UNIVERSITATEA DE MEDICINĂ,
FARMACIE, ŞTIINŢE ŞI TEHNOLOGIE
„GEORGE EMIