Este simetrică în jurul mediei sale.
   c. Are o înălțime uniformă.

4. Distribuția Student se utilizează atunci când:

   a. Dimensiunea eșantionului este mare și dispersia populației este cunoscută.
   b. Mărimea eșantionului este mică și dispersia populației este necunoscută.
   c. Populația urmează o distribuție uniformă.

5. Pe măsură ce gradele de libertate cresc pentru distribuția Student, distribuția:

   a. Devine mai largă și mai plană.
   b. Se apropie de o distribuție normală.
   c. Rămâne constantă, indiferent de gradele de libertate.

6. O distribuție Chi-pătrat nu este simetrică, ci are o coadă mai lungă spre:

   a. Stânga.
   b. Dreapta.
   c. Atât la stânga, cât și la dreapta, în funcție de situație.

7. Distribuția Chi-pătrat este utilizată în principal pentru:

   a. Estimarea parametrilor altor distribuții.
   b. Testarea ipotezelor privind dispersia populației.
   c. Testarea ipotezelor privind mediile populației.

8. Distribuția normală este complet descrisă de:

a. Medie și mediană.
b. Medie și dispersie.
c. Dispersie și asimetrie.

9. Distribuția continuă uniformă este adesea utilizată pentru a modela:

a. Variabile cu un interval cunoscut și probabilitate egală pentru orice valoare din interval.
b. Variabile cu o asimetrie naturală.
c. Variabile care se grupează în jurul unei valori centrale.

10. Forma distribuției Chi pătrat depinde de:

a. Media eșantionului.
b. Numărul de încercări.
c. Gradele de libertate.

Răspunsuri corecte

1. b. Aceeași probabilitate pentru toate rezultatele.

2. a. Un singur vârf la medie și simetrie în jurul mediei.

3. b. Este simetrică în jurul mediei sale.

4. b. Mărimea eșantionului este mică și varianța populației este necunoscută.

5. b. Se apropie de o distribuție normală.

6. b. Dreapta.

7. b. Testează ipoteze despre dispersia populației.

8. b. Medie și dispersie.

9. a. Variabile cu un interval cunoscut și probabilitate egală pentru orice valoare din interval.

10. c. Gradele de libertate.

8. Estimarea

În domeniul statisticii sunt des folosite conceptele de *populație* și *eșantion*. O ***populație*** este un set complet de elemente sau evenimente care sunt relevante pentru o anumită cercetare sau analiză. De exemplu, dacă ne interesează să studiem înălțimea adulților din România, populația ar consta din toți adulții din țară.

De obicei, populațiile sunt foarte mari, cu mii și milioane de elemente și de aceea, nu putem colecta date de la întreaga populație din cauza costurilor și efortului necesar. Prin urmare, de obicei, se lucrează cu un subset al populației numit ***eșantion***. Eșantionul trebuie să fie reprezentativ pentru populație, astfel încât să putem face inferențe statistice.

Există diferite metode de eșantionare care pot fi utilizate pentru a colecta un eșantion reprezentativ. Două metode comune sunt:

- **Eșantionare Aleatorie**: Acesta este cel mai simplu și cel mai direct mod de a colecta un eșantion. Fiecare membru al populației are o șansă egală de a fi selectat în eșantion elementele fiind alese la întâmplare.
- **Eșantionare Stratificată**: Populația este împărțită în mai multe subgrupuri sau "straturi", iar apoi se extrage un eșantion aleatoriu din fiecare strat. Această metodă este utilă atunci când populația este heterogenă și dorim să ne asigurăm că eșantionul reflectă această diversitate.

Există diferențe de caracteristici între populație și eșantion, dar, uneori, singura modalitate de a studia o populație este de a studia unul sau mai multe eșantioane luate dintr-o populație.

Odată obținut un eșantion, putem face o ***estimare punctuală***, adică să determinăm un singur număr care servește drept estimare "cea mai bună" pentru un parametru al populației, cum ar fi media. De exemplu, putem calcula media înălțimii în eșantionul nostru și să folosim acest număr ca o estimare punctuală a mediei înălțimii în întreaga populație.

Cu toate acestea, orice estimare punctuală vine cu un anumit grad de incertitudine. Putem cuantifica această incertitudine folosind în schimb o ***estimare cu intervale de încredere***. În acest caz determinăm un interval de valori în care parametrul populației este probabil să se afle. De exemplu, putem spune că suntem 95% siguri că media înălțimii adulților din România este între 165 cm și 175 cm. În acest exemplu, intervalul [165, 175] este intervalul de încredere iar 95% este nivelul de încredere.

Deoarece estimarea cu un interval de încredere este mai frecventă, ne vom concentra pe ea în continuare. În acest curs vom aborda estimarea a doi parametri importanți ai populației, *media* și *dispersia* (și implicit abaterea standard). În Tabelul 8.1.

sunt notații pe care le-am folosit de-a lungul acestui curs și pe care le vom folosi pentru a distinge între parametrii populației și statisticile eșantionului.

Tabel 8.1. Notații folosite pentru parametrii populației și eșantionului

Parametru	Populației	Eșantion
Media	μ	$\bar{x}$
Abaterea standard	σ	s
Dispersia	σ^2	s^2

Dintr-o populație putem extrage mai multe eșantioane (Figura 8.1). Datorită tendinței de grupare centrale știm că cele mai multe valori ale populației se află în vecinătatea mediei populației (μ) și atunci eșantioanele vor avea de asemenea valori, și implicit medii ($\bar{x}_1 - \bar{x}_4$), în vecinătatea mediei populației. Bineînțeles că există și probabilitatea de a obține eșantioane cu valori foarte mari sau valori foarte mici, dar acestea sunt mai rare deoarece valorile extreme în sine sunt mai rare.

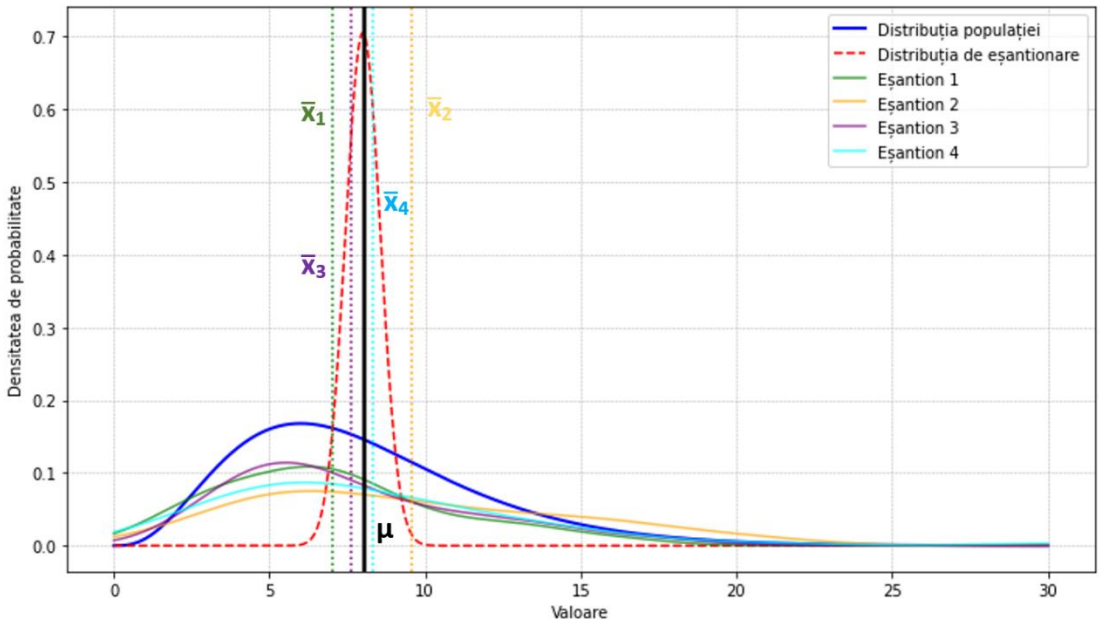


Fig. 8.1. Exemplu de distribuție a unei populații și a mai multor eșantioane

Dacă luăm un număr foarte mare de eșantioane, și trasăm distribuția mediilor acestora, vom obține o distribuție normală numită ***distribuția de eșantionare***. Această distribuției are proprietatea că media ei va fi identică cu media populației din care am făcut eșantionarea, indiferent de forma distribuției populației.

Când estimăm parametrul unei populații, de fapt determinăm niște limite (una sau două) ale unui interval în care presupunem cu o anumită probabilitate că se află parametrul estimat. Intervalul se numește ***interval de încredere*** (Figura 8.2). Acesta

poate avea una sau două limite, cea din stânga fiind numită ***limita inferioară*** iar cea din dreapta ***limita superioară*** a intervalului de încredere. Probabilitatea ca parametrul pe care îl estimăm să se afle între aceste limite se numește ***nivelul de încredere*** iar ce se află înafara limitelor se numește ***risc***. Vom nota riscul cu litera grecească α. Deoarece aria de sub curba unei distribuții reprezintă o probabilitate, ea va fi egală cu 1. Drept urmare, nivelul de încredere va fi toată probabilitatea din care eliminăm riscul $(1 - \alpha)$.

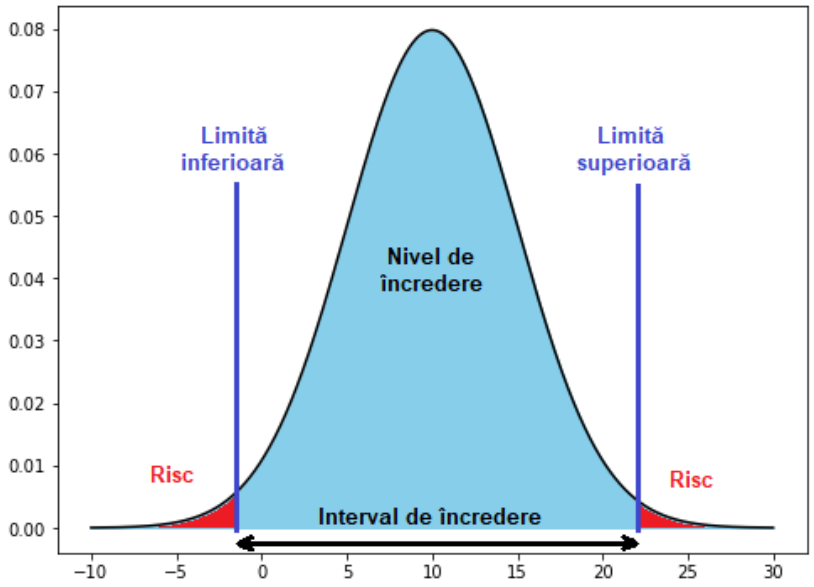
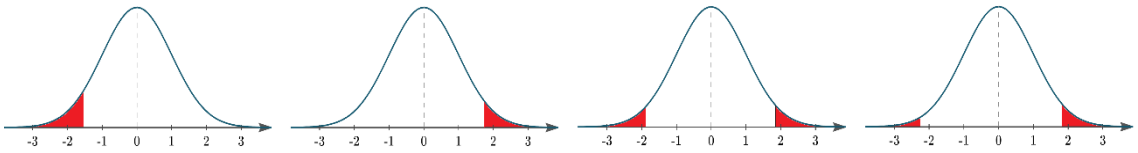


Fig. 8.2. Estimarea cu interval de încredere (risc bilateral)

Deoarece putem avea una sau două limite pentru intervalul de încredere, putem avea următoarele tipuri de risc (Figura 8.3):

- Riscul Unilateral Dreapta (RUD)
- Riscul Unilateral Stânga (RUS)
- Riscul Bilateral Simetric (RBS)
- Risc Bilateral Asimetric (RBA)



a) *b)* *c)* *d)*

Fig. 8.3. Tipuri de risc: a) Risc Unilateral Stânga; b) Risc Unilateral Dreapta; c) Risc Bilateral Simetric; d) Risc Bilateral Asimetric

Odată definiți acești termeni, putem să trecem la estimarea parametrilor populației: media și dispersia. Un element important în estimarea mediei este dacă dispersia populației este cunoscută sau nu. Drept urmare vom considera aceste două cazuri când vorbim de estimarea mediei.

8.1. Estimarea medie (dispersia populației cunoscută)

Atunci când dispersia populației este cunoscută, estimarea mediei populației devine oarecum simplă. În astfel de cazuri, una dintre cele mai des folosite tehnici este utilizarea distribuției Normale (z) pentru construirea intervalelor de încredere și testarea ipotezelor.

Formula de calcul a intervalului de încredere atunci când se cunoaște dispersia populației este următoarea:

$$CI = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

CI – limitele intervalului de încredere

z – este scorul z, care poate fi căutat în tabelul de distribuție normală standard și corespunde nivelului de încredere dorit (de obicei 90%, 95% sau 99%).

σ – este abaterea standard a populației (considerată cunoscută în acest caz)

n – este dimensiunea eșantionului.

Haideți să vedem cum se aplică pe un exemplu concret. O companie produce câteva mii de piese în fiecare zi. Ne interesează, să zicem, diametrul exterior mediu al acestor piese. Deoarece măsurarea tuturor pieselor nu ar fi fezabilă, un angajat ia un eșantion de $n = 100$ piese. El calculează apoi media eșantionului și obține $\bar{x} = 10.25$. Deoarece procesul de producție este cunoscut, abaterea standard a populației (și implicit și dispersia) pentru acest proces este cunoscută $\sigma = 0.1$. Să spunem că vrem să estimăm media cu un nivel de încredere de 95%, risc bilateral simetric. Riscul este de $\alpha = 5\%$.

Dacă cele două limite ale intervalului de încredere sunt x_1 și x_2, atunci putem spune că media (μ) se află între x_1 și x_2 cu o probabilitate de 95% și putem scrie:

$$P(x_1 < \mu < x_2) = 0.95$$

De asemenea, știm că riscul este de $\alpha = 5\%$ și este bilateral simetric, vom avea jumătate din risc ($\alpha/2 = 2.5\%$) sub limita inferioară a intervalului de încredere și cealaltă jumătate, deasupra limitei superioare ($\alpha/2 = 2.5\%$).

Pentru a găsi x_1 și x_2, va trebui să folosim o instanță specială a distribuției normale denumită **Distribuția Normală Standard** (Figura 8.4). Această distribuție are valoarea medie 0 și abaterea standard egală cu 1. Valorile de pe axă sunt numite **scoruri z**.

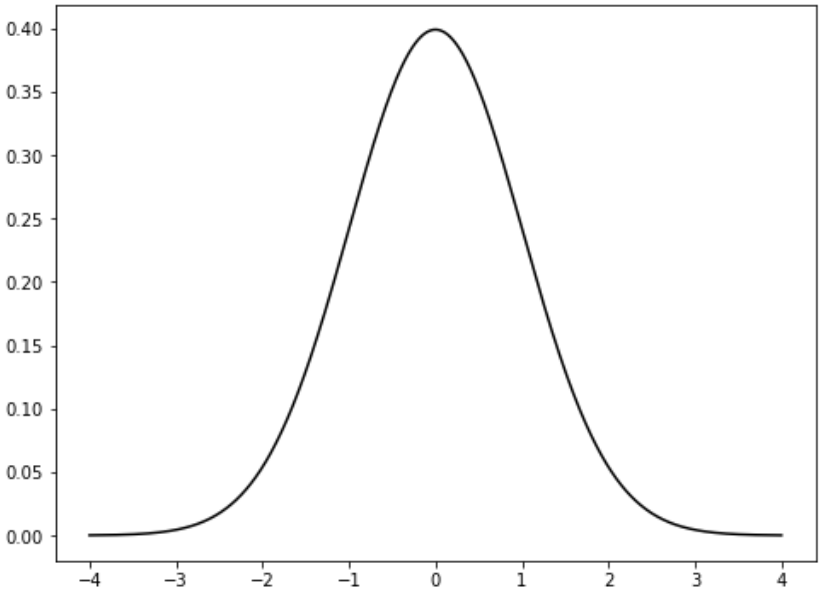


Fig. 8.4. Distribuția Normală Standard

Scorurile z care împart Distribuția Normală Standard în aceleași proporții ca în cazul nostru (2.5% - 95% - 2.5%) sunt echivalentele valorilor x_1 și x_2 pe distribuția eșantionului nostru.

Scorurile z sunt citite dintr-un tabel. Tabelul conține valorile z „rupte” în două părți: prima parte până la prima zecimală se găsește pe prima coloană din tabel, iar a doua parte corespunzătoare celei de-a doua zecimale se găsește în primul rând din tabel (Figura 8.5). Ariile corespunzătoare probabilităților asociate fiecărei valori z se află în corpul tabelului.

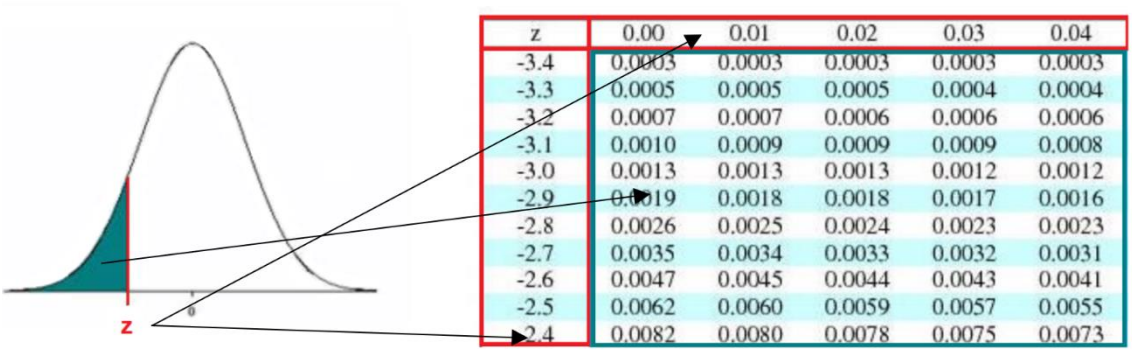
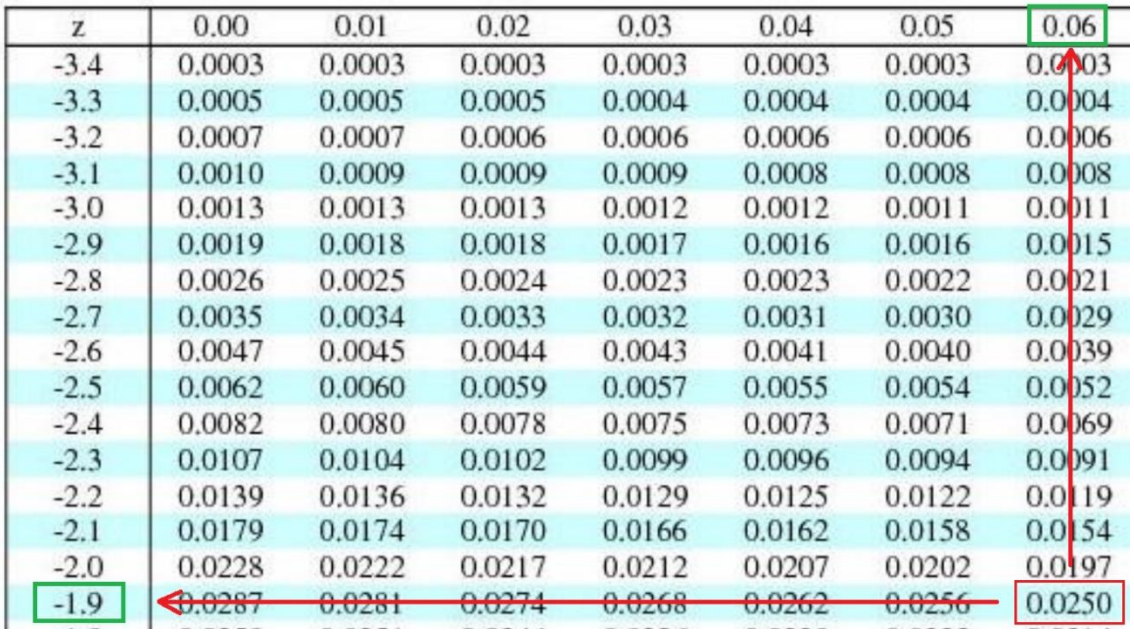


Fig. 8.5. Tabelul scorurilor z

Pentru a citi din tabelul Z trebuie mai întâi să găsim în corpul tabelului cea mai apropiată valoare de probabilitatea pe care o căutăm. Mergem orizontal pe rândul valorii găsite și identificăm prima parte din valoarea lui z de pe prima coloană. Apoi mergem de la valoarea ariei găsite în sus până ajungem pe primul rând al tabelului și citim a doua parte a valorii lui z. De exemplu, în Figura 8.6, căutăm valoarea 0.0250 corespunzătoare riscului de 2.5% la stânga. Odată găsită mergem orizontal pe rând în stânga și găsim valoarea -1.9 corespunzătoare primei părți a lui z. De la valoarea 0.0250

mergem apoi pe coloană în sus și găsim valoarea 0.06 corespunzătoare celei de-a doua părți a valorii lui z. Combinând prima parte (-1.9) cu cea de-a doua parte (0.06) obținem valoarea lui z =-1.96.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250

Fig. 8.6. Citirea scorurilor z din tabel

În Anexa 1 găsiți tabelul complet. Veți observa că sunt de fapt două tabele: unul pentru riscul din stânga, unde z ia valori negative și unul pentru riscul din dreapta unde z ia valori pozitive. Un lucru de care trebuie să ținem cont când citim acest tabel este că aria corespunzătoare probabilității pe care o căutăm este definită între $-\infty$ și z. Acest lucru înseamnă că aria crește pe măsură ce z crește. Când z este 0, având în vedere că distribuția normală este simetrică, aria va fi exact 50% (0.5000). Drept urmare pentru a afla care este valoarea lui z corespunzătoare riscului la dreapta de 2.5%, va trebui să găsim aria complementară lui 2.5%, respectiv 97.5%. Uitându-ne în tabel vom observa că z va avea o valoare de 1.96 pentru un risc de 2.5% la dreapta. Acest rezultat nu ar trebui să fie surprinzător având în vedere că distribuția normală este simetrică și riscurile sunt și ele la rândul lor simetrice.

Înlocuind valorile obținute din tabel, împreună cu celelalte valori cunoscute în formula de calcul al limitelor intervalului de încredere, obținem:

$$x_1 = 10.25 - 1.96\frac{0.1}{\sqrt{100}} = 10.2304$$

$$x_2 = 10.25 + 1.96\frac{0.1}{\sqrt{100}} = 10.2696$$

Drept urmare, putem spune cu un nivel de încredere de 95% că media populației este între 10.2304 și 10.2696:

$$P(10.2304 < \mu < 10.2696) = 0.95$$

Acest exemplu acoperă cazul de risc simetric. Pentru riscurile unilaterale trebuie să determinați doar una dintre cele două limite (fie la stânga, fie la dreapta) iar pentru riscul bilateral asimetric calculul se face similar, luând fiecare limită pe rând.

Pentru a aplica această metodă de estimare, facem câteva presupuneri legate de eșantionul nostru și de distribuția acestuia:

- **Distribuția eșantionului este normală**: Pentru o dimensiune mare a eșantionului (de obicei n≥30), teorema limitei centrale spune că distribuția mediei eșantionului este aproximativ normală. Pentru eșantioane mai mici, această metodă se aplică numai dacă populația de bază este normală.
- **Dispersia populației este cunoscută**: Această metodă presupune că dispersia populației este cunoscută, ceea ce este rareori cazul în scenariile din lumea reală.
- **Eșantionare a fost făcută aleatoriu**: Metoda presupune că eșantionul este obținut prin utilizarea unei metode de eșantionare aleatorie.
- **Independența observațiilor**: Observațiile din eșantion trebuie să fie independente unele de altele.

8.2. Estimarea medie atunci când dispersia populației este necunoscută

În cazul în care dispersia populației este necunoscută, eșantionul nu este distribuit în mod normal. Distribuția pe care o vom folosi în acest caz se numește distribuția Student. Deoarece abaterea standard a populației este necunoscută, vom folosi în schimb abaterea standard a eșantionului, aceasta fiind cea mai bună aproximare pe care o avem la îndemână.

Problema are o formulare și o rezolvare similară cu cea a estimării când cunoaștem dispersia populație. Începem de la definirea nivelului de încredere:

$$P(x_1 < \mu < x_2) = 1 - \alpha$$

Formula pe care o vom folosi pentru determinarea limitelor intervalului de încredere va fi:

$$CI = \bar{x} \pm t\frac{s}{\sqrt{n}}$$

unde:

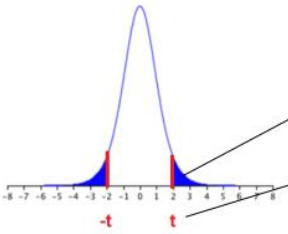
CI – limitele intervalului de încredere

t - este scorul t, care poate fi căutat în tabelul de distribuție Student și corespunde nivelului de încredere dorit (de obicei 90%, 95% sau 99%).

s - este abaterea standard a eșantionului

n - este dimensiunea eșantionului.

Pentru exemplificare vom lua cazul unui eșantion de n=5 valori, media eșantionului $\bar{x} = 7.5$ abaterea standard a eșantionului $s = 1.5$ și risc bilateral asimetric cu un risc stânga $\alpha_1 = 2\%$ și risc dreapta $\alpha_1 = 5\%$. Obținem nivelul de încredere scăzând din 100% valoarea totală a riscului (2%+5%=7%) și anume 93%. Ca și înainte, trebuie să găsim scoruri t dintr-un tabel pentru a determina limitele intervalului de încredere (Anexa 2).



Grade de libertate	0.005 (1) 0.01 (2)	0.01 (1) 0.02 (2)	0.025 (1) 0.05 (2)	0.05 (1) 0.10 (2)	0.10 (1) 0.20 (2)	0.25 (1) 0.50 (2)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718

Fig. 8.7. Tabelul scorurilor t

În tabelul t, aria corespunzătoare riscului este în scrisă în primul rând al tabelului (pentru riscul unilateral) și în al doilea rând al tabelului (pentru riscul bilateral simetric). Valoarea t corespunzătoare riscului ales este în corpul tabelului. Prima coloană conține ***gradele de libertate*** care sunt determinate cu formula:

$$\upsilon = n - 1$$

Citim scorul t la intersecția coloanei care are valoarea corespunzătoare riscului și a rândului cu gradele de libertate corespunzătoare.

În cazul exemplului nostru, avem 5 valori, deci vom avea 4 grade de libertate. Pe rândul 4 vom căuta valorile de pe coloanele corespunzătoare celor două riscuri. Deoarece vom considera fiecare capăt al intervalului separat, ne vom uita pe primul rând al tabelului pentru a citi valorile riscului. În tabel nu avem valoarea de 2% și vom alege cea mai apropiată valoare, respectiv 2.5% (0.025). La intersecția rândului cu 4 grade de liberate și coloanei cu riscul de 2.5% avem valoarea scorului t egală cu 2.776. Deoarece vorbim de risc stânga și media distribuției este 0, vom lua valoarea cu minus $t = -2.776$. Pentru riscul dreapta, avem valoarea de 5% (0.05) și aceleași 4 grade de liberate, deci valoarea lui t este $t = 2.132$.

Odată ce valorile scorurilor t au fost determinate, le introducem în formulă:

$$x_1 = 7.5 - 2.776\frac{1.5}{\sqrt{5}} = 5.638$$

$$x_2 = 7.5 + 2.776\frac{1.5}{\sqrt{5}} = 9.362$$

Drept urmare, putem spune că media populației în acest caz se află între 5.638 și 9.362 cu o probabilitate de 93%.

Ca și în cazul estimării cu valori z, există o serie de presupuneri pe care pe facem când estimăm media și nu cunoaștem dispersia:

- **Eșantionarea aleatorie**: eșantionul trebuie selectat aleatoriu din populație.
- **Independență**: Observațiile trebuie să fie independente unele de celelalte. Această ipoteză este adesea îndeplinită prin eșantionare aleatorie.
- **Normalitate**: Eșantionul ar trebui să provină dintr-o populație distribuită normal.
- **Dimensiunea eșantionului**: În timp ce distribuția t este utilă în special pentru eșantioane mici, dimensiunea eșantionului nu trebuie să fie extrem de mică. Un minim comun este de cel puțin 5.
- **Scara de măsurare**: Datele ar trebui să fie cel puțin scala de interval sau rație, deoarece aceste tipuri de date permit interpretarea semnificativă a mediei aritmetice.
- **Fără valori aberante**: valorile aberante pot denatura media și pot afecta împrăștierea, influențând astfel atât media estimată, cât și dispersia. Valorile aberante trebuie examinate cu atenție și gestionate în mod corespunzător.
- **Simetria**: testul t este mai sensibil la abaterea de la normalitate atunci când distribuția este deformată. În astfel de cazuri, testele neparametrice ar putea fi mai adecvate.

Înainte de realizarea estimării ar trebui testate ipotezele menționate mai sus. În caz contrar, rezultatele obținute pot fi denaturate.

8.3. Estimarea dispersiei populației

Uneori am putea dori să estimăm împrăștierea valorilor populației. Dispersia este unul dintre indicatorii care caracterizează împrăștierea valorilor.

Estimarea dispersiei poate fi utilă în controlul calității unde stabilitatea variației este adesea la fel de importantă ca stabilitatea medie. Un produs poate îndeplini cerințele de calitate în medie, dar dacă variația este mare, multe articole individuale vor fi defecte. În finanțe, dispersia (sau abaterea standard) este o măsură a volatilității sau a riscului unei investiții. Cunoașterea dispersiei ajută la testarea ipotezelor și la construirea de intervale de încredere pentru alți parametri.

Pentru a estima dispersia populației, trebuie să cunoaștem numărul de valori din eșantion (n) și dispersia (s^2) sau abaterea standard (s) a eșantionului. Modul de lucru este similar cu cel de la estimarea medie.

Definim problema noastră după urmează:

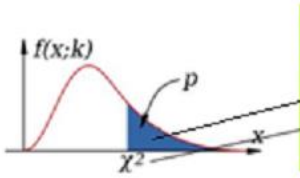
$$P(x_1 < \sigma^2 < x_2) = 1 - \alpha$$

Ceea ce înseamnă că σ^2 este între x_1 și x_2 , cu o probabilitate de $1 - \alpha$.

Formula pe care o vom folosi în acest caz este:

$$CI = \left[(n-1)\frac{s^2}{\chi^2_{superior}}, (n-1)\frac{s^2}{\chi^2_{inferior}}\right]$$

Valorile scorurilor χ^2 de care avem nevoie pentru a găsi x_1 și x_2 sunt luate din tabelul distribuției χ^2(Chi-pătrat). Un extras din tabelul χ^2 este afișat în Figura 8.8 iar acesta poate fi găsit integral în Anexa 3.



Grade de libertate	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086

Fig. 8.8. Citirea tabelului cu scoruri Chi-pătrat

Probabilitatea (aria de sub curbă) se găsește în primul rând din tabel iar valorile χ^2 sunt în corpul tabelului. În prima coloană avem din nou gradele de libertate ca și în cazul tabelului cu scoruri t. Valoarea χ^2 este situată la intersecția coloanei cu probabilitatea corespunzătoare riscului și rândul corespunzător gradelor de libertate. Când citim tabelul trebuie să ținem conte de câteva aspecte. Aria corespunzătoare riscului va crește de la $+\infty$ spre stânga. De asemenea, observați că pentru valoarea inferioară a intervalului de încredere vom folosi valoarea superioară a lui χ^2 și pentru valoarea superioară a intervalului χ^2 inferior. Acest lucru se datorează faptului că χ^2 se află la numitorul fracției.

Deci, un risc de 1% la dreapta este pe coloana 0.99, în timp ce un risc de 1% la stânga este pe coloana 0.01.

Să considerăm un exemplu unde vrem să estimăm dispersia populației cu un risc bilateral simetric de 10% știind că abaterea standard a eșantionului este $s = 0.1$ și avem $n = 5$ valori. Din riscul bilateral de 10% deducem două lucruri: nivelul de încredere este 90% iar fiecare risc are o valoare de 5% (stânga și dreapta). Din tabelul χ^2 de pe rândul cu 4 grade de libertate și coloana lui 0.05 obținem valoarea $\chi^2_{superior} = 9.488$ iar pentru $\chi^2_{inferior} = 0.711$ de pe coloana lui 0.95. Introducând datele în formulă, obținem:

$$CI = \left[(5-1)\frac{0.1^2}{9.488}, (5-1)\frac{0.1^2}{0.711}\right]$$

Din care rezultă $CI = [0.0042, 0.0562]$ indicând că dispersia populației în acest caz se află în acest interval cu o probabilitate de 90%.

Pentru a estima dispersia corect, trebuie satisfăcute înainte următoarele ipoteze:

- **Eșantionare aleatorie**: Eșantionul trebuie selectat aleatoriu din populație. Acest lucru asigură că eșantionul este reprezentativ pentru populație, făcând estimarea mai precisă.
- **Dimensiunea eșantionului**: În mod ideal, dimensiunea eșantionului ar trebui să fie suficient de mare pentru a permite o estimare robustă. Deși nu există o regulă simplă pentru ceea ce constituie un eșantion „mare”, o recomandare comună este cel puțin 30 de observații pentru ca teorema limită centrală să intre în joc. Cu toate acestea, metoda chi-pătrat pentru estimarea variației este încă aplicabilă și în cazul eșantioanelor mai mici, atâta timp cât datele sunt distribuite în mod normal.
- **Normalitate**: datele ar trebui să fie distribuite aproximativ normal, mai ales dacă dimensiunea eșantionului este mică.
- **Independența observațiilor**: Valorile din eșantion trebuie să fie independente unele de altele. Această ipoteză este adesea îndeplinită prin procesul de eșantionare aleatorie. Dacă datele sunt dependente (de exemplu, date din seria temporală), metodele alternative pot fi mai adecvate pentru estimarea dispersiei.
- **Date numerice**: datele ar trebui să fie la nivel de interval sau rație, deoarece acestea sunt tipurile de date în care este logic să discutăm despre dispersie.
- **Fără valori aberante**: valorile aberante pot avea un efect disproporționat asupra dispersie, modificând astfel estimarea. Datele ar trebui verificate pentru valori aberante, care ar trebui gestionate în mod corespunzător înainte de estimare.

Încălcarea acestor ipoteze poate duce la rezultate părtinitoare sau înșelătoare. Prin urmare, este important fie să se valideze ipotezele în măsura în care este posibil înainte de a continua cu analiza, fie să se utilizeze tehnici statistice care sunt robuste la încălcările acestor ipoteze.

Putem rezuma cele trei cazuri în Tabelul 8.2.

Tabel 8.2. Estimarea parametrilor populației (rezumat)

Estimare	Parametrii cunoscuți	Distribuție	Statistica	Formula de calcul
Media populației	n, $\bar{x}$, σ^2, σ, α	Normal	z	$CI = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$
	n, $\bar{x}$, s^2, s, α	Student	t	$CI = \bar{x} \pm t\frac{s}{\sqrt{n}}$
Dispersia populației	n, s^2, s, α	Chi-pătrat	χ^2	$CI = (n-1)\frac{s^2}{\sigma^2}$

Estimarea parametrilor populației ne ajută în procesul decizional prin furnizarea unor informații suplimentare despre populația noastră. Chiar dacă valorile obținute nu sunt exacte, aceste tot ne pot ajuta să luăm decizii dându-ne niște valori limită cu o anumită probabilitate.

Distribuția normală este un instrument practic pentru o multitudine de discipline. Prevalența sa în fenomenele naturale o face esențială pentru științele sociale în standardizarea testelor și a scorurilor, permițându-ne să comparăm diferite seturi de date pe o scară comună.

Distribuția Student joacă un rol esențial atunci când se tratează eșantioane de dimensiuni mici, un scenariu comun în studiile experimentale și clinice în care obținerea unor eșantioane mari poate fi nepractică sau imposibilă. Ea permite cercetătorilor să facă deducții cu privire la mediile populației cu un anumit nivel de încredere, în ciuda faptului că dispun de date limitate. Acest lucru este deosebit de valoros în domenii precum psihologia și medicina, în care variabilitatea individuală este ridicată, iar dimensiunile eșantioanelor sunt adesea limitate.

Distribuția chi-pătrat este un actor cheie în analiza datelor categoriale. Este piatra de temelie a testului de independență chi pătrat, care permite cercetătorilor să determine relația dintre variabilele categoriale. Acest lucru este extrem de util în domenii precum genetica pentru a evalua asocierea trăsăturilor genetice, în marketing pentru a evalua preferințele consumatorilor și în ecologie pentru a studia distribuția speciilor.

Aceste distribuții nu sunt doar abstracții matematice; ele sunt lentilele prin care privim și dăm sens lumii. Ele le permit statisticienilor să tragă concluzii semnificative din date, să testeze ipoteze și, în cele din urmă, să contribuie la avansarea cunoștințelor în diverse domenii. Ca atare, înțelegerea proprietăților și aplicațiilor acestora este importantă pentru orice statistician, om de știință sau cercetător care dorește să ia decizii informate pe baza datelor.

8.4. Verificarea cunoştinţelor

1. Ce parametru este estimat de media eșantionului?

a. Mediana populației
b. Modul populației
c. Media populației

2. Teorema limitei centrale este importantă în estimare deoarece:

a. Permite ca media eșantionului să fie utilizată ca o estimare punctuală pentru media populației.
b. Afirmă că distribuția mediilor eșantionului va fi distribuită normal, indiferent de mărimea eșantionului.
c. Asigură faptul că varianța eșantionului este un estimator nepărtinitor al variației populației.

3. Atunci când dispersia populației este necunoscută și dimensiunea eșantionului este mică, ce distribuție ar trebui utilizată pentru estimare?

a. Distribuția normală
b. Distribuția binomială
c. Distribuția Student

4. Care este scopul construirii unui interval de încredere?

a. Pentru a furniza un interval de valori care este probabil să conțină parametrul populației.
b. Pentru a determina cu precizie parametrul populației.
c. Pentru a testa o ipoteză cu privire la parametrul populației.

5. Lățimea unui interval de încredere pentru estimarea mediei unei populații va:

a. Crește pe măsură ce mărimea eșantionului crește.
b. Scade pe măsură ce scade dispersia eșantionului.
c. Rămâne constantă indiferent de mărimea eșantionului.

6. Estimarea punctuală pentru dispersia populației este:

a. Intervalul de variație al datelor din eșantion.
b. Abaterea standard a eșantionului.
c. Dispersia eșantionului.

7. Care dintre următoarele este un exemplu de estimare punctuală?

a. Media eșantionului
b. Intervalul de încredere
c. Testul de ipoteză

8. Nivelul de încredere (de exemplu, 95%) în contextul unui interval de încredere se referă la:

a. Procentul din datele eșantionului care se încadrează în interval.
b. Probabilitatea ca intervalul să includă adevăratul parametru al populației.
c. Proporția din populația din care a fost extras eșantionul.

9. Dacă ați dori să estimați media populației cu un grad mai mare de încredere, ce ați face?

a. Măriți dimensiunea eșantionului.
b. Micșorați lățimea intervalului de încredere.
c. Utilizați un nivel alfa mai mare.

10. Pentru a estima media unei populații atunci când dispersia este cunoscută, folosim:

a. Distribuția normală
b. Distribuția binomială
c. Distribuția Student

Răspunsuri corecte

1. c. Media populației

2. a. Permite utilizarea mediei eșantionului ca o estimare punctuală pentru media populației.

3. c. Distribuția Student

4. a. Pentru a furniza un interval de valori care este probabil să conțină parametrul populației.

5. b. Scade pe măsură ce scade dispersia eșantionului.

6. c. Dispersia eșantionului.

7. a. Media eșantionului

8. b. Probabilitatea ca intervalul să includă adevăratul parametru al populației.

9. a. Creșteți dimensiunea eșantionului.

10. a. Distribuția normală

9. Controlul statistic al proceselor

Controlul statistic al proceselor (SPC) este o metodă de monitorizare, control și îmbunătățire a proceselor prin analiză statistică. Aceasta conține șase instrumente care pot fi folosite pentru a identifica problemele și cauzele lor, pentru o mai bună înțelegere și monitorizare a procesului:

- Histograma
- Diagrama Pareto
- Diagrama cu puncte
- Cartelele de control
- Diagrama cauză-efect
- Diagrama de proces

În continuare vom explora fiecare dintre aceste instrumente, unele mai detaliat altele mai superficial.

9.1. Histograma

Histogramele sunt utilizate pentru a vizualiza distribuția datelor continue. Prin cunoașterea modului în care datele sunt distribuite putem trage concluzii cu privire la acest proces. Histograma a fost discutată în secțiunile anterioare și nu vom insista asupra ei. Figura 9.1 ilustrează o histogramă a diametrelor pieselor fabricate într-o fabrică.

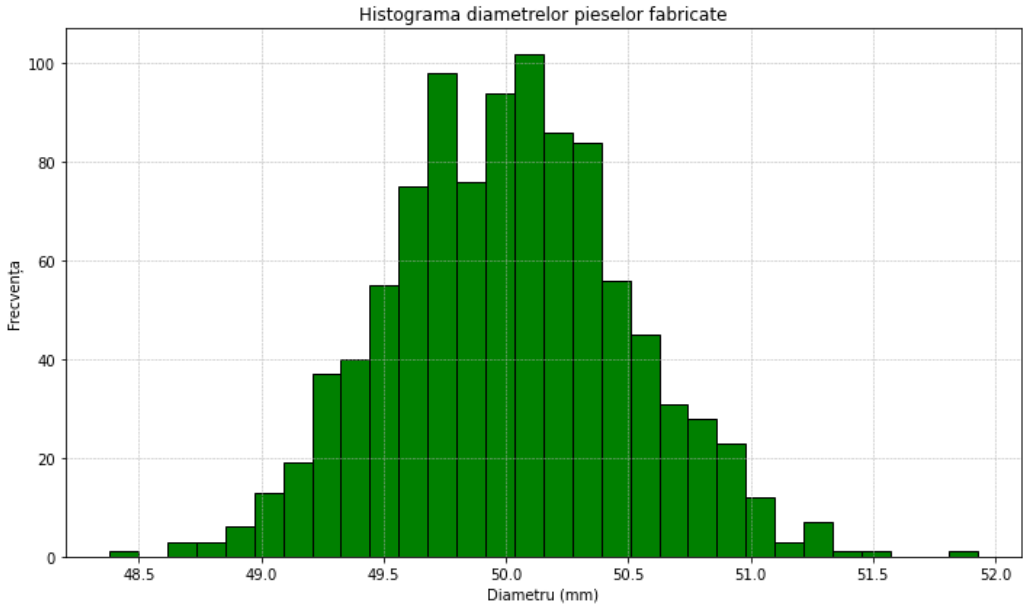


Fig. 9.1. Un exemplu de histogramă

Cu ajutorul histogramei putem observa forma distribuției datelor și dacă aceasta se apropie de o distribuție dorită, cum ar fi cea Normală.

9.2. Diagrama Pareto

Scopul principal al diagramei Pareto este de a evidenția cei mai importanți factori dintr-un set de date. Diagrama Pareto a fost dezvoltată de Vilfredo Pareto în anii 1800 și se bazează pe principiul Pareto. Acesta, denumit adesea regula 80/20, este un principiu care afirmă că 80% din rezultate (sau ieșiri) sunt rezultatul a 20% din toate cauzele (sau intrările) unui anumit eveniment. Cu alte cuvinte, în multe situații, un număr mic de cauze va produce majoritatea rezultatelor sau efectelor. De exemplu: 80% din profiturile unei companii pot proveni de la 20% dintre clienții săi; 80% din reclamații ar putea proveni de la 20% dintre clienți; 80% din erorile software ar putea fi cauzate de 20% din problemele cunoscute. Valorile (80/20) nu sunt proporții stricte, ci doar orientative și ilustrează faptul că există o disproporționalitate între intrări și ieșiri. Principiul Pareto este un instrument puternic care poate ajuta la creșterea eficienței și eficacității. Înțelegând care factori (cei câțiva esențiali) contribuie la majoritatea rezultatelor, se pot concentra eforturile și resursele asupra acelor domenii critice, ceea ce duce la rezultate mai bune și la o mai bună utilizare a resurselor.

În controlul calității, diagrama Pareto este utilizată pentru a vizualiza acest principiu, ajutând la prioritizarea problemelor sau cauzelor care trebuie abordate. Abordând cele câteva cauze majore care conduc la majoritatea problemelor, organizațiile pot obține îmbunătățiri semnificative cu un efort relativ limitat.

Diagrama Pareto (Figura 9.2), conține atât bare, cât și un grafic cu linii. Valorile individuale sunt reprezentate în ordine descrescătoare prin bare, în timp ce totalul cumulativ este reprezentat de linie.

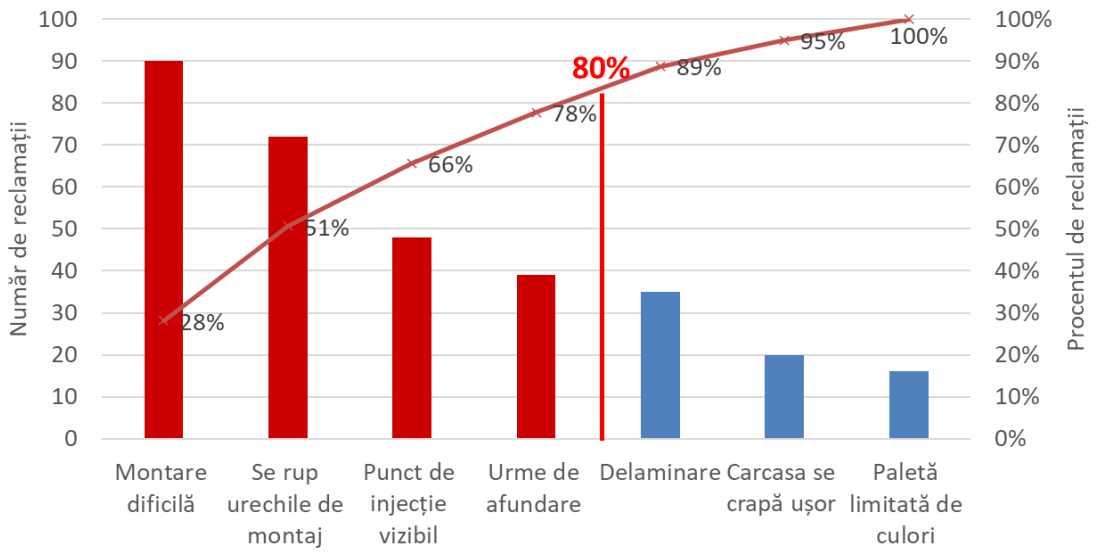


Fig. 9.2. Un exemplu de o diagramă Pareto

Crearea unei diagrame Pareto implică niște pași sistematici pentru a se asigura că aceasta vizualizează în mod eficient cei mai importanți factori dintr-un set de date:

1. **Identificarea categoriilor**: Decideți care sunt categoriile sau factorii pe care doriți să îi analizați. Acestea ar putea fi tipuri de defecte, surse de reclamații sau orice alte clasificări relevante pentru problema în cauză.
2. **Colectarea datelor**: Adunați date pentru fiecare categorie. Acest lucru ar putea implica numărarea defectelor de fiecare tip, numărarea plângerilor pentru fiecare categorie etc.
3. **Ordonarea categoriilor**: Aranjați categoriile în ordine descrescătoare în funcție de frecvență.
4. **Calcularea totalurilor și procentelor cumulate**: Pentru fiecare categorie, calculați totalul cumulat și procentul cumulat.
5. **Trasarea graficului**: Reprezentați categoriile pe axa orizontală, începând din stânga cu categoria cu frecvența cea mai mare. Reprezentați frecvența pentru fiecare categorie sub formă de bare. Înălțimea barei reprezintă frecvența categoriei respective. Folosind o a doua axă verticală în dreapta, reprezentați procentul cumulat sub forma unui grafic cu linii. Această linie ar trebui să înceapă în partea de sus a primei bare și să se încheie la 100% pe ultima coloană.
6. **Trasarea liniilor de referință (opțional)**: Puteți adăuga o linie de referință la marcajul de 80% pe axa procentului cumulativ pentru a identifica cu ușurință cele câteva categorii vitale care contribuie la 80% din efect.
7. **Analiza graficului**: Cele mai din stânga coloane (și categoriile pe care le reprezintă) sunt de obicei cele mai semnificative contribuții la problemă sau situație. Barele din dreapta reprezintă categoriile care au un efect mai mic, dar care ar putea totuși să necesite atenție în contexte specifice.
8. **Luarea măsurilor**: Folosiți informațiile din graficul Pareto pentru a prioritiza acțiunile. Concentrați-vă asupra categoriilor care au cel mai semnificativ impact. Abordarea problemelor din aceste domenii poate duce la cele mai substanțiale îmbunătățiri. Atenție, această prioritizare se face strict prin prisma frecvenței de apariție și nu a importanței efectului. Unele cauze, deși rare, pot avea efecte semnificative în problema pe care încercați să o soluționați. Folosiți-vă judecata în prioritizarea acțiunilor.
9. **Revizuirea și actualizarea graficului**: De-a lungul timpului, pe măsură ce se întreprind acțiuni și se schimbă procesele, distribuția problemelor pe categorii se poate schimba și ea. Este important să revizuiți și să actualizați periodic graficul Pareto pentru a reflecta situația actuală și pentru a vă asigura că eforturile rămân concentrate asupra domeniilor critice.

Categoriile care alcătuiesc 80% din total, sunt cele mai frecvente și care ar trebui să fie abordate. În exemplul de mai sus, primele patru categorii alcătuiesc aproximativ 80%.

9.3. Diagrama cu puncte

O diagramă cu puncte este o reprezentare grafică în care fiecare observație din setul de date este reprezentată ca un punct. Poziția fiecărui punct este determinată de valoarea a două variabile: o variabilă determină poziția pe axa x, iar cealaltă variabilă determină poziția pe axa y.

Diagrama are ca scop vizualizarea și compararea relației sau corelației dintre două variabile cantitative. Se poate folosi de asemenea pentru a identifica modele, tendințe, grupuri sau valori aberante în date.

Diagrama este compusă din (Figura 9.3):

- Axa X (axa orizontală): Reprezintă valorile unei variabile.
- Axa Y (axa verticală): Reprezintă valorile celei de-a doua variabile.
- Puncte: Fiecare punct de pe grafic reprezintă o singură observație din setul de date.

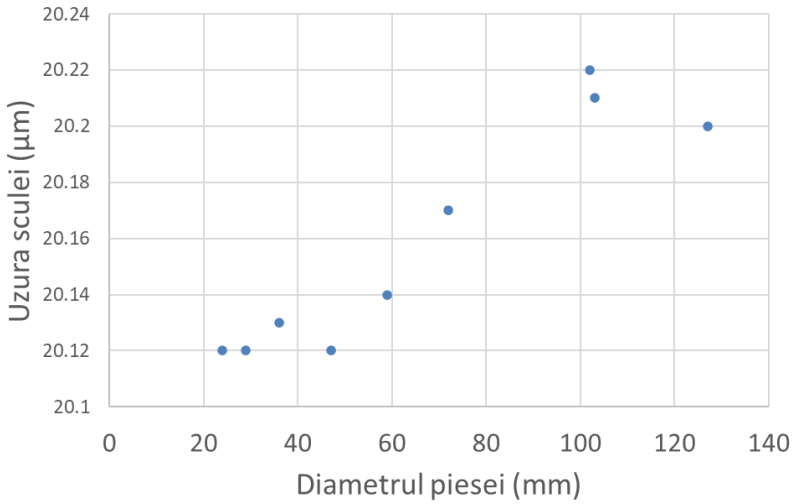


Fig. 9.3. Un exemplu de un diagramă cu puncte

În exemplul din Figura 9.3. avem pe axa X diametrul unei piese exprimat în milimetri iar pe axa verticală uzura sculei folosite pentru obținerea piesei în micrometri. Fiecare punct reprezintă o pereche de valori diametru-uzură colectată pentru fiecare piesă în parte. Reprezentând astfel mai multe perechi de valori, putem observa că pe măsură ce diametrul piesei crește și uzura sculei crește. Drept urmare, punctele ilustrează o tendință crescătoare vizibilă pe grafic.

Pentru realizarea unui astfel de grafic trebuie realizați următorii pași:

1. Alegeți variabilele pe care doriți să le analizați.
2. Reprezentați valorile unei variabile pe axa x.
3. Reprezentați valorile celei de-a doua variabile pe axa y.
4. Pentru fiecare observație din setul de date, marcați un punct unde valorile x și y se intersectează.

Graficele cu puncte, deși simple, pot să ne ajute în înțelegerea relațiilor dintre două variabile, să identificăm tendințe și să identificăm valori aberante.

9.4. Cartele de control

Procesele de producție pot fi supuse diferitelor variații datorate fluctuațiilor în calitatea materiei prime, variațiilor în setările de mașină, erori umane și multe altele. Dacă aceste variații nu sunt monitorizate și controlate, ele pot duce la un produs sau serviciu de calitate inferioară. Cartelele sau diagramele de control ne ajută să identificăm și să corectăm aceste variații înainte ca acestea să devină probleme serioase.

Cartelele de control sunt larg utilizate în ingineria calității și managementul proceselor pentru monitorizarea proceselor și identificarea deviațiilor semnificative de la un standard stabilit. Diagramele de control pot fi utilizate pentru a măsura parametri multipli, cum ar fi media sau mediană și amplitudinea sau abaterea standard. Ele ajută la identificarea variațiilor din proces, și ajută la a face distincția între fluctuațiile naturale și anomaliile care necesită acțiuni corective. Acestea permit o abordare proactivă a controlului calității, minimizând defectele și reducând risipa.

9.4.1. Capabilitatea procesului

Diagrame de control pot fi utilizate și pentru a determina dacă un proces poate îndeplini cerințele specificate. În acest caz vorbim de capabilitatea procesului care este abilitatea unui proces sau a unei mașini de a funcționa într-un mod care satisface specificațiile de calitate. Pentru a determina capabilitatea procesului trebuie să știm:

- **LIS** – limita inferioară de specificație
- **LSS** – limita superioară de specificație
- **s** – abaterea standard a procesului

Limitele de specificație pot fi limite de toleranță sau limitele de control ale unui proces. Formula pentru capabilitatea procesului atunci când folosim limitele de toleranță este:

$$Cp = \frac{LST - LIT}{6s}$$

Unde:

LST – limita superioară de toleranță

LIT – limita inferioară de toleranță

s – abaterea standard a datelor colectate din proces

Un alt indicator al capabilității este Cpk, indicele capabilității procesului. Pentru a-l determina trebuie să calculăm distanța față de media valorilor $\bar{x}$ și apoi capabilitatea superioară (Cps) și inferioară (Cpi):

$$Cps = \frac{LST - \bar{x}}{3s}$$

$$Cpi = \frac{\bar{x} - LIT}{3s}$$

Din aceasta derivăm Cpk cu formula:

$$Cpk = Min(Cpi, Cps)$$

Cu cât capabilitatea procesului este mai mare cu atât procesul este mai performant. Un proces cu o capabilitate mai mare decât 1.33 este considerat capabil, în timp ce unul cu o capabilitate mai mică decât 1 este considerat incapabil. Un proces cu o capabilitate între 1 și 1.33 este la limită capabil și are nevoie de unele ajustări pentru a deveni capabil. Unele companii folosesc limite mai stricte decât valoarea limită de 1.33.

9.4.2. Elemente ale unei diagrame de control

O diagramă de control monitorizează evoluția mediei (sau a altui parametru de tendință centrală) și răspândirea datelor. Ținta este plasată pe linia centrală, iar limitele sunt de fiecare parte a acestuia (limitele superioare și inferioare de control). Limitele de control sunt de obicei plasate la distanța de 3 abateri standard de o parte și de alta a mediei. Limitele de avertizare pot fi de asemenea puse (Figura 9.4.). Dacă se adaugă limitele de avertizare, acestea sunt poziționate la 90% din distanța dintre linia centrală și limitele de control.

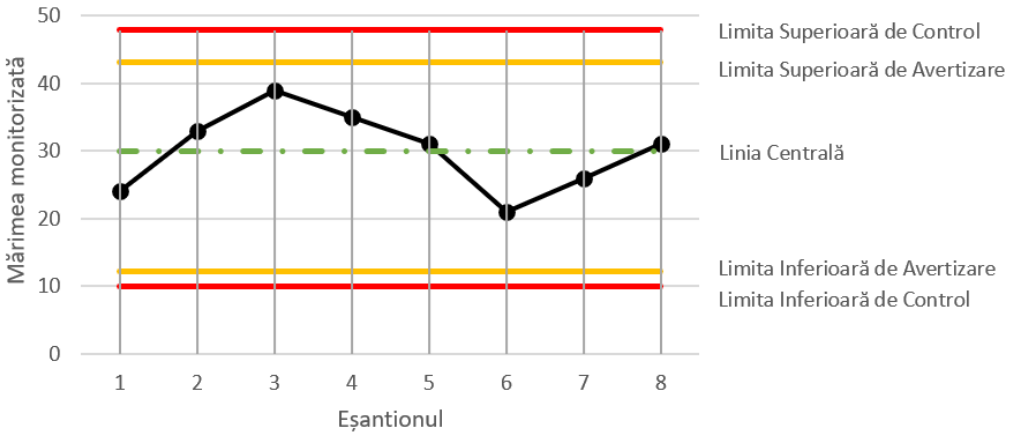


Fig. 9.4. Elementele unei diagrame de control

Fiecare abatere standard de la medie împarte graficul în 3 zone simetrice: A, B și C. Zona A este cea mai îndepărtată de media și zona C este zona centrală (Figura 9.5.). Aceste zone ne vor ajuta să identificăm eventuale probleme în proces.

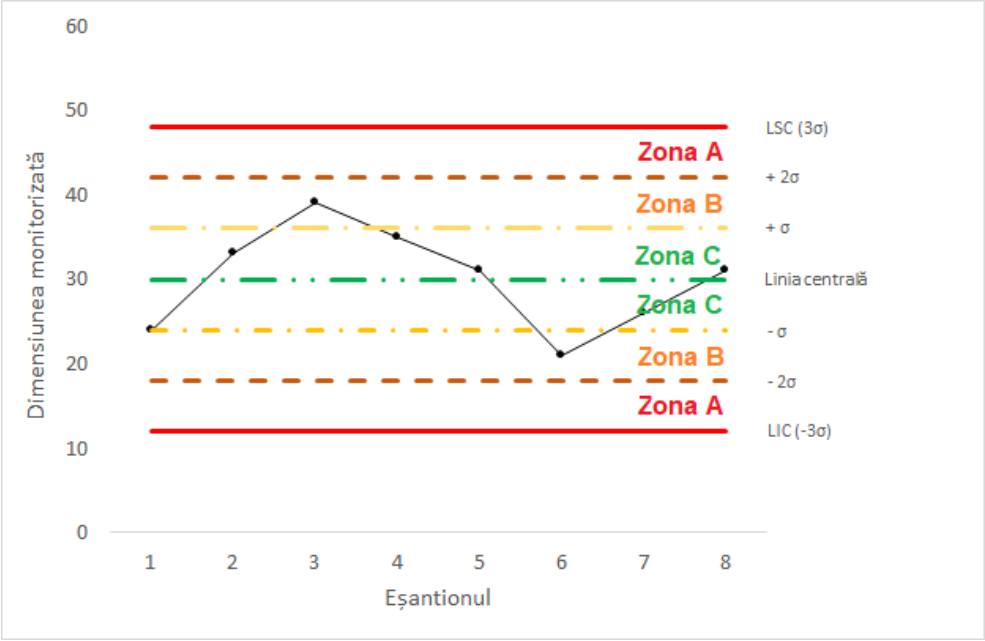


Fig. 9.5. Elemente ale unei diagrame de control tipice și cele trei zone

Pentru a utiliza corect diagramele de control, trebuie să avem un proces stabil, care să nu varieze foarte mult în timp, cu date normal distribuite, iar limitele de control trebuie să cadă de o parte și de alta a liniei de centru.

9.4.3. Tipuri de cartele de control

Există diferite diagrame de control în funcție de tipul de date și de dimensiunea eșantionului. Figura 9.6 prezintă diferitele tipuri de diagrame de control.

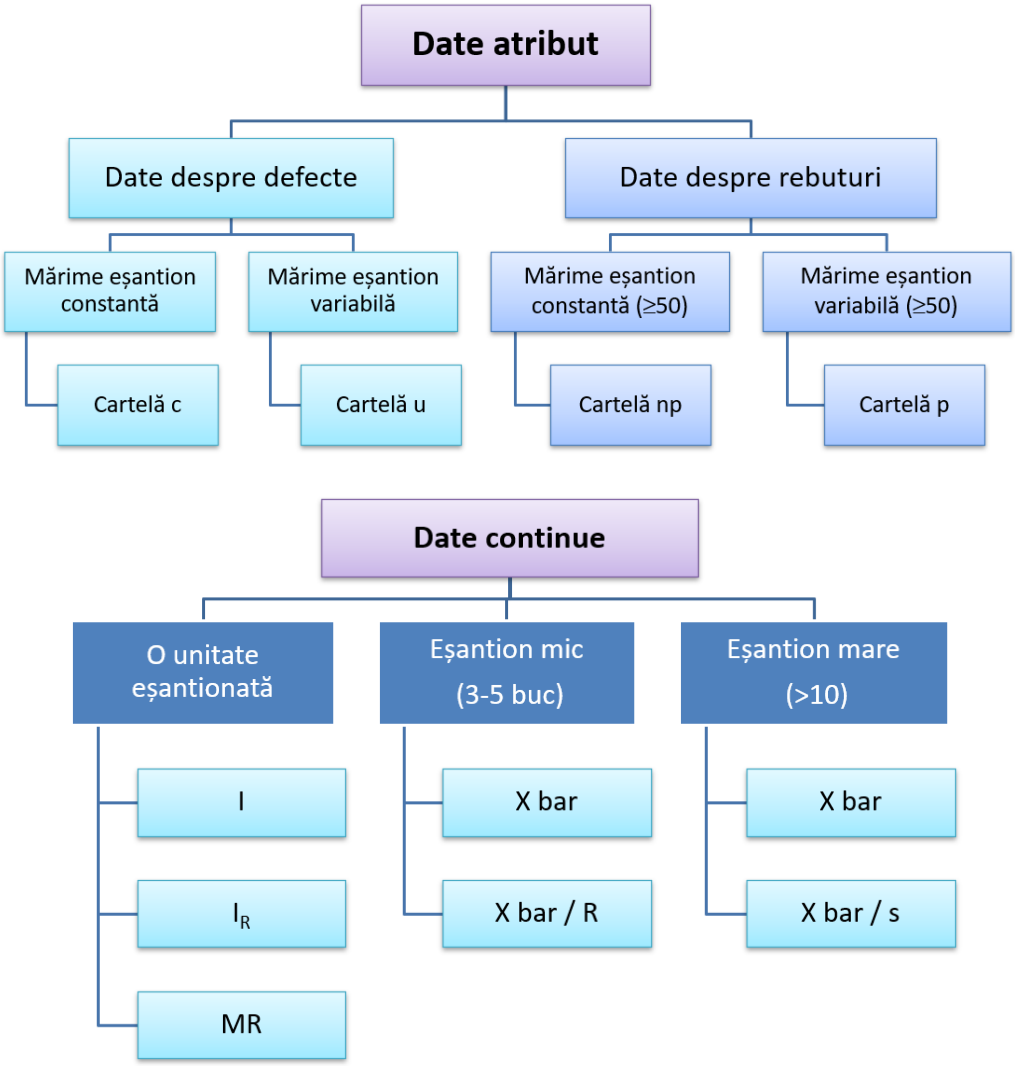


Fig. 9.6.Cartele de control în funcție de tipul de date și dimensiunea eșantionului

Putem avea date continue (sau variabile) și date discrete (sau atribut). Pentru date continue avem I/MR, X-Bar/R și X-Bar/S diagrame și pentru datele atribut avem c, u, NP și p diagrame. Diagramele I/MR folosesc valori individuale, în timp ce diagramele X-Bar se folosesc când avem grupuri de eșantioane. În cazul în care subgrupul este mic, utilizăm amplitudinea (R) pentru a caracteriza răspândirea și atunci când dimensiunea eșantionului este mai mare folosim abaterea standard (s).

Diagramele pentru atribute utilizează date discrete (numărabile). În analiza de control al calității, aceste date numărabile se încadrează într-una din două categorii:

- **Defecte** - reprezintă numărul de neconformități ele unui element, cum ar fi o piesă. Nu există nicio limită a numărului de defecte posibile. Diagramele de defecte contorizează numărul de defecte din unitatea de inspecție.
- **Rebuturi** - cazul în care întregul articol este considerat a nu fi conform cu specificațiile produsului. Fiecărui element îi poate fi asociat un singur număr: 1 sau 0. Diagramele de rebuturi contorizează numărul de rebuturi dintr-un subgrup.

În funcție de mărimea eșantionului putem avea diferite cartele. Cartela de tip ***c*** se folosește pentru monitorizarea defectelor iar eșantioane au mărime constantă. Dacă eșantionul are mărime variabilă, vom folosi cartelele de tip ***u***. În cazul rebuturilor, pentru eșantioane cu mărime constantă vom folosi cartela de tip ***n*** iar pentru eșantioane de mărime variabilă, cartela ***np***.

Cartela de tip c

Cartela de tip C, este utilizată pentru a monitoriza numărul de defecte dintr-o anumită unitate în timp. Acest tip de grafic este deosebit de util atunci când ne interesează mai degrabă numărul total de defecte pe unitate decât proporția defectelor. De exemplu, ați putea utiliza o diagramă C pentru a monitoriza numărul de zgârieturi de pe piesele de mașină vopsite sau numărul de sigilii rupte în loturile de conserve alimentare.

Deoarece cartela ilustrează număr de defecte, datele dintr-o astfel de diagramă nu pot fi negative. Acesta este motivul pentru care limita inferioară de control (LIC) este adesea zero. Cartela de tip C presupune de obicei că defectele urmează o distribuție Poisson. Această ipoteză este importantă pentru interpretarea corectă a limitelor de control și a altor proprietăți statistice. Deoarece folosește date obținute prin numărare, o cartelă de tip C poate fi mai sensibilă la schimbările din proces care duc la o creștere a numărului de defecte pe unitate, ceea ce o face valoroasă pentru detectarea din timp a problemelor de calitate.

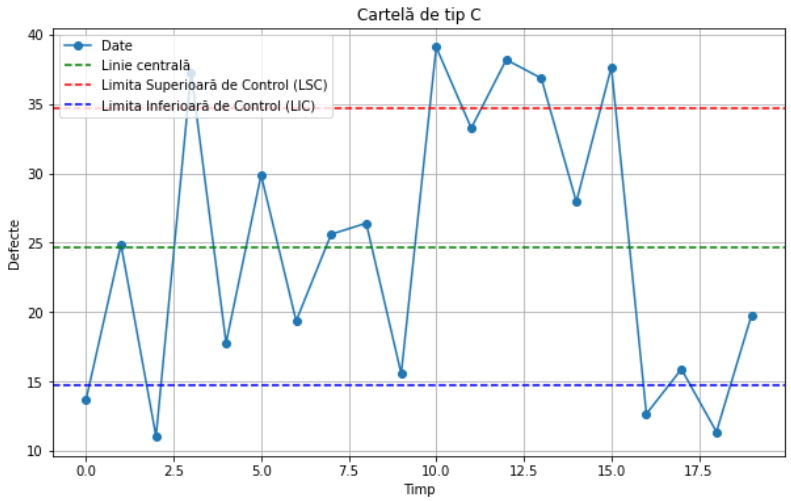


Fig. 9.7.Un exemplu cartelă de tip c

Cartela este formată din următoarele elemente:

1. **Punctele** care reprezintă date colectate: Acestea reprezintă numărul de defecte pe unitate în cadrul procesului. De exemplu, dacă monitorizați calitatea lucrărilor de vopsire a mașinilor, fiecare punct ar putea reprezenta numărul de zgârieturi de pe o singură mașină.

2. **Linia centrală**: Această linie reprezintă numărul mediu de defecte pe toate unitățile. Ea servește drept linie de bază pentru evaluarea stabilității procesului.
3. **Limita superioară de control** (LSC): Această linie este calculată pe baza numărului mediu de defecte și reprezintă pragul superior pentru ceea ce se consideră variație normală în cadrul procesului [18], [19]:

$$LSC = \bar{c} + 3\sqrt{\bar{c}}$$

unde:

$$\bar{c} = c/m$$

iar

c- numărul de defecte

m – numărul de eșantioane

4. **Limita inferioară de control** (LIC): În mod similar, această linie este calculată pe baza numărului mediu de defecte și reprezintă pragul inferior pentru ceea ce este considerat variație normală. [18], [19]:

$$LIC = \bar{c} - 3\sqrt{\bar{c}}$$

Deoarece nu putem avea rezultate negative, această limită este considerată 0 dacă rezultatul calculului este negativ:

$$LIC = \max(0, \quad \bar{c} - 3\sqrt{\bar{c}})$$

5. **Axa timpului**: Axa orizontală reprezintă secvența în care sunt colectate datele, care este de obicei bazată pe timp.

Cartela de tip u

O cartelă u (u este pentru unitate) este o diagramă de control atribut care afișează modul în care frecvența de defecte, sau neconformități, se schimbă în timp pentru un proces sau sistem. Numărul de defecte este colectat pentru zona de oportunitate în fiecare subgrup. Zona de oportunitate poate fi fie un grup de elemente, fie doar un element individual pe care se efectuează numărarea defectelor. Diagrama u este un indicator al consecvenței și predictibilității nivelului de defecte în proces. O diagramă u este adecvată atunci când zona de oportunitate pentru un defect variază de la subgrup la subgrup

Cartela de tip u (Figura 9.8) este cea mai potrivită pentru monitorizarea ratei de defecte pe unitate atunci când mărimea eșantionului variază. Spre deosebire de cartela de tip C, care se concentrează pe numărul total de defecte, cartela de tip U este concentrată pe rata defectelor, permițând o comparație standardizată între eșantioane

de diferite dimensiuni. Este deosebit de utilă în scenariile în care volumele de producție sau dimensiunile loturilor fluctuează.

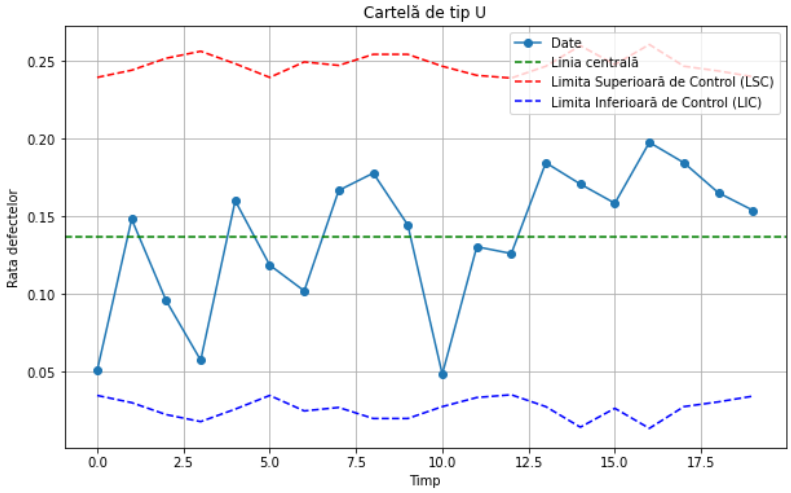


Fig. 9.8. Un exemplu cartelă de tip u

Cartela de tip np

Diagrama np (Figura 9.9) este concepută pentru monitorizarea numărului de articole defecte într-un eșantion de dimensiuni constante. Este un indicator al consecvenței și predictibilității nivelului de defecte dintr-un proces. Diagrama conține aceleași elemente ca în cazul cartelei de tip u, precum linia centrală și limitele de control, cu precizarea că fiecare punct reprezintă numărul de defecte dintr-un eșantion nu dintr-o unitate.

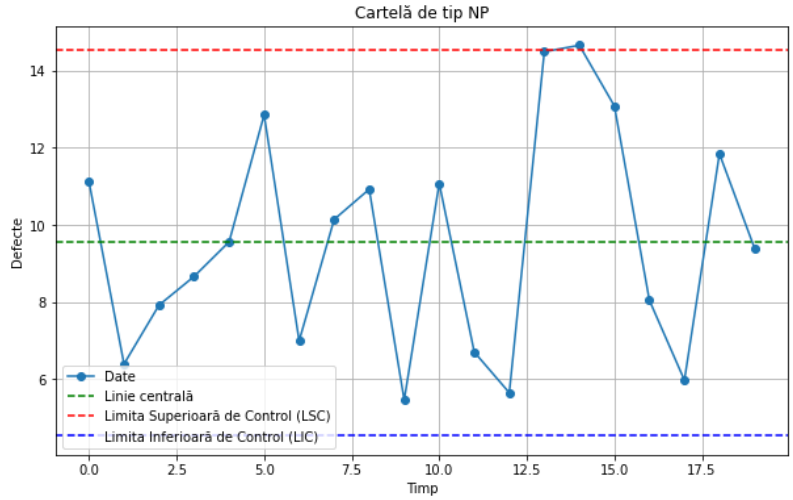


Fig. 9.9.Un exemplu de cartelă np

Imaginați-vă un scenariu în care verificați calitatea comprimatelor produse într-un lot de 500. Diagrama np va afișa numărul de tablete defecte din fiecare lot, ajutându-vă să identificați orice problemă de calitate.

Cartela de tip p

Cartela de tip p (Figura 9.10) este utilizat pentru a monitoriza proporția de articole defecte dintr-un eșantion. Aceasta este utilizată atunci când dimensiunea subgrupului variază, iar diagrama afișează proporția sau fracția elementelor respinse, mai degrabă decât numărul respins.

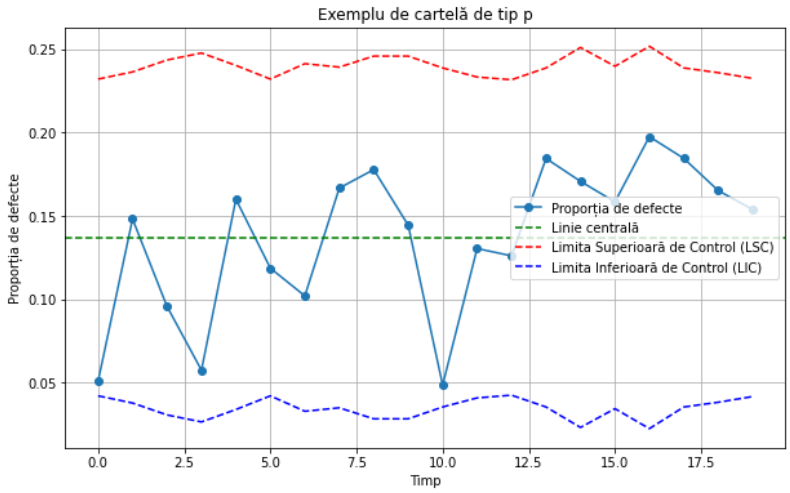


Fig. 9.10.Un exemplu de cartelă p

Cartelele de tip P sunt concentrate pe proporții sau procente, și nu pe numărul total de defecte ca în cartelele de tip C.

Aceste grafice presupun adesea o distribuție binomială a datelor, ceea ce le face adecvate pentru eșantioane mari și proporții care nu sunt extrem de mici sau mari.

Cartela de tip I

Cartela de tip I (Figura 9.11) este utilizat pentru a monitoriza variația unor valori continue, individuale în timp. În această cartelă fiecare punct este o observație, ca spre exemplu durata unei operații în minute.

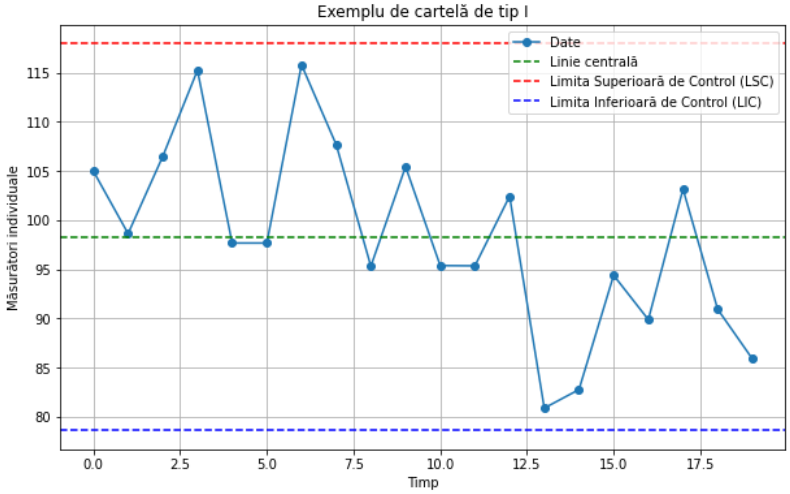


Fig. 9.11.Un exemplu de cartelă de tip I

Aceasta este o cartelă pentru valori continue și, spre deosebire de cartelele pentru date de tip atribut, va avea și valori care nu sunt întregi. Spre exemplu, putem avea diametrul unei piese de 23.57 mm. Deoarece fiecare punct reprezintă o valoare individuală, graficul este sensibil la schimbări mici în proces spre deosebire de alte grafice care folosesc media. Acest tip de grafic este folosit când datele noastre urmează sau se apropie de o distribuție normală. De obicei este folosită în paralel cu o cartelă care urmărește stabilitatea variației (cum ar fi cartela de tip MR).

Cartela de tip X-Bar

Cartela X-Bar (Figura 9.12) este utilizată pentru a monitoriza evoluția în timp a mediei unor valori dintr-un subgrup. Fiecare punct de pe grafic reprezintă o medie a unui eșantion (sau subgrup). În funcție de dimensiunea eșantionului, aceasta poate fi folosită împreună cu cartela de tip R sau S. Graficul arată cât de consecvent și previzibil este un proces.

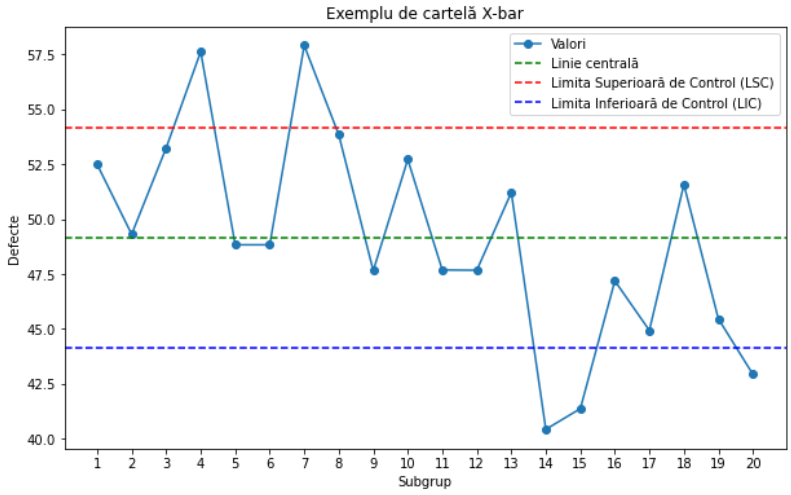


Fig. 9.12.Un exemplu de cartelă X-Bar

Cartela de tip MR

Cartela MR (Moving-Range) este utilizată pentru a monitoriza variația datelor (Figura 9.13). Se utilizează împreună cu diagrama I. Fiecare valoare de pe grafic reprezintă diferența dintre două observații consecutive. Deoarece diferența reprezintă amplitudinea (range), iar aceasta se aplică succesiv fiecărui cuplu de valori, „mișcându-se" astfel pe grafic de la stânga la dreapta, cartela este numită Moving-Range („amplitudine mobilă").

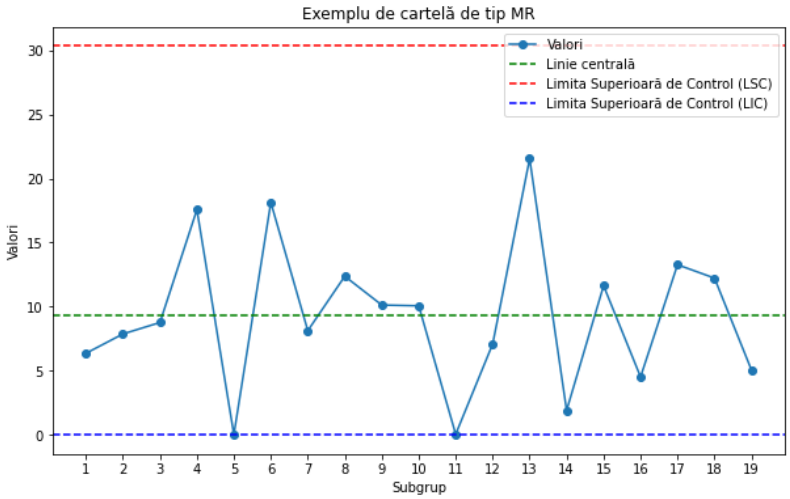


Fig. 9.13.Un exemplu de cartelă MR

Cartela de tip R

Cartela de tip R (Figura 9.14) este folosită în analiza variației datelor dintr-un grup de eșantioane utilizând amplitudinea (range). Este utilizată adesea împreună cu cartela X-bar. Dimensiunea eșantionului este de obicei mică (mai puțin de 5 valori). Fiecare valoare din diagramă reprezintă amplitudinea subgrupului de eșantioane.

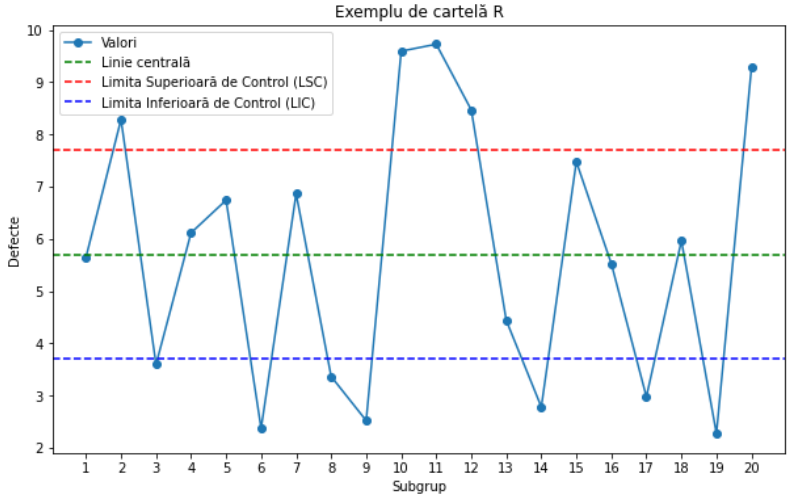


Fig. 9.14. Un exemplu de cartelă R

Cartela de tip S

Cartela S (Figura 9.15) este utilizată ca și cartela R pentru a monitoriza variația datelor, dar este utilizată atunci când dimensiunea subgrupului este mai mare (mai mult de 5 valori într-un subgrup). Pentru fiecare subgrup, se calculează abaterea standard a valorilor din subgrup care este apoi trasată pe grafic. Se poate folosi împreună cu diagrama X-bar.

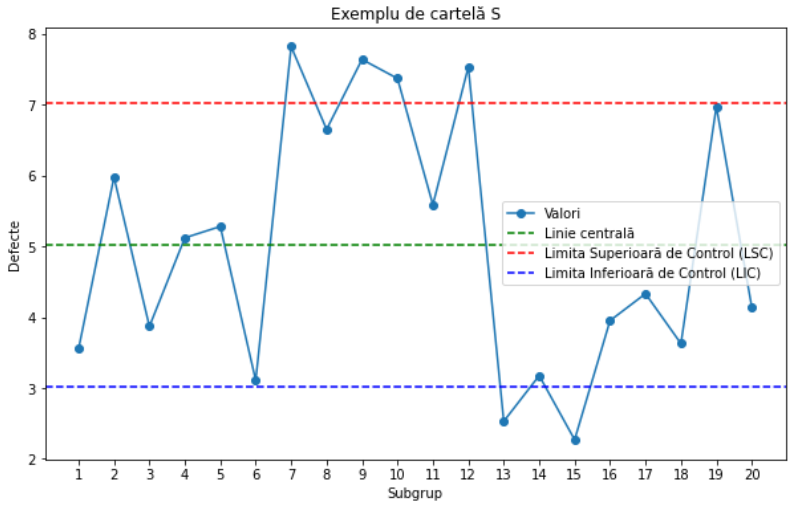


Fig. 9.15.Un exemplu de cartelă S

9.4.4. Interpretarea unei cartele de control

Cartelele de control sunt folosite pentru a identifica potențiale probleme în proces, având ca scop evitarea obținerii de rebuturi sau dereglări majore ale procesului. În interpretarea cartelelor de control există câteva reguli care ne ajută în identificarea potențialelor probleme în procesul pe care îl monitorizăm. În tabelul 9.1 sunt prezentate pe scurt aceste reguli.

Tabel 9.1. Detecția problemelor într-un proces cu ajurul cartelei de control

Regula	**Descriere**
1. Puncte înafara limitelor	Unul sau mai multe puncte sunt dincolo de limite
2. Test Zona A	2 din 3 puncte consecutive sunt în zona A sau mai departe
3. Test Zona B	4 din 5 puncte consecutive sunt în zona B sau mai departe
4. Test Zona C	7 sau mai multe puncte consecutive sunt de o singură parte a mediei (în Zona C sau mai departe)
5. Tendință	7 puncte consecutive au o tendință în sus sau în jos
6. Amestecare	8 puncte consecutive fără nici un punct în zona C
7. Stratificare	15 puncte consecutive în zona C
8. Supra-control	14 puncte consecutive alternând

Dacă avem punte înafara limitelor de control (Figura 9.16) acest lucru indică o problemă cu procesul monitorizat. Acest lucru se poate întâmpla datorită unor variații mari față de parametri normali și se pot datora configurării greșite a echipamentului de producție, erori de măsurare sau chiar prezența unui angajat nou. Un alt indicator al unei variații mari este dacă 2 din 3 puncte consecutive se află în zona A sau mai departe (Figura 9.16). Acesta se numește testul pentru zona A și poate indica omiterea unui pas de producție, defectarea unui echipament etc.

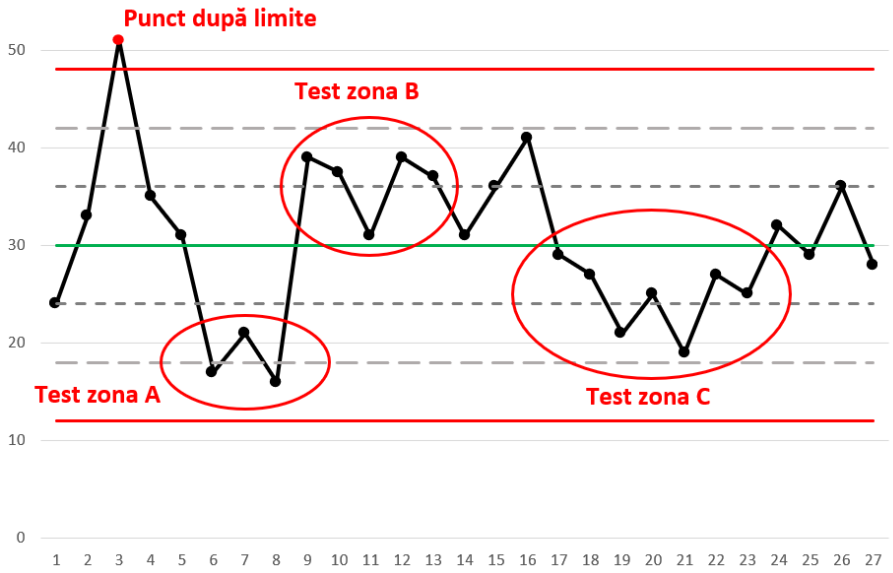


Fig. 9.16. Regula 1-4

Dacă observăm 4 din 5 puncte consecutive în zona B sau mai departe sau 7 sau mai multe puncte consecutive sunt de o singură parte a mediei (în Zona C sau mai departe), acestea pot indica variații mici și medii de la valoarea nominală. Acestea reprezintă testele pentru zona B respectiv C (Figura 9.16). Cauze posibile pot fi o modificare a instrucțiunilor de lucru, a dispozitivelor de măsură sau chiar a materialului.

Când pe cartelă observăm 7 puncte consecutive au o tendință în sus sau în jos (Figura 9.17) acest lucru poate indica efectele unui fenomen care evoluează gradual, cum ar fi uzura sculei sau încălzirea sculei de prelucrat.

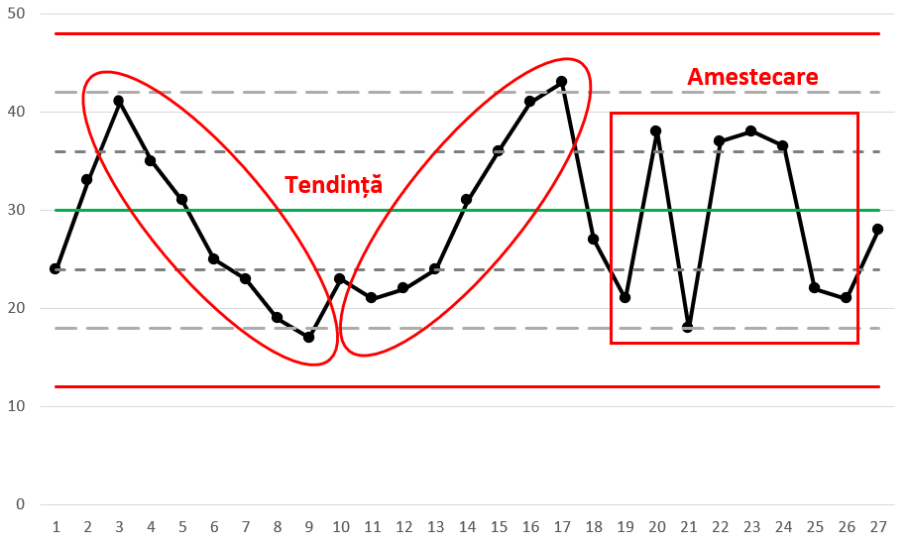


Fig. 9.17. Regula 5-6

Prezența a 8 puncte consecutive fără nici un punct în zona C indică amestecarea valorilor provenite din două procese distincte cum ar fi de la două mașini diferite sau de la folosirea a două materiale diferite (Figura 9.17). Aceleași cauze pot fi identificate atunci când avem 15 puncte consecutive în zona C. Acest fenomen se numește stratificare (Figura 9.18).

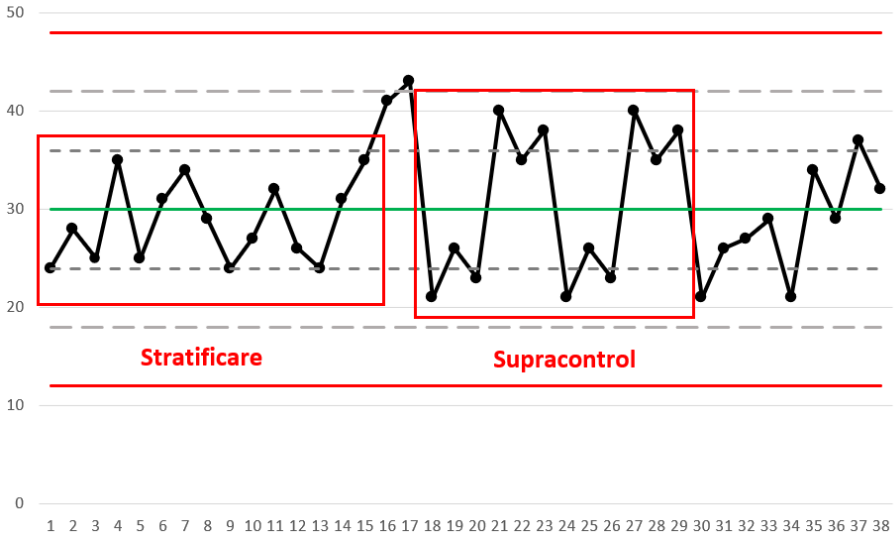


Fig. 9.18. Regula 7-8

Când avem 14 puncte consecutive alternând sau există anumite modele repetitive poate fi o indicație de supracontrol cauzată de manipularea datelor de către operator sau folosirea alternativă a mai multor materiale.

Cauzele posibile descrise pe scurt pentru regulile mai sus menționate sunt prezentate în tabelul 9.2

Tabel 9.2. Cauzele posibile ale regulilor observate în cartela de control

Descrierea manifestării	**Regulile**	**Cauze posibile**
Variații mari de la medie	1, 2	Angajat nou; configurare greșită; eroare de măsurare; a fost sărit un pas de producție; un pas nu a fost terminat; pană de curent; Echipament defect
Variații mici de la medie	3, 4	Schimbarea materialului; modificare instrucțiunilor de lucru; dispozitiv diferit de măsură; schimb de lucru diferit; îmbunătățirea abilităților muncitorului; schimbare în programul de mentenanță; schimbarea procedurii de instalare

Tendințe	5	Uzura sculei; efecte termice (răcire, încălzire)
Amestecare	6	Existența mai multor procese (schimburi, mașini, materiale)
Stratificare	7	Existența mai multor procese (schimburi, mașini, materiale)
Supracontrol	8	Manipularea datelor de către operator; Folosirea alternativă a mai multor materiale

Oricare ar fi modelele identificate cu ajutorul cartelei de control, un prim pas este oprirea procesului monitorizat și identificarea cauzelor. Odată identificate cauzele, procesul se poate relua.

9.5. Diagrame cauză-efect

Diagrama cauza-efect, de asemenea cunoscută sub numele de diagramă Ishikawa sau diagrama „os de pește”, este folosită pentru a determina cauza rădăcină pentru un anumit efect sau problemă (Figura 9.19). Fiecare cauză rădăcină este legată de coloana vertebrală a diagramei. Cauzele obișnuite pentru un proces tehnic sunt: mașină, metodă, măsură, materiale, oameni (forța de muncă) și mediu, dar acestea sunt pot fi adaptate procesului analizat. Fiecare cauză rădăcină are sub-cauze care, la rândul său, pot avea alte sub-cauze. Acest instrument ajută la vizualizarea modului în care diferite sisteme interacționează între ele și ce cauza rădăcină ar putea exista pentru un anumit efect.

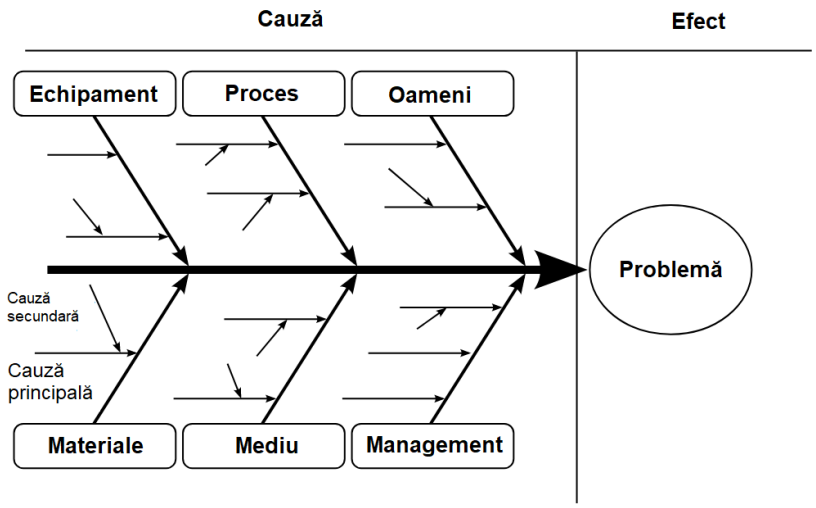


Fig. 9.19. Un exemplu de diagramă cauză-efect [20]

Spre exemplu, putem folosi diagrama Ishikawa pentru a determina care sunt cauzele pentru care produsele noastre prezintă un procent de defecte mai mare decât limita acceptată. Analizând procesul de producție din fiecare perspectivă în parte

(Mașină/Echipament, Materiale, Proces, Mediu, Personal și Management) putem identifica cauze principale, secundare și chiar terțiare care duc la rata mare de defecte.

Diagrama nu are ca scop identificarea unei singure cauze, ci oferă o imagine de ansamblu asupra potențialelor cauze și a influenței lor asupra efectului observat.

9.6. Diagrame de proces

Diagrama de proces (Flowchart) este un instrument care ajută la vizualizarea pașilor unui proces. Acesta ajută la o mai bună înțelegere a procesului și pentru a vedea modul în care diferitele lui părți interacționează (Figura 9.20).

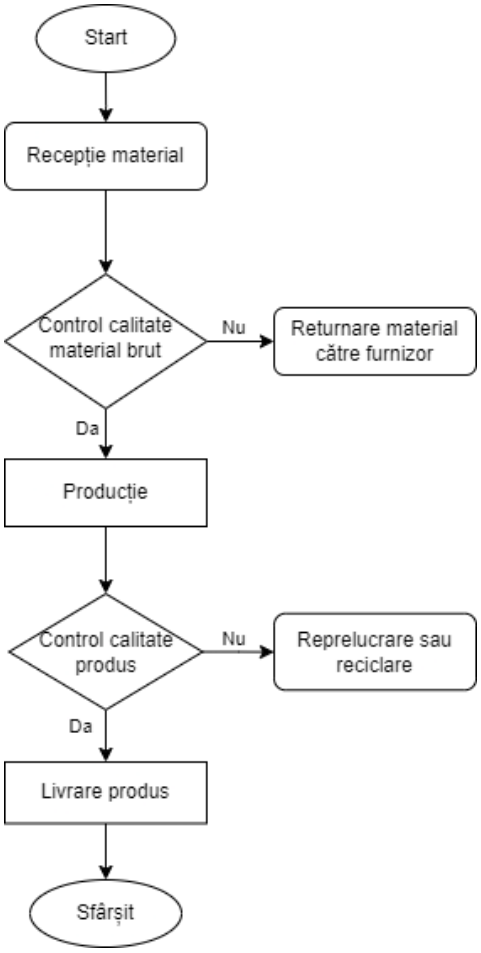


Fig. 9.20. Un exemplu de diagramă de proces

Fiecare proces poate fi la rândul său împărțit în sub-procese care pot avea propriile diagrame. Diagrama trebuie să fie simplă și ușor de înțeles, altfel nu își îndeplinește scopul de a oferi o imagine de ansamblu asupra procesului.

9.7. Verificarea cunoștințelor

1. Ce este o cartelă de control?
   a. Un instrument muzical
   b. Un grafic folosit pentru a urmări performanța unui proces
   c. O hartă geografică
   d. Un tip de diagramă de pescuit

2. Care este scopul principal al unui cartel de control?
   a. Îmbunătățirea esteticii datelor
   b. Monitorizarea și controlul calității proceselor
   c. Crearea de animații
   d. Niciuna dintre acestea

3. Ce reprezintă linia centrală (CL) într-o cartelă de control?
   a. Limita de control de sus
   b. Limita de control de jos
   c. Media procesului
   d. Variația procesului

4. Ce tip de date este cel mai potrivit pentru o cartelă de tip P?
   a. Date continue
   b. Date de tip atribut
   c. Date categorice
   d. Date complexe

5. În ce situație este util o cartelă de tip X-bar?
   a. Când se urmăresc defectele de pe o piesă
   b. Când se urmărește media subgrupurilor
   c. Când se urmărește variația între subgrupuri
   d. Toate cele de mai sus

6. Ce reprezintă LSC și LIC?
   a. Unitatea Centrală de Lucru
   b. Limita Sustenabilității Centrale și Limita de Inovare Considerată
   c. Limita Superioară de Control și Limita Inferioară de Control
   d. Niciuna dintre acestea

7. Care dintre următoarele este un indicator al unui proces instabil?
   a. Puncte aleatorii pe tot parcursul cartelei
   b. Puncte concentrându-se în jurul liniei centrale
   c. 7 puncte sau mai multe consecutiv deasupra sau sub linia centrală
   d. Toate punctele se află între limitele de control

8. Ce tip de cartelă este utilizată pentru a monitoriza numărul de defecte pe unitate?

a. Cartelă P
b. Cartelă C
c. Cartelă X-bar
d. Cartelă R

9. Ce tip de cartelă este utilizat pentru a monitoriza variația procesului?

a. Cartelă P
b. Cartelă X-bar
c. Cartelă U
d. Cartelă S

10. Care dintre următoarele afirmații este adevărată pentru o cartelă I-MR?

a. Utilizează media subgrupurilor
b. Utilizează datele individuale
c. Este folosit numai pentru date categorice
d. Niciuna dintre acestea

Răspunsuri Corecte

1. b. Un grafic folosit pentru a urmări performanța unui proces
2. b. Monitorizarea și controlul calității proceselor
3. c. Media procesului
4. b. Date de tip atribut
5. b. Când se urmărește media subgrupurilor
6. c. Limita Superioară de Control și Limita Inferioară de Control
7. c. 7 puncte sau mai multe consecutiv deasupra sau sub linia centrală
8. b. Cartelă C
9. d. Cartelă S
10. b. Utilizează datele individuale

10. Corelația și regresia

Acest capitol vă va ghida prin principiile de bază ale corelației și regresiei care sunt instrumente statistice puternice, va ilustra aplicarea lor cu exemple practice și va pune bazele unei modelări statistice mai avansate. Fie că preziceți tendințe economice, analizați date științifice sau pur și simplu încercați să înțelegeți lumea un pic mai bine, o înțelegere a corelației și regresiei este indispensabilă.

Corelația oferă o măsură cuantificabilă a modului în care variabilele se schimbă în tandem. Vom discuta despre coeficientul de corelație, o statistică care surprinde gradul în care două variabile sunt legate liniar. Acest concept este esențial pentru numeroase aplicații, de la finanțe, unde ajută la diversificarea portofoliului, până la sănătate, unde ajută la identificarea factorilor de risc pentru boli.

Regresia liniară ne ajută să modelăm relația dintre o variabilă independentă și o variabilă dependentă. Analiza de regresie ne permite nu doar să descriem asocierea, ci și să facem predicții, să controlăm variabilele de confuzie și chiar să deducem cauzalitatea în condițiile potrivite.

10.1. Corelația

Corelația Pearson este utilizată pentru a determina puterea și direcția unei relații liniare între două variabile continue [21]. Mai precis, testul folosește un coeficient numit **coeficient de corelație Pearson**, notat ca *r*. Valoarea acestui coeficient poate varia de la -1, pentru o relație liniară negativă perfectă, la +1, pentru o relație liniară pozitivă perfectă. O valoare de 0 (zero) indică că nu există nicio relație între cele două variabile. Corelația poate fi vizualizată cu ajutorul unui grafic cu puncte (Figura 10.1).

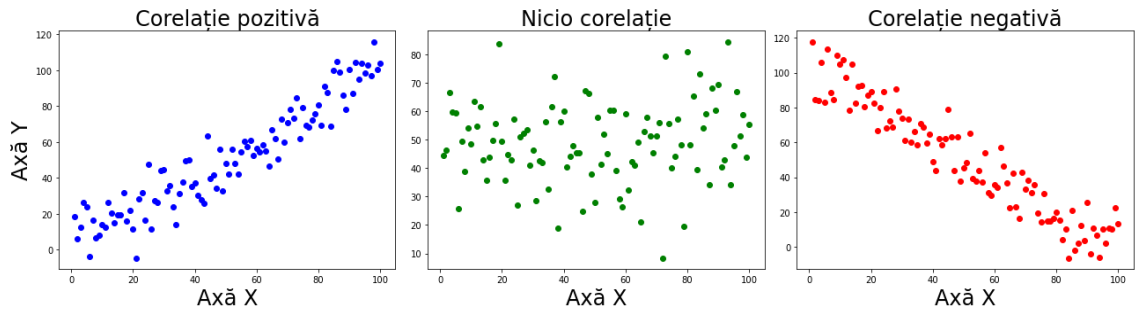


Fig. 10.1. Tipuri de corelații

În cazul corelației pozitive punctele sunt dispuse crescător. Cu cât punctele sunt mai puțin împrăștiate și tind să se dispună pe o dreaptă, cu atât coeficientul r se apropie de 1. În mod similar dacă punctele sunt dispuse descrescător, corelația este negativă.

Dacă nu putem distinge o tendință crescătoare sau descrescătoare, atunci nu avem corelație.

Pentru calculul corect al coeficientului de corelație trebuie îndeplinite niște condiții preliminare:

- Nivelul de măsură trebuie să fie interval sau rație pentru ambele variabile
- Variabilele trebuie să fie aproximativ normal distribuite
- Asocierea între cele două variabile trebui să fie liniară
- Datele ar trebui să nu conțină valori aberante

Spre exemplu, am putea investiga dacă există o legătură (sau corelație) între înălțimea și greutatea unor persoane dintr-un grup de interes. În acest caz, avem două variabile măsurate la pe o scală de tip rație. Dacă nu există factori perturbatori, de obicei înălțimea și greutatea sunt normal distribuite într-o populație. Un mod în care ne putem spori șansele de a avea date normal distribuite este de a avea eșantioane de dimensiuni mai mari (>30). Greutatea și înălțimea sunt de obicei liniar asociate. Persoanele mai înalte au tendința de a avea greutate mai mare iar cele mai scunde greutate mai mică. Bineînțeles că nu este o regulă strictă și putem avea și excepții. Valorile aberante pot fi eliminate prin mai multe metode.

Relația de calcul a coeficientului de corelație *r* este:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

unde:

x_i – valorile primei variabile

y_i – valorile celei de-a doua variabile

$\bar{x}$ – media valorilor primei variabile

$\bar{y}$ – media valorilor celei de-a doua variabile

Există mai multe tipuri de corelații (Spearman, Kendal etc) dar corelația Pearson este cea mai des folosită. Dacă sunt încălcate una sau mai multe dintre condițiile preliminare atunci se folosesc alte tipuri de corelație. Spre exemplu, dacă datele nu sunt normal distribuite sau dacă lucrăm cu date ordinale, putem folosi coeficientul de corelație Spearman care nu face presupuneri legate de distribuția datelor și folosește ranguri pentru calcul în locul valorilor absolute.

Este foarte important de înțeles că existența corelației nu implică cauzalitate. Altfel spus, dacă două variabile sunt corelate, nu înseamnă că una o determină pe cealaltă. Pot exista alți factori ascunși care să cauzeze corelația între variabilele analizate. Spre exemplu, dacă pe perioada verii avem mai multe internări de persoane

cu insolație pe de o parte iar pe de altă parte un consum mai mare de înghețată, deși putem spune că cele două variabile sunt corelate, nu putem spune că dacă consumăm mai multă înghețată facem insolație.

10.2. Regresia liniară

Regresia liniară simplă ne ajută să prezicem valoare unei variabile bazându-ne pe valoarea unei alte variabile [22]. Variabila pe care o prezicem se numește variabilă dependentă sau prezisă iar variabila folosită pentru prezicere se numește variabilă independentă sau predictoare. Folosim regresia liniară atunci când relația dintre cele două variabile este liniară

Putem folosi regresia pentru:

- a determina dacă relația liniară dintre două variabile este semnificativă statistic,
- a calcula cât din variația variabilei dependente este explicată de variabila independentă,
- a înțelege direcția și oricărei relații,
- a prezice valorile variabilelor dependente pe baza valorilor diferite ale variabilei independente.

Ideea de bază constă în a obține o dreaptă care se potrivește cel mai bine cu datele noastre. Linia care se potrivește cel mai bine pe date este cea pentru care distanțele dintre puncte și linia de regresie sunt minime. Distanța dintre fiecare punct și dreapta de regresie se numește eroarea totală de predicție.

Dreapta de regresie are ecuația:

$$Y = a * X + b$$

unde:

Y – variabila dependentă

X – variabila independentă

a, b – parametrii dreptei de regresie

Eroare de predicție se calculează făcând suma pătratelor diferențelor dintre valorile observate (din setul de date) și valorile prezise:

$$\Delta = \sum (y - \hat{y})^2$$

unde:

y – valorile observate

$\hat{y}$ – valorile prezise

Reprezentarea grafică a unei drepte de regresie pentru două variabile X și Y este reprezentată în figura 10.2.

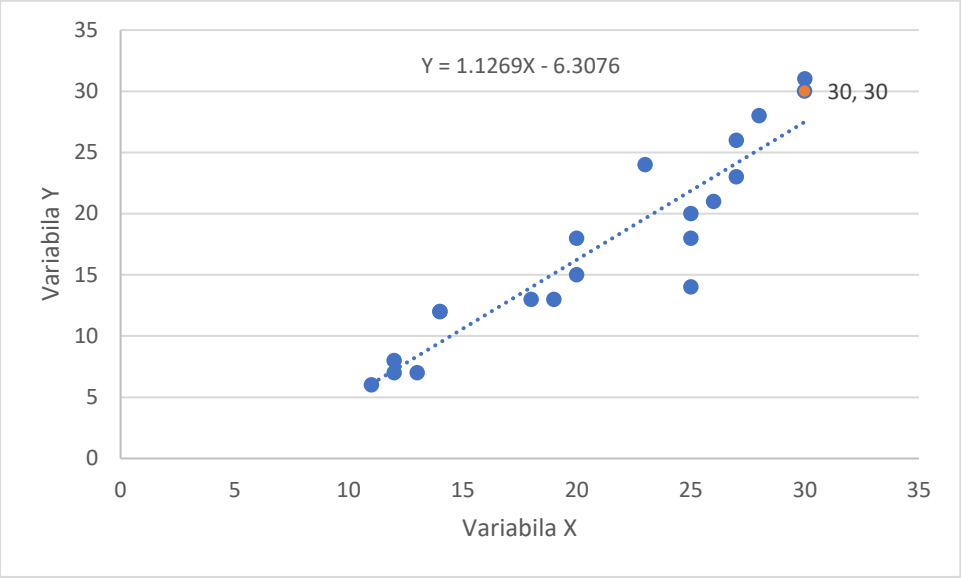


Fig. 10.2 Dreapta de regresie pentru două variabile X și Y.

Fiecare observație este reprezentată printr-un punct care are valorile variabilelor X și respectiv Y. Spre exemplu, punctul marcat cu roșu, are valorile X=30 și Y=30. Dreapta care se potrivește cel mai bine pe aceste date are ecuația:

$$Y = 1.1269 * X - 6.3076$$

Acesta se numește modelul de regresie. În acest model, coeficientul lui X se numește panta de regresie și ne indică variația medie a lui Y pentru o variație de o unitate a lui X. Coeficientul liber reprezintă valoarea lui Y atunci când X este 0 și se numește termenul liber.

Calitatea modelului obținut poate fi evaluat cu ajutorul coeficientului de determinare (R^2). Acesta este o măsură statistică a cât de aproape sunt datele de linia de regresie găsită. R^2 reprezintă procentul variației variabilei dependente (de răspuns) explicat de modelul liniar:

$$R^2 = \frac{Variație\ explicată}{Variație\ totală}$$

Acest indicator ia valori între 0 și 1, unde 0 indică faptul că modelul nu explică variabilitatea datelor de răspuns în jurul mediei sale iar 1 indică faptul că modelul explică toată variabilitatea datelor de răspuns în jurul valorii medii. În general, cu cât R^2 este mai mare, cu atât modelul se potrivește mai bine cu datele. Un model de calitate are valori mari ale coeficientului (0.8 – 0.9) dar pragul de calitate variază în funcție de domeniu, de aplicație etc.

Regresia liniară se bazează pe mai multe ipoteze cheie, care sunt necesare pentru obținerea unor rezultate de încredere:

1. **Liniaritate**: Relația dintre variabilele independente și variabila dependentă este liniară. Acest lucru poate fi verificat cu ajutorul graficelor cu puncte sau prin evaluarea coeficientului de corelație.
2. **Independența**: Observațiile sunt independente una de cealaltă. Aceasta este o ipoteză esențială în modelele de regresie care presupune că observațiile variabilei dependente sunt colectate fără nicio influență conexă.
3. **Homoscedasticitate**: Erorile prezintă o varianță constantă la fiecare nivel al unei variabile independente. Cu alte cuvinte, dispersia sau "împrăștierea" erorilor ar trebui să fie aproximativ aceeași la toate nivelurile variabilelor independente (homoscedasticitate), spre deosebire de a avea o dispersie care crește sau scade odată cu valorile ajustate (heteroscedasticitate).
4. **Normalitatea erorilor**: Erorile din model ar trebui să fie aproximativ normal distribuite. Această ipoteză nu este necesară pentru estimarea coeficienților în sine, deoarece metoda celor mai mici pătrate care estimează coeficienții este neparametrică. Cu toate acestea, normalitatea este necesară pentru construirea corectă a intervalelor de încredere și a testelor de ipoteză.
5. **Nu există sau există puțină multicoliniaritate**: Multicoliniaritatea apare atunci când variabilele independente sunt foarte corelate între ele. Acest lucru poate face dificilă determinarea efectului individual al fiecărei variabile independente asupra variabilei dependente din cauza redundanței informațiilor. Este de preferat să aveți variabile independente care nu sunt puternic corelate (coeficienți mai mari de 0.8-0.9).
6. **Nu există endogenitate**: Regresorii (X) nu trebuie să fie corelați cu termenul de eroare (ε). Endogeneitatea poate să apară din cauza prejudecății variabilelor omise, a erorii de măsurare sau a unui anumit tip de cauzalitate simultană între variabilele independente și cele dependente.
7. **Erorile sunt independente de predictori**: Erorile ε sunt necorelate cu variabilele predictive. Acest lucru garantează că predictorii au o influență consecventă și imparțială asupra variabilei dependente Y.

În cazul în care aceste ipoteze nu sunt îndeplinite, rezultatele analizei de regresie pot fi nesigure sau invalide. Este important să testăm aceste ipoteze și să aplicăm măsuri de remediere, după caz, care pot include transformarea variabilelor, adăugarea de variabile în model sau utilizarea unor tehnici de estimare alternative.

10.3. Verificarea cunoștințelor

1. Ce reprezintă panta într-o ecuație de regresie liniară simplă?
   a. Valoarea prezisă a lui Y atunci când X este zero
   b. Variația medie a lui Y pentru o variație de o unitate a lui X
   c. Corelația dintre X și Y

2. În contextul analizei de regresie, ce reprezintă R pătrat (R^2)?
   a. Proporția de variație a variabilei dependente care poate fi prezisă de variabila independentă.
   b. Varianța reziduurilor.
   c. Coeficientul de corelație dintre X și Y.

3. Care dintre următoarele indică cea mai puternică relație dintre două variabile într-un model de regresie liniară?
   a. $R^2 = -0.8$
   b. $R^2 = 0$
   c. $R^2 = 0.9$

4. Care este principalul scop al utilizării analizei de regresie?
   a. Pentru a descrie asocierea dintre variabile.
   b. Pentru a prezice valoarea unei variabile dependente pe baza valorii a cel puțin unei variabile independente.
   c. Pentru a dovedi relația cauză-efect între variabile.

5. Care dintre următoarele ipoteze NU este necesară pentru analiza regresiei liniare?
   a. Homoscedasticitate
   b. Distribuția normală a variabilelor
   c. Independența erorilor

6. Ce metodă este utilizată în mod obișnuit pentru a găsi linia cea mai bine adaptată într-o regresie liniară simplă?
   a. Estimarea celor mai mici pătrate
   b. Estimarea de maximă verosimilitate
   c. Estimarea modului

7. Ce înseamnă un coeficient de corelație zero?
   a. Corelație pozitivă perfectă
   b. Corelație negativă perfectă
   c. Nicio corelație

8. Dacă graficul cu puncte a două variabile formează o linie dreaptă perfectă, înclinată în jos, care este coeficientul de corelație?
   a. 0

b. 1
c. -1

9. Care dintre următoarele este o afirmație adevărată despre corelație?
   a. Corelația implică cauzalitatea.
   b. Corelația măsoară puterea și direcția unei relații liniare între două variabile.
   c. Corelația poate fi găsită numai în modele liniare.

10. Când nu poate fi utilizat coeficientul de corelație Pearson?
   a. Atunci când relația este neliniară.
   b. Atunci când variabilele se află pe scări diferite.
   c. Atunci când setul de date conține valori aberante.

Răspunsuri corecte

1. b. Variația medie a lui Y pentru o variație de o unitate a lui X

2. a. Proporția de variație a variabilei dependente care poate fi prezisă de variabila independentă.

3. c. $R^2 = 0,9$

4. b. Predicția valorii unei variabile dependente pe baza valorii a cel puțin unei variabile independente.

5. b. Distribuția normală a variabilelor

6. a. Estimare prin metoda celor mai mici pătrate

7. c. Nici o corelație

8. c. -1

9. b. Corelația măsoară puterea și direcția unei relații liniare între două variabile.

10. a. Când relația este neliniară.

11. Referințe

[1] Cambridge Dictionary, “data,” Cambridge Dictionary. Accesat: Jul. 24, 2022. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/data

[2] C. Zins, “Conceptual approaches for defining data, information, and knowledge,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 4, pp. 479–493, 2007, doi: 10.1002/asi.20508.

[3] R. L. Ackoff, “From data to wisdom,” *J. Appl. Syst. Anal.*, vol. 16, pp. 3–9, 1989.

[4] J. Rowley, “The wisdom hierarchy: representations of the DIKW hierarchy,” *J. Inf. Sci.*, vol. 33, no. 2, pp. 163–180, Apr. 2007, doi: 10.1177/0165551506070706.

[5] D. Chaffey and S. Wood, *Business Information Management: Improving Performance Using Information Systems by Chaffey, Dave, Wood, Steve (2004) Paperback*.

[6] “Definition of VARIABLE.” Accesat: Aug. 10, 2022. [Online]. Disponibil la: https://www.merriam-webster.com/dictionary/variable

[7] O.Theobald, *Statistics for Absolute Beginners: A Plain English Introduction*. Amazon Digital Services LLC - KDP Print US, 2017.

[8] C. Mckay, *Probability and Statistics*. Scientific e-Resources, 2019.

[9] Incnis Mrsi, “Discrete probability distribution illustration.” Wikimedia. Accesat: May 15, 2023. [Online]. Disponibil la: https://commons.wikimedia.org/wiki/File:Discrete probability distribution illustration.svg

[10] “Binomial distribution,” *Wikipedia*. May 06, 2023. Accesat: May 15, 2023. [Online]. Disponibil la: https://en.wikipedia.org/w/index.php?title=Binomial_distribution&oldid=1153481251

[11] “Hypergeometric distribution,” *Wikipedia*. May 15, 2023. Accesat: May 15, 2023. [Online]. Disponibil la: https://en.wikipedia.org/w/index.php?title=Hypergeometric_distribution&oldid=1154916959

[12] “Continuous uniform distribution,” *Wikipedia*. Mar. 13, 2023. Accesat: May 15, 2023. [Online]. Disponibil la: https://en.wikipedia.org/w/index.php?title=Continuous_uniform_distribution&oldid=1144473074

[13] “1.3.6.6.1. Normal Distribution.” Accesat: May 15, 2023. [Online]. Disponibil la: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm

[14] “Normal distribution - Wikipedia.” Accesat: May 15, 2023. [Online]. Disponibil la: https://en.wikipedia.org/wiki/Normal_distribution

[15] “Empirical Rule (68-95-99.7) Explained | Built In.” Accesat: May 15, 2023. [Online]. Disponibil la: https://builtin.com/data-science/empirical-rule

[16] “Student’s *t*-distribution,” *Wikipedia*. May 09, 2023. Accesat: May 15, 2023. [Online]. Disponibil la: https://en.wikipedia.org/w/index.php?title=Student%27s_t-distribution&oldid=1153912628

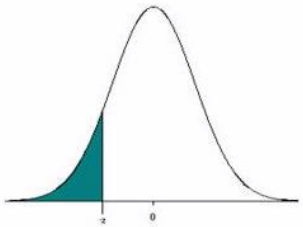
[17] “Chi-squared distribution - Wikipedia.” Accesat: May 15, 2023. [Online]. Disponibil la: https://en.wikipedia.org/wiki/Chi-squared_distribution

[18] L. A. Doty, *Statistical Process Control*, 2nd edition. New York: Industrial Press, Inc., 1996.

[19] S. Glen, “C Chart: Definition, Formulas,” Statistics How To. Accesat: Oct. 10, 2023. [Online]. Disponibil la: https://www.statisticshowto.com/c-chart/
[20] F. at de.wikipedia, *Ishikawa fishbone-type cause-and-effect diagram*. 2008. Accesat: Oct. 25, 2023. [Online]. Disponibil la: https://commons.wikimedia.org/wiki/File:Ishikawa_Fishbone_Diagram.svg
[21] “Pearson’s Product-Moment Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example.” Accesat: Oct. 25, 2023. [Online]. Disponibil la: https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php
[22] “Linear Regression Analysis in SPSS Statistics - Procedure, assumptions and reporting the output.” Accesat: Oct. 25, 2023. [Online]. Disponibil la: https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

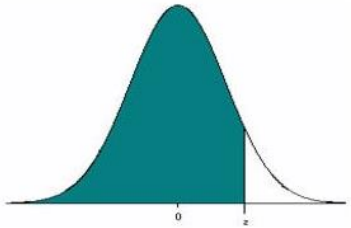
Anexa 1 - Tabel pentru distribuția normal (valori z)

Tabelul pentru valorile negative ale lui Z



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

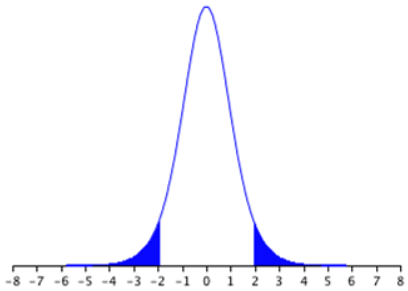
Tabelul pentru valorile pozitive ale lui Z



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Notă: Valorile probabilităților din acest tabel reprezintă aria din partea STÂNGĂ a valorii lui Z.
Pentru a calcula aria din DREAPTA trebuie calculat A_dreapta = 1 - A_stânga (A_stânga reprezentând aria din STÂNGA)

Anexa 2 - Tabelul distribuției Student (valori t)

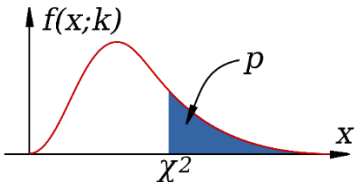


Nivelul de semnificație (α)

Grade de libertate	0.005 (1)	0.01 (1)	0.025 (1)	0.05 (1)	0.10 (1)	0.25 (1)
	0.01 (2)	0.02 (2)	0.05 (2)	0.10 (2)	0.20 (2)	0.50 (2)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.132	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.540	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.878	2.485	2.060	1.708	1.316	.684
26	2.779	2.479	2.056	1.706	1.315	.684
27	2.771	2.473	2.052	1.703	1.314	.684
28	2.763	2.467	2.048	1.701	1.313	.683
29	2.756	2.462	2.045	1.699	1.311	.683
≥30	2.575	2.327	1.960	1.645	1.282	.675

*** (1) - unilateral; (2) - bilateral simetric**

Anexa 3 - Tabelul distribuției Chi-pătrat (valori χ2)

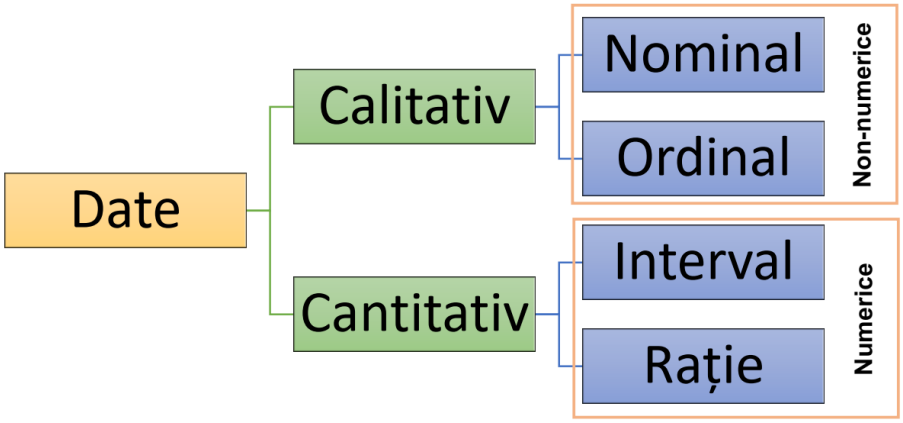


Grade de libertate	Aria de la dreapta valorii critice (χ^2)							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Anexa 4 – Sinteza noțiunilor

Statistică descriptivă

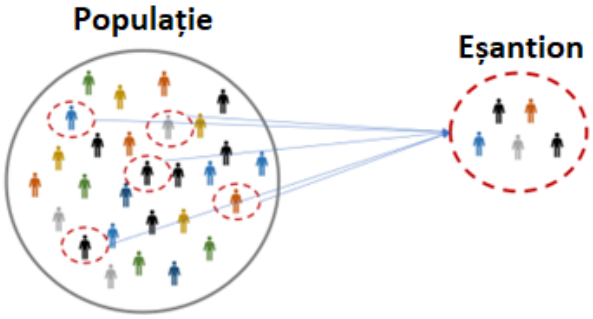
Niveluri de măsurare



Frecvența

Absolută	Relativă
Numărul de elemente	Proporția (procentul) de elemente din total $Frecvența\ relativă = \frac{Frecvența\ absolută}{Nr.total\ de\ elemente}$

Populație și eșantion

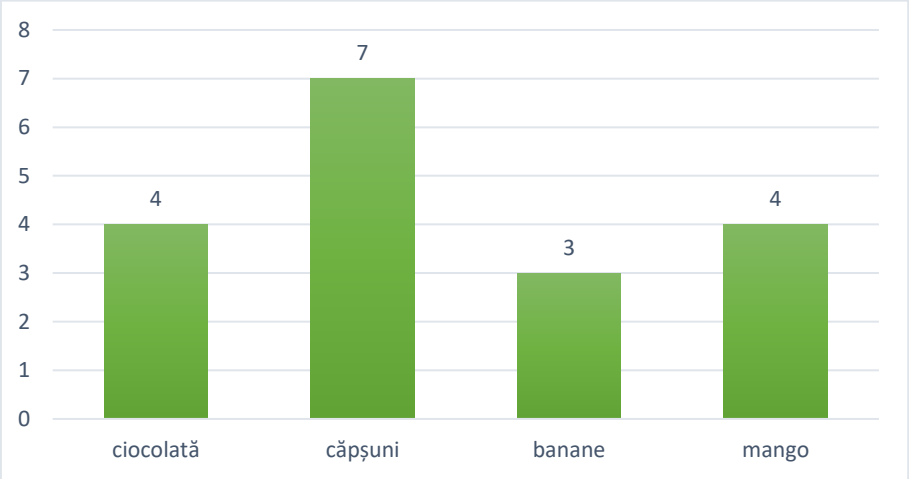


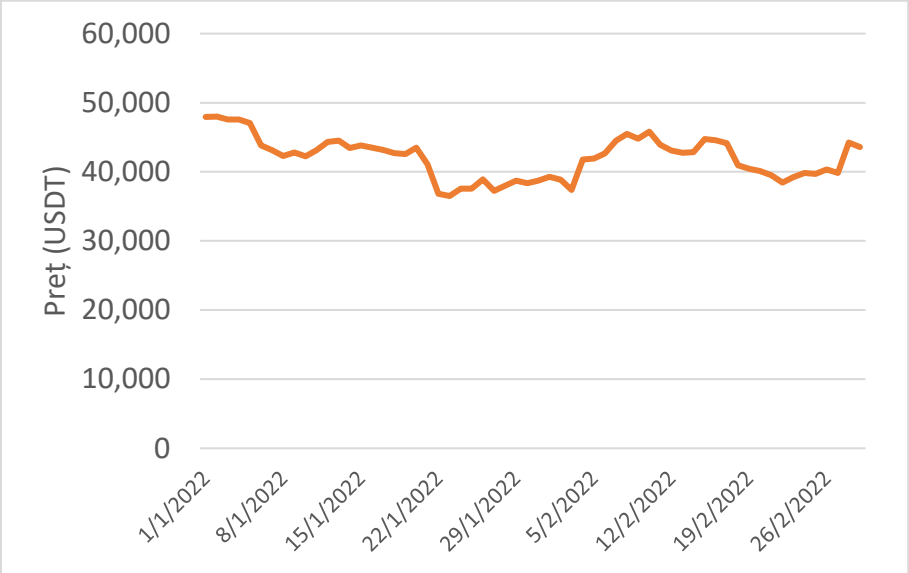
Parametri populației	**Caracteristicile eșantionului:**
μ – media populației	$\bar{x}$ – media eșantionului
σ^2 – dispersia populației	s^2 – dispersia eșantionului
σ – abaterea standard a populației	s – abaterea standard a eșantionului

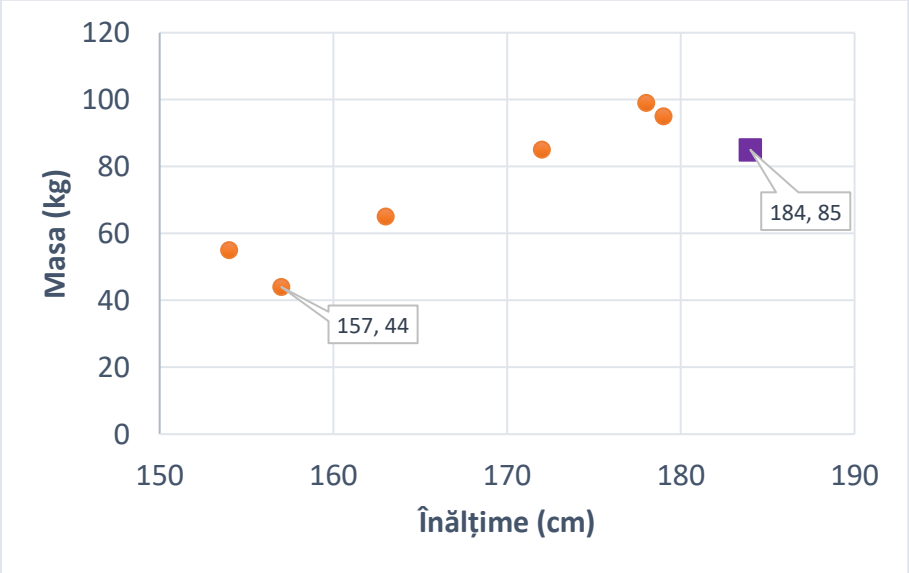
Indicatori

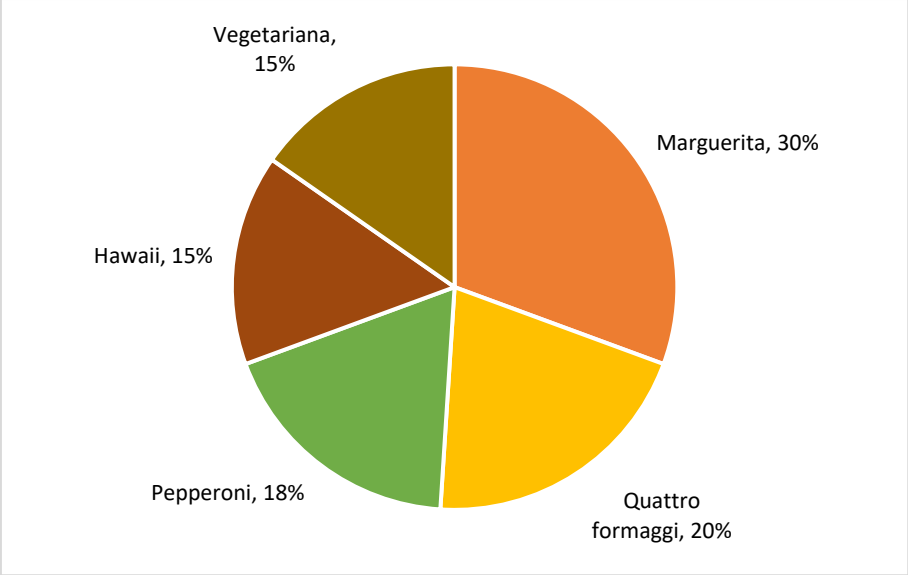
Măsură	Indicator	Formulă	
		Populație	**Eșantion**
Tendința centrală	Medie aritmetică	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
	Mediană	Valoarea din mijlocul șirului ordonat	
	Modală	Categoria/intervalul cu cea mai mare frecvență	
	Medie pătratică	$M_p = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2}{n}}$	
	Medie geometrică	$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$	
	Medie armonică	$M_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	
	Valoarea centrală	$x_c = \frac{Max - Min}{2}$	
Împrăștiere	Minim	Cea mai mică valoare din șir	
	Maxim	Cea mai mare valoare din șir	
	Amplitudine	$R = x_{max} - x_{min}$	
	Dispersie	$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
	Abatere standard	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$	$\mathrm{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
Formă	Asimetrie	$g_1 = \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3}$	
	Aplatizare	$k = \frac{\sum_{i=1}^{n}(x_i - \mu)^4}{(\sum_{i=1}^{n}(x_i - \mu)^2)^2}$	

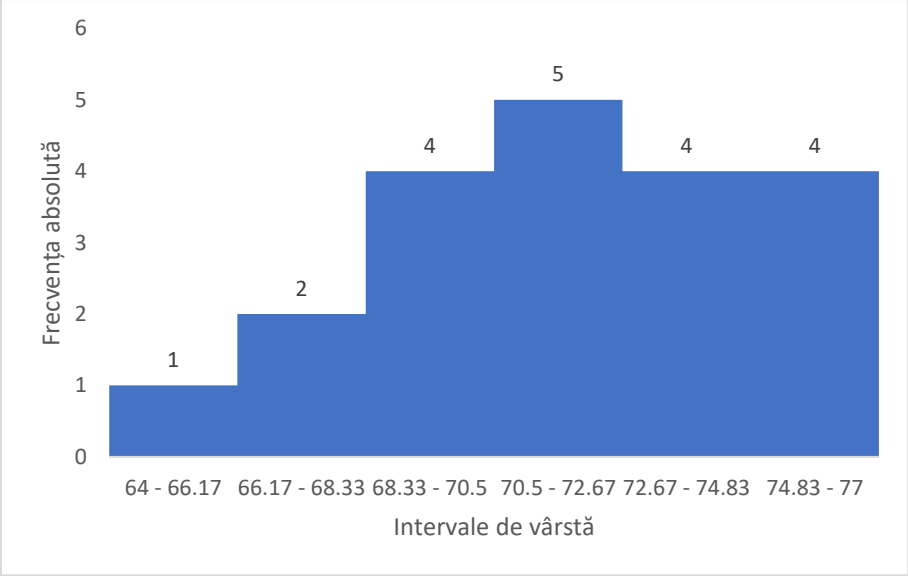
Vizualizări

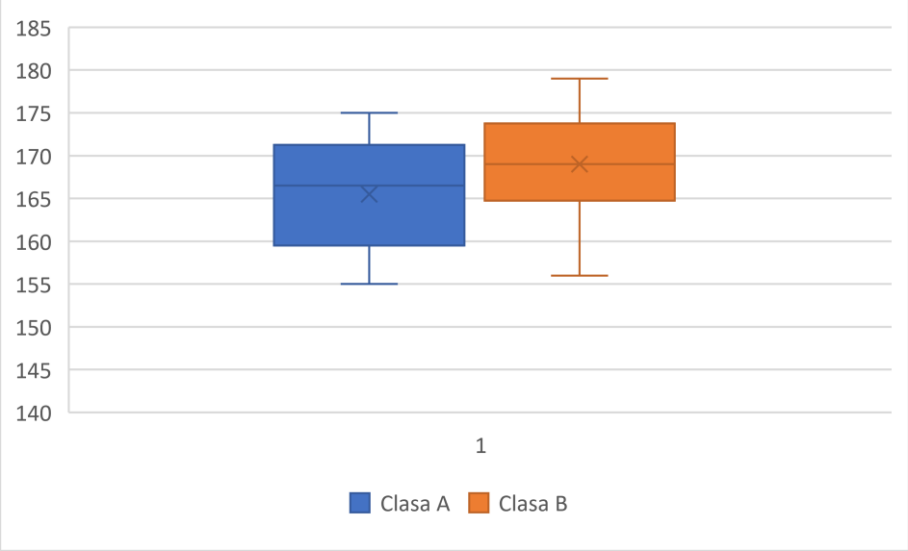







Vegetariana, 15%
Marguerita, 30%
Hawaii, 15%
Pepperoni, 18%
Quattro formaggi, 20%

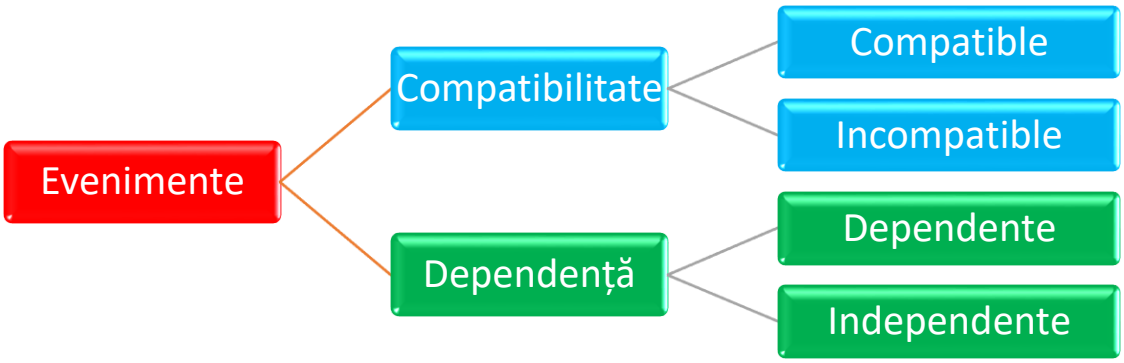

Frecvența absolută
0
1
2
3
4
5
6
1
2
4
5
4
4
64 - 66.17
66.17 - 68.33
68.33 - 70.5
70.5 - 72.67
72.67 - 74.83
74.83 - 77
Intervale de vârstă


185
180
175
170
165
160
155
150
145
140
1
Clasa A
Clasa B

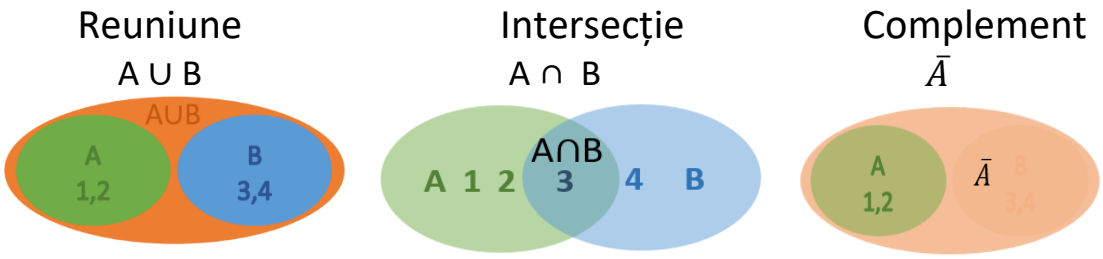
Probabilități

Tipuri de evenimente

- Imposibil: Ø
- Aleatoriu: A = {1, 2, 3}
- Sigur: E



Operațiuni cu evenimente



Probabilitate

$$P(X) = \frac{nr.evenimente\ favorabile}{nr.evenimente\ egal\ posibile}$$

Probabilitate condiționată: **P(B| A)** <- probabilitatea lui B, dat fiind că a avut loc A

Operație	Tipul evenimentului	Probabilitate
Reuniune	Incompatibil	$P(A \cup B) = P(A) + P(B)$
	Compatibil și independent	$P(A \cup B) = P(A) + P(B) - P(A) * P(B)$
	Compatibil și dependent	$P(A \cup B) = P(A) + P(B) - P(A) * P(B\|A)$
Intersecție	Compatibil și independent	$P(A \cap B) = P(A) * P(B)$
	Compatibil și dependent	$P(A \cap B) = P(A) * P(B\|A)$

Legile probabilității

Independența evenimentelor

$$P(B|A) = P(B)$$

Regula adunării (generalizată)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Regula înmulțirii

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n)$$

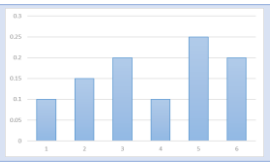
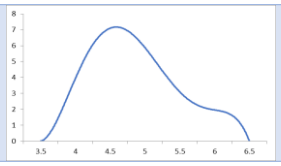
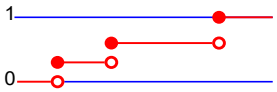
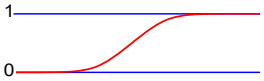
Legea probabilității totale

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

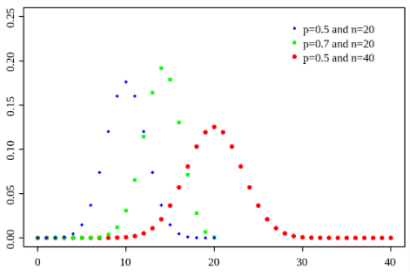
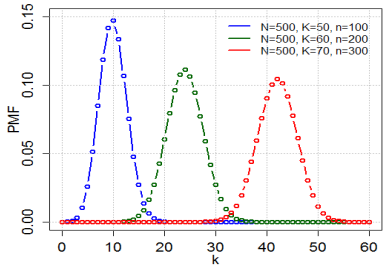
Regula lui Bayes

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Variabile aleatorii

	Variabile aleatoare discrete	Variabile aleatoare continue
Probabilitățile pot fi scrise într-un tabel	da	nu
Putem determina probabilitatea unei anumite valori	da	nu
Reprezentare grafică		
Funcția de probabilitate		

Distribuții discrete

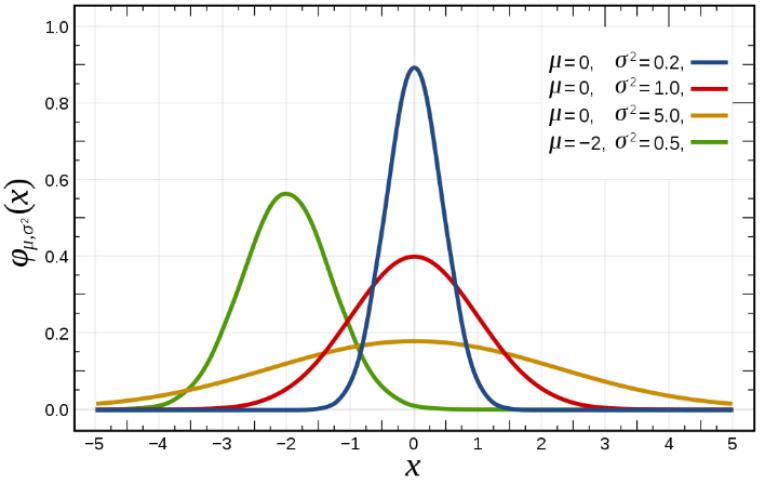
Distribuția binomială	***Distribuția hipergeometrică***
Numărul de succese la extragerea unui eșantion de dimensiune **n** dintr-o populație de dimensiunea **N,** <u>punând de fiecare dată piesa la loc.</u> **p** – probabilitatea succesului **q** – probabilitatea de eșec **n** – numărul de extrageri **Notă:** $$C_n^k = \frac{n!}{(n-k)! * k!}$$	Numărul de succese la extragerea unui eșantion de dimensiune **m** dintr-o populație de dimensiunea **n,** <u>fără a pune de fiecare dată piesa la loc.</u> **p** – probabilitatea succesului **q** – probabilitatea de eșec **m** – numărul de extrageri **a** – numărul de succese (a=n*p) **b** – numărul de eșecuri (b=n*q)
$$P(X = k) = C_n^k * p^k * q^{n-k}$$	$$P(X = k) = \frac{C_a^k * C_b^{m-k}}{C_n^m}$$
$$F(k) = P(X \leq k) = \sum C_n^k p^k q^{n-}$$	$$F(x) = P(X \leq k) = \frac{1}{C_n^m} \sum_{i=0}^{k} C_a^i * C_b^{m-}$$
	

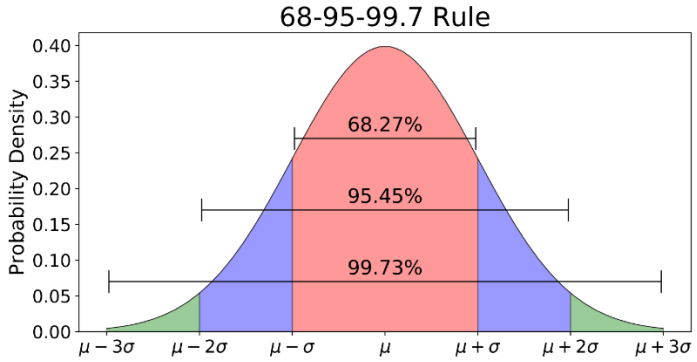
Distribuţii continue

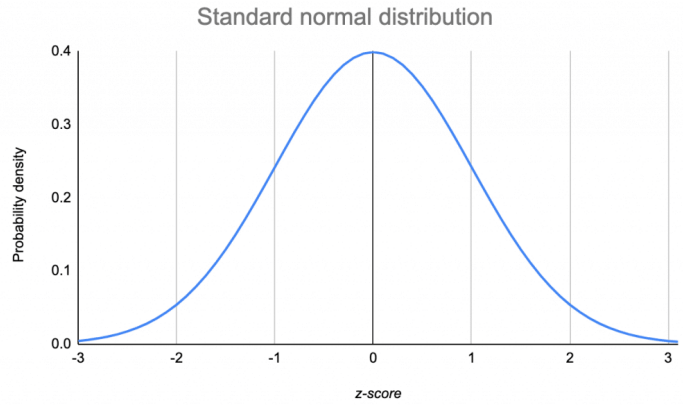
Distribuția normală

Parametri: media (μ) și abaterea standard (σ)

Simetrică în jurul mediei



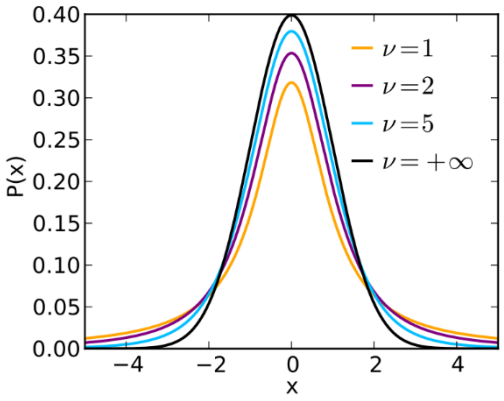




Distribuția Student

Parametru ν = n-1

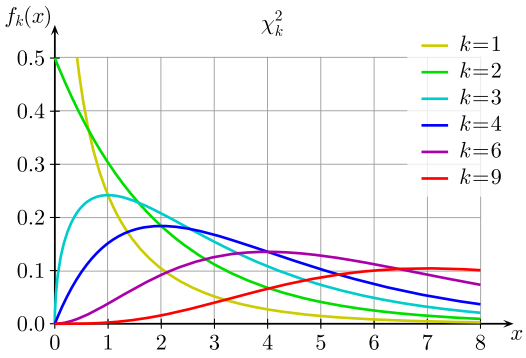
Simetrică în jurul mediei



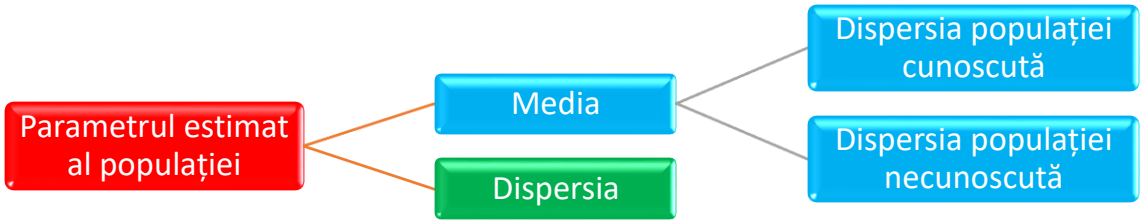
Distribuția Chi-pătrat (χ2)

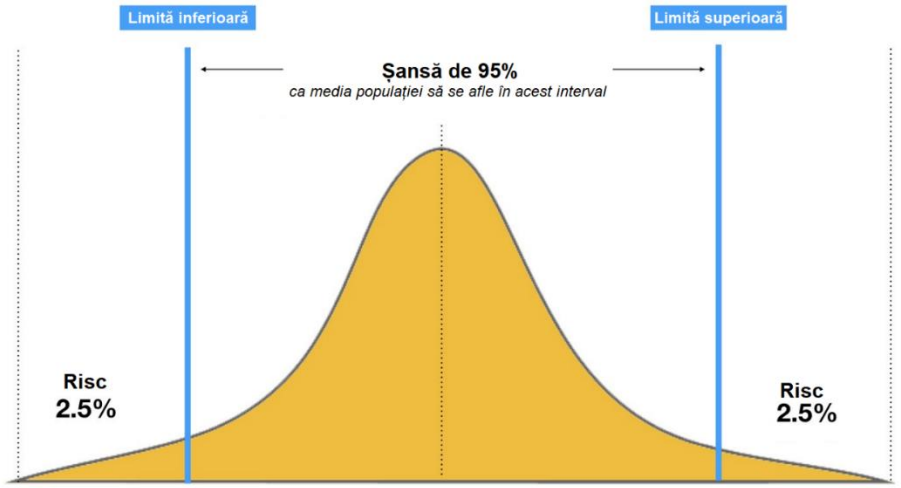
Parametrul k = n-1 (grade de libertate)

Asimetric(grade de libertate)ă și strict pozitiv

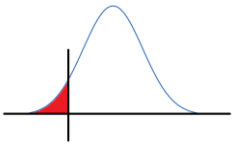
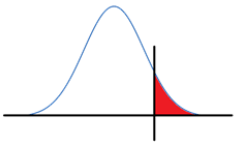
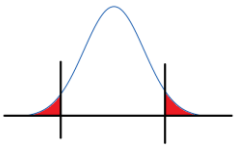
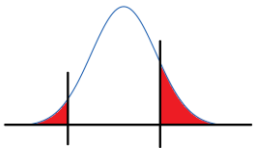


Estimarea

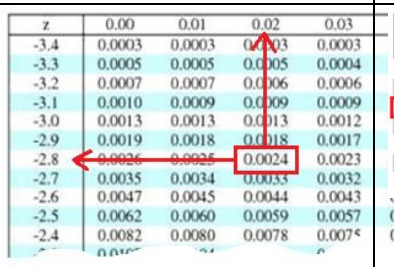
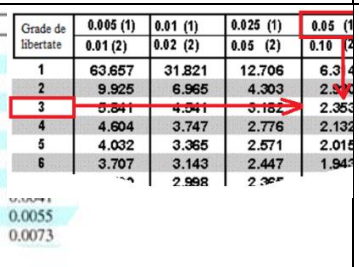
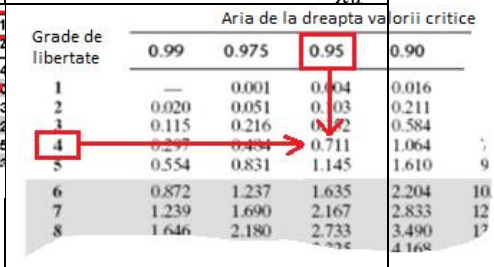




Tipuri de risc

Risc unilateral stânga(RUS)	Riscul unilateral dreapta (RUD)	Risc bilateral simetric (RBS)	Riscul bilateral asimetric (RBA)
			

Estimarea

<table>
<tr><th colspan="2">Estimarea mediei</th><th rowspan="2">Estimarea dispersiei</th></tr>
<tr><th>Dispersia este cunoscută</th><th>Dispersia necunoscută</th></tr>
<tr><td>Distribuția normală
(valori z)</td><td>Distribuția Student
(valori t)</td><td>Distribuție Chi-pătrat
(valori χ^2)</td></tr>
<tr><td>$\bar{x} - z_{\alpha\_st}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha\_dr}\frac{\sigma}{\sqrt{n}}$</td><td>$\bar{x} - t_{\alpha\_st}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha\_dr}\frac{s}{\sqrt{n}}$</td><td>$(n-1)\frac{s^2}{\chi^2_{1-\alpha}} < \sigma^2 < (n-1)\frac{s^2}{\chi^2_{\alpha}}$</td></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>Știind că: σ=1.5, $\bar{x}$=10.3, n=30

Estimați μ cu un risc unilateral stânga de 4%.

Din tabel z = -1.75

$10.3 - 1.75\frac{1.5}{\sqrt{30}} = 9.82$

Rezultat: μ este mai mare decât 9.82 cu o probabilitate de 96%. Există un risc de 4% ca μ să fie mai mic decât 9.82</td><td>Știind că: s=1.2, $\bar{x}$=5.1, n=20

Estimați μ cu un risc unilateral dreapta de 2.5%.
Grade de libertate ν = 20-1 = 19
Din tabel t = 2.093

$5.1 + 2.093\frac{1.2}{\sqrt{20}} = 5.66$

Rezultat: μ este mai mic decât 5.66 cu o probabilitate de 97.5%. Există un risc de 2.5 ca μ să fie mai mare decât 5.66</td><td>Știind că: s=2, $\bar{x}$=7, n=25
Estimați σ^2 cu un risc bilateral simetric de 10%.
Grade de libertate ν = 25-1 = 24
Risc stânga (1-α/2) = 5%
Risc dreapta (α/2) = 5%
Din tabel:
$\chi^2_{1-\alpha/2}$ = 13.848 $\chi^2_{\alpha/2}$ = 36.415
$Lim\_stânga = 24\frac{4}{36.415} = 2.63$
$Lim\_dreapta = 24\frac{4}{13.848} = 6.93$
Rezultat: σ^2 este între 2.63 și 6.93 cu o probabilitate de 90%. Există un risc de 10% ca σ^2 să fie înafara intervalului.</td></tr>
</table>

Controlul statistic al proceselor (SPC)

Instrumente SPC

- Histograma
- Diagrama Pareto
- Diagrama cu puncte
- Cartelele de control
- Diagrama cauză-efect
- Diagrama de proces

Histograma

Paşi în construirea unei histograme:

1. Determinați valorile minime și maxime și calculați amplitudinea
2. Împărțiți amplitudinea în numărul de intervale stabilit pentru a găsi lungimea intervalului
3. Utilizați lungimea intervalului pentru a determina capetele intervalului pentru fiecare interval
4. Numărați câte valori din șir sunt în fiecare interval
5. Utilizați o diagramă cu coloane pentru a vizualiza numărul de valori din fiecare interval.

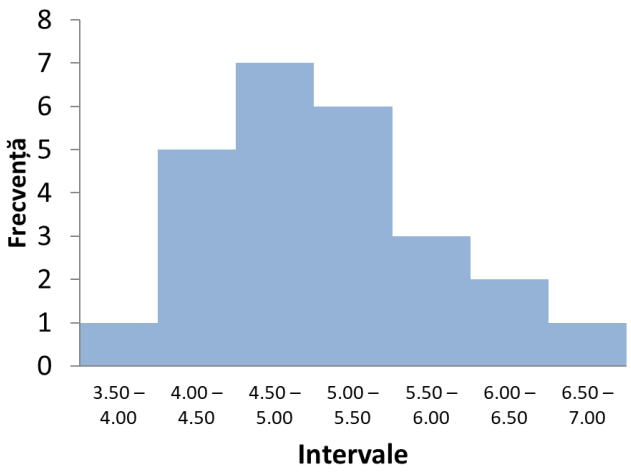


Diagrama Pareto

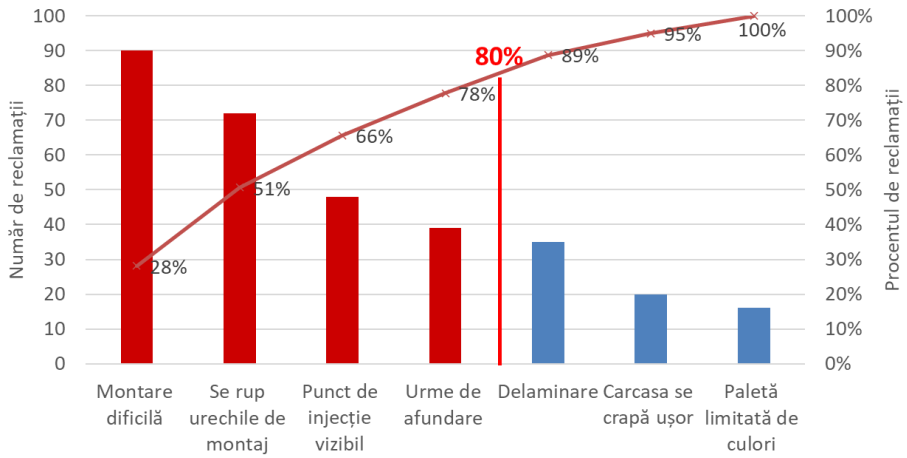
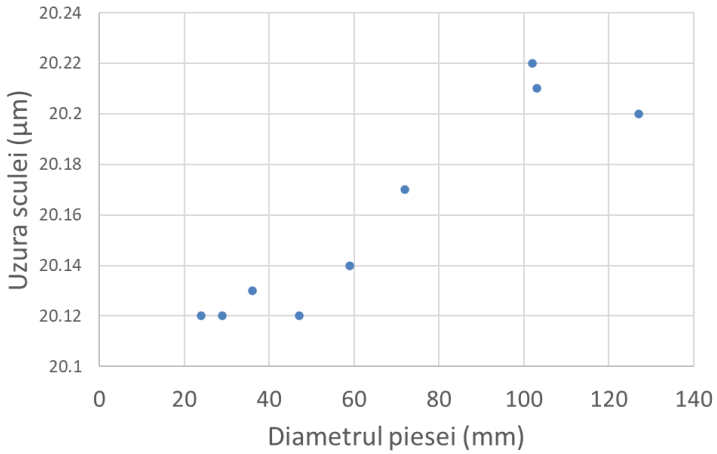
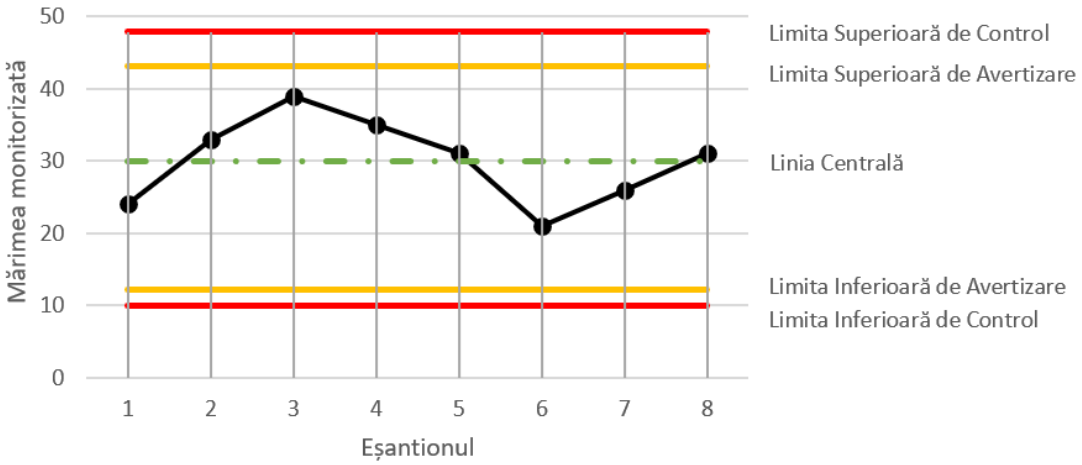


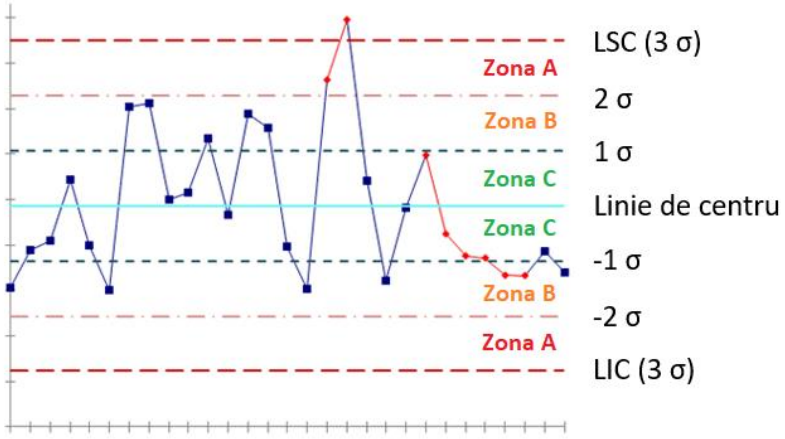
Diagrama cu puncte



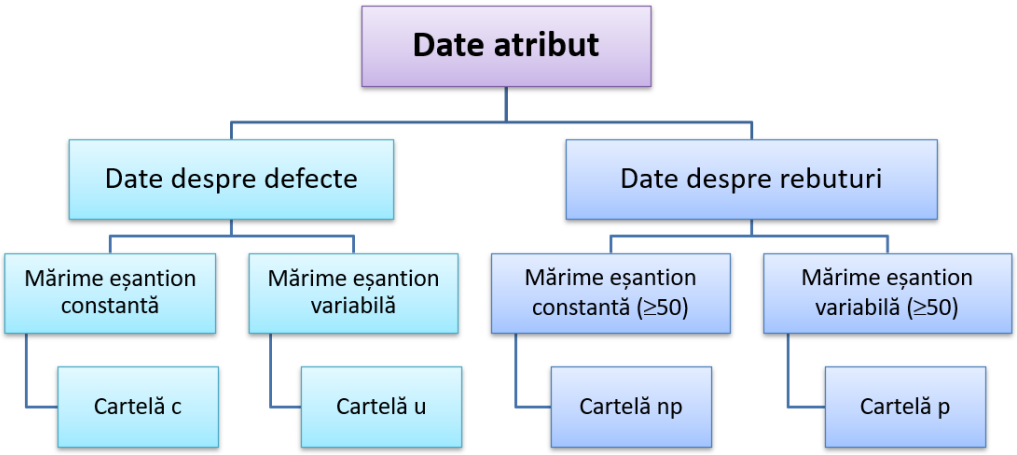
Cartele de control

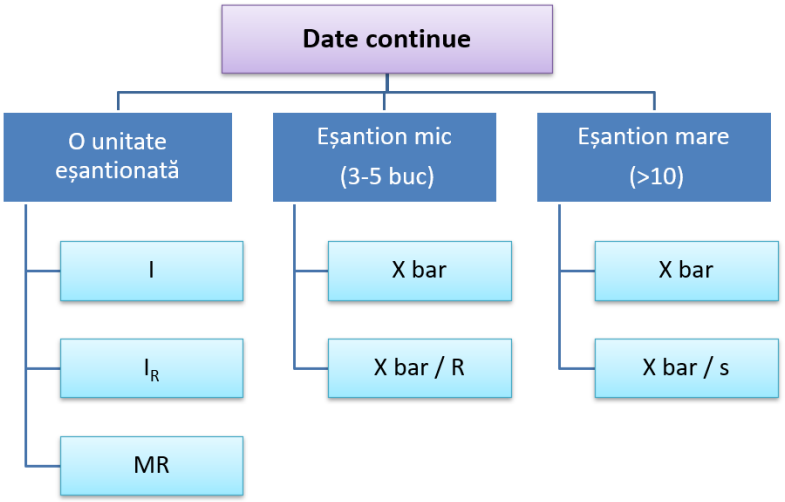
Elementele unei cartele de control





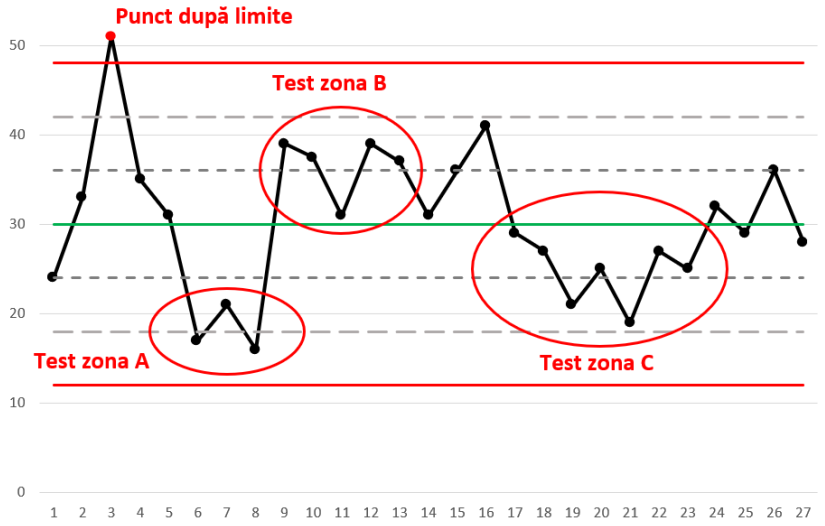
Tipuri de cartela de control

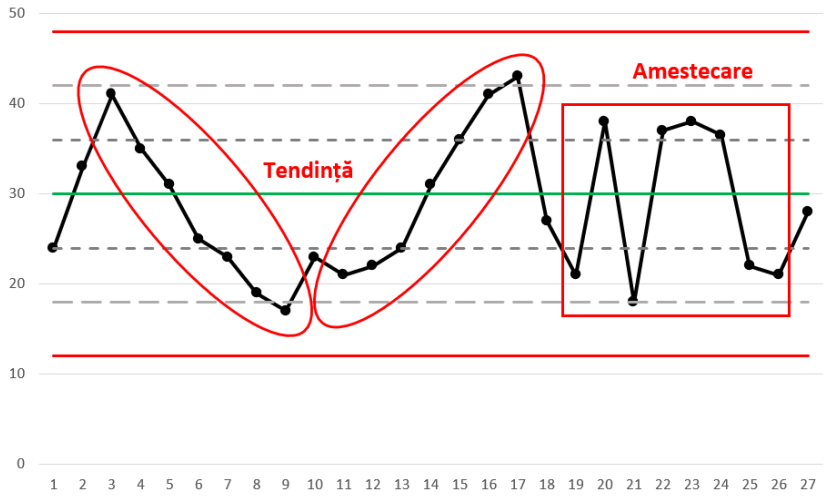


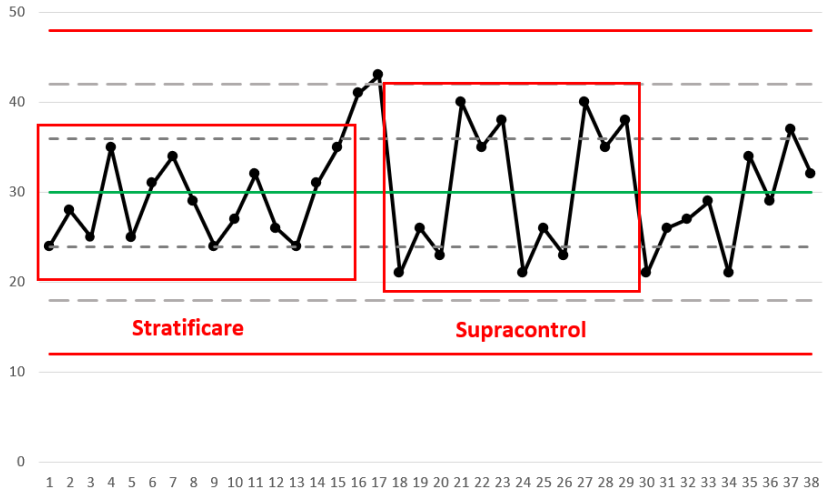


Cum să identificați problemele cu procesul

Regula	Descriere
1. Puncte înafara limitelor	Unul sau mai multe puncte sunt dincolo de limite
2. Test Zona A	2 din 3 puncte consecutive sunt în zona A sau mai departe
3. Test Zona B	4 din 5 puncte consecutive sunt în zona B sau mai departe
4. Test Zona C	7 sau mai multe puncte consecutive sunt de o singură parte a mediei (în Zona C sau mai departe)
5. Tendință	7 puncte consecutive au o tendință în sus sau în jos
6. Amestecare	8 puncte consecutive fără nici un punct în zona C
7. Stratificare	15 puncte consecutive în zona C
8. Supra-control	14 puncte consecutive alternând

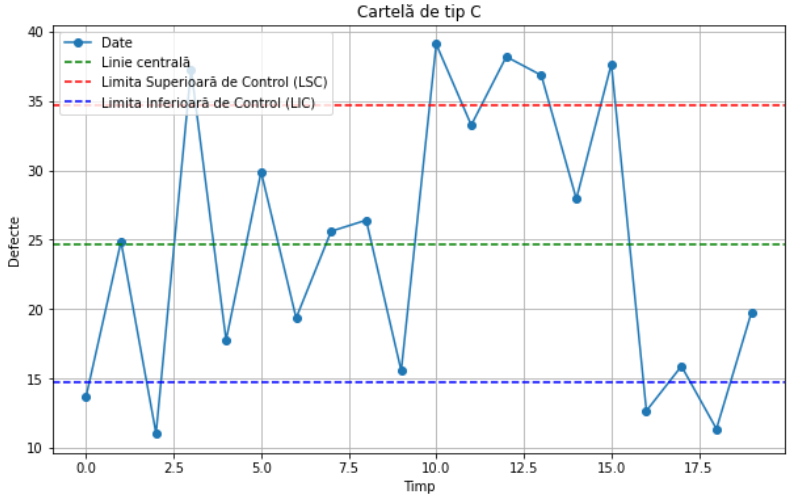


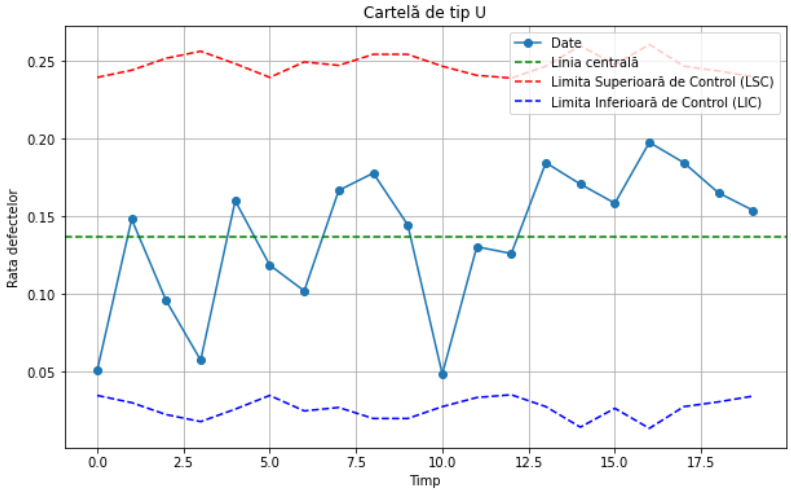


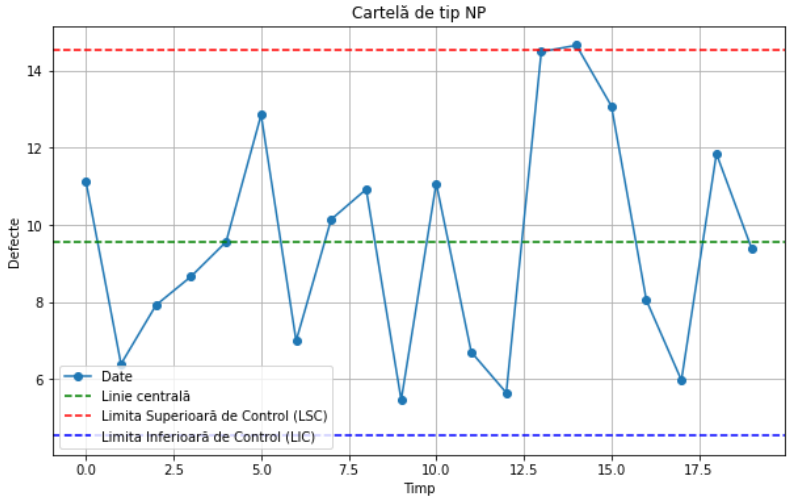


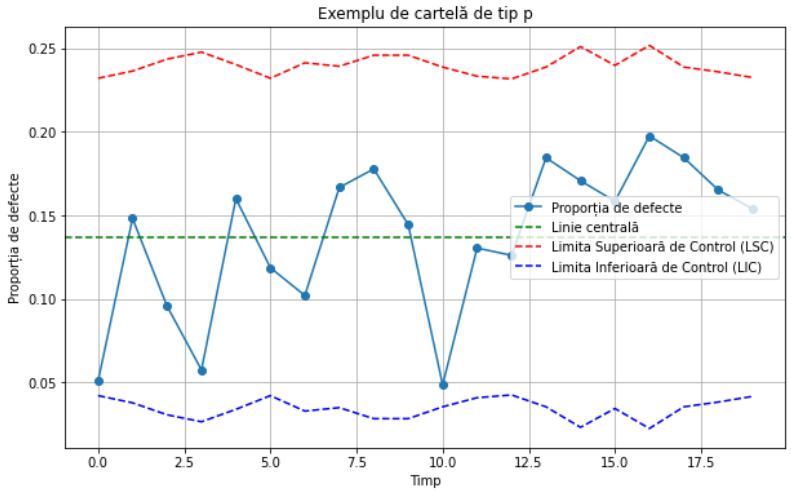
Descrierea manifestării	Regulile	Cauze posibile
Variații mari de la medie	1, 2	Angajat nou; configurare greșită; eroare de măsurare; a fost sărit un pas de producție; un pas nu a fost terminat; pană de curent; Echipament defect
Variații mici de la medie	3, 4	Schimbarea materialului; modificare instrucțiunilor de lucru; dispozitiv diferit de măsură; schimb de lucru diferit; îmbunătățirea abilităților muncitorului; schimbare în programul de mentenanță; schimbarea procedurii de instalare
Tendințe	5	Uzura sculei; efecte termice (răcire, încălzire)
Amestecare	6	Existența mai multor procese (schimburi, mașini, materiale)
Stratificare	7	Existența mai multor procese (schimburi, mașini, materiale)
Supracontrol	8	Manipularea datelor de către operator; Folosirea alternativă a mai multor materiale

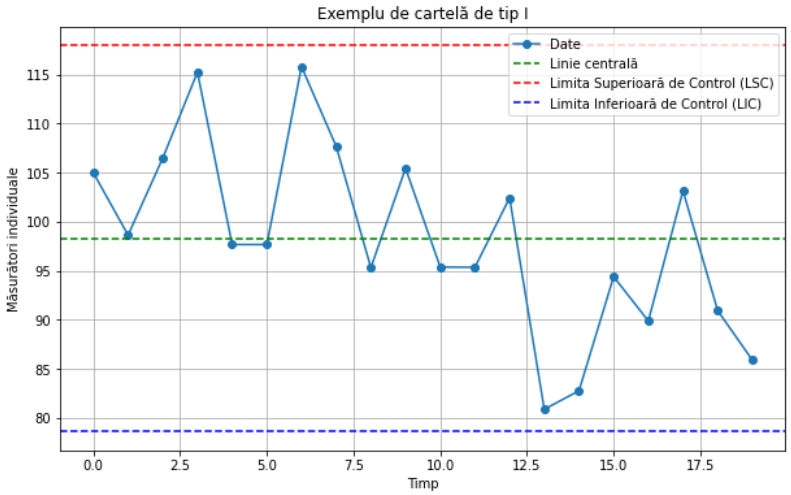
Tipuri de cartele de control

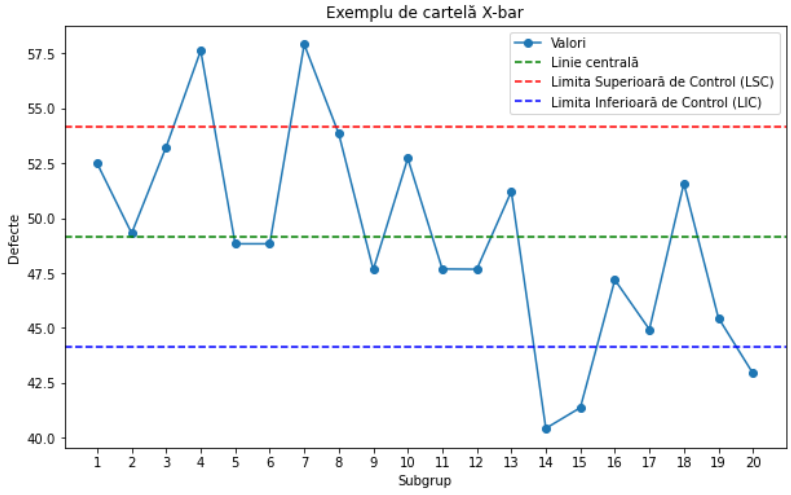


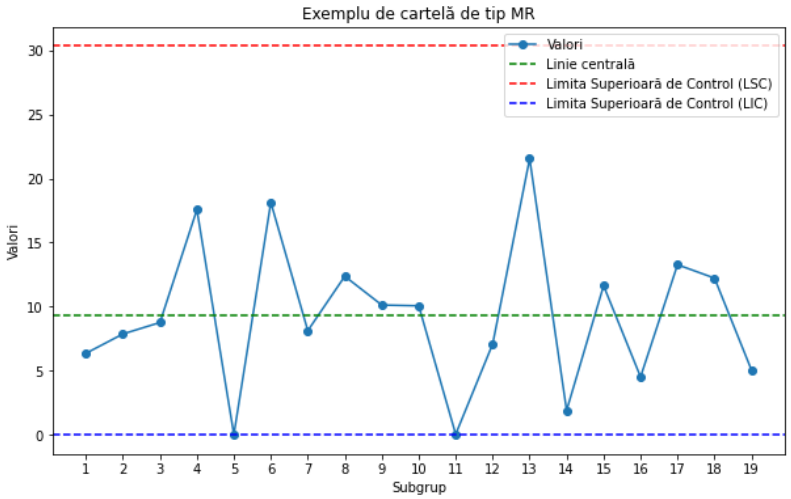


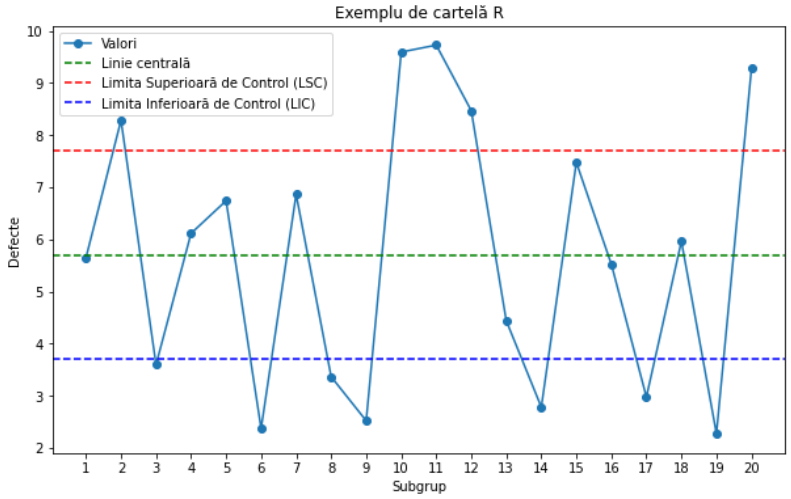



Exemplu de cartelă de tip p
Proporția de defecte
Timp
Proporția de defecte
Linie centrală
Limita Superioară de Control (LSC)
Limita Inferioară de Control (LIC)


Exemplu de cartelă de tip I
Măsurători individuale
Timp
Date
Linie centrală
Limita Superioară de Control (LSC)
Limita Inferioară de Control (LIC)


Exemplu de cartelă X-bar
Defecte
Subgrup
Valori
Linie centrală
Limita Superioară de Control (LSC)
Limita Inferioară de Control (LIC)


Exemplu de cartelă de tip MR
Valori
Linie centrală
Limita Superioară de Control (LSC)
Limita Superioară de Control (LIC)
Valori
Subgrup


Exemplu de cartelă R
Valori
Linie centrală
Limita Superioară de Control (LSC)
Limita Inferioară de Control (LIC)
Defecte
Subgrup

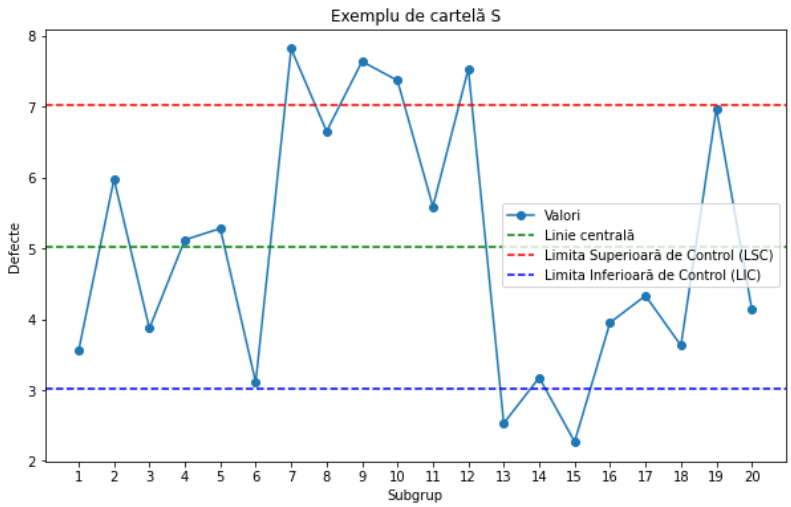

Exemplu de cartelă S
Valori
Linie centrală
Limita Superioară de Control (LSC)
Limita Inferioară de Control (LIC)
Defecte
Subgrup

Diagrama Ishikawa

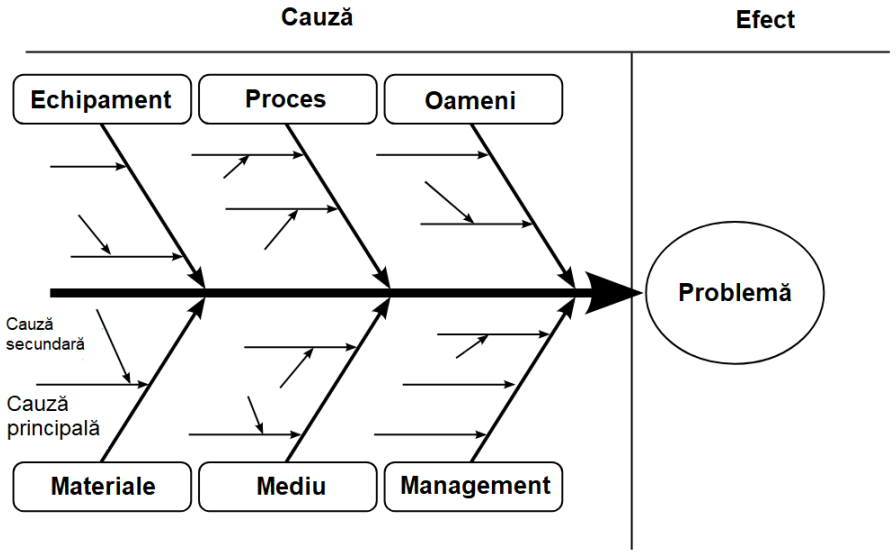
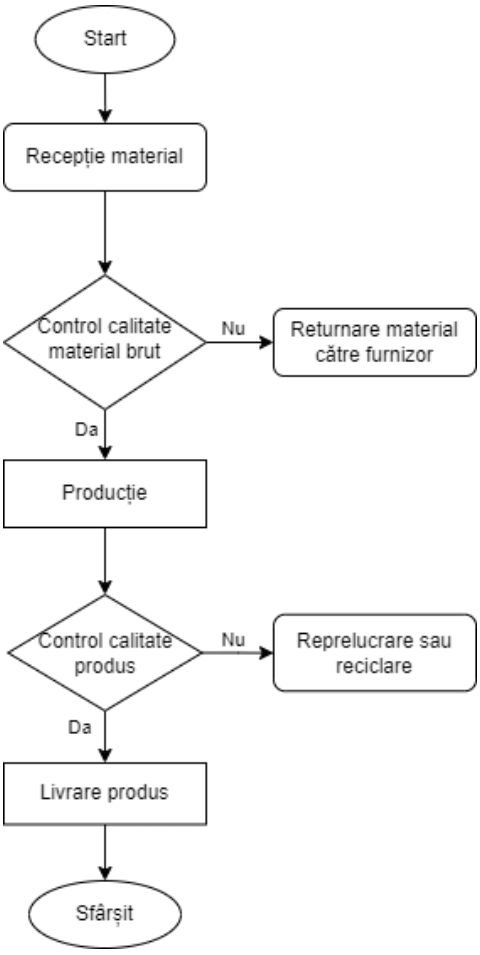


Diagrama fluxului de proces



Corelație și regresie

Coeficientul de corelație Pearson

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Ecuația liniei de regresie:

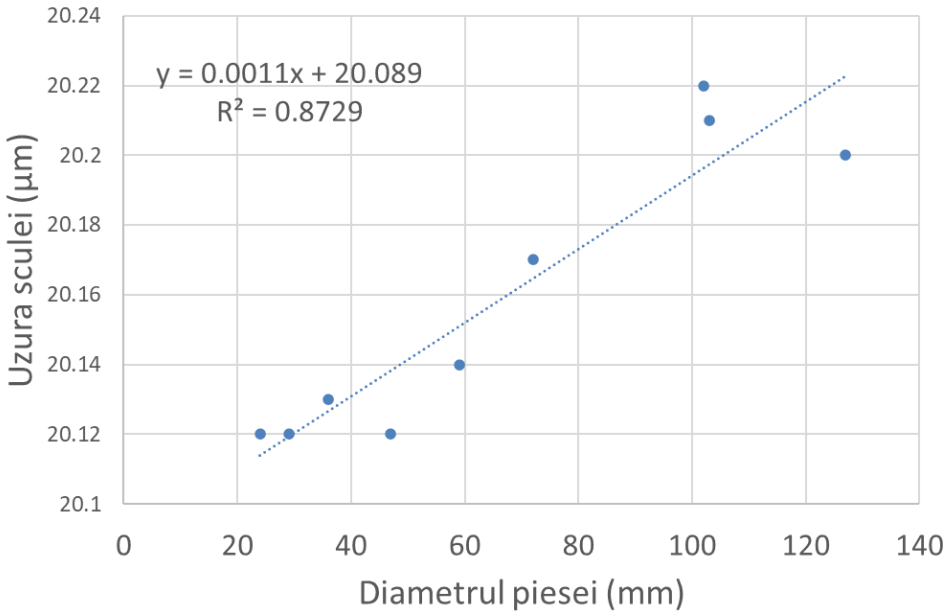
$$Y = a * X + b$$

Eroare de predicție:

$$\Delta = \sum (y - \hat{y})^2$$

Coeficientul de determinare (R^2):

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$



Index de termeni

Notație	Denumire / Descriere
a	Număr de succese
A	Amplitudine
ABR	Risc bilateral asimetric
ai	Frecvența absolută
α (alpha)	Riscul
b	Număr de eșecuri
χ^2 (chi pătrat)	Statistica pentru distribuția χ^2
f_i	Frecvența relativă
Ø (fi)	Mulțimea vidă / Elementul imposibil
IQR	Intervalul intercuartilic
k (distribuția Student)	Grade de libertate
k (distribuții discrete)	Observația curentă
m (distribuția hipergeometrică)	Număr de încercări/extrageri
M_a	Media armonică
M_e	Mediana
M_g	Media geometrică
μ	Media populației
M_o	Modala
M_p	Media pătratică
M_x	Media aritmetică (general)
n (distribuția binomială)	Număr de încercări/extrageri
n (distribuția hipergeometrică)	Numărul total de elemente
v (niu)	Grade de libertate
p	Probabilitatea de succes la o încercare
P(A|B)	Probabilitatea lui A dat fiind B
P(X)	Probabilitatea lui X
q	Probabilitatea de eșec la o încercare
Q1, Q2, Q3	Prima, a doua, a treia cuartilă
r	Coeficientul de corelație Pearson
R^2	Coeficientul de determinare
RBS	Risc bilateral simetric
RUD	Risc unilateral dreapta
RUS	Risc unilateral stânga
s	Abaterea standard (eșantion)
S	Câmpul de evenimente
s^2	Dispersia (eșantion)
σ (sigma)	Abaterea standard (populație)
σ^2	Dispersia (populație)
t	Statistica pentru distribuția Student
$\bar{x}$ (x_bar)	Media eșantionului
x_{max}	Valoarea maximă
x_{min}	Valoarea minimă
z	Statistica pentru distribuția Normală

Listă figuri

Listă tabele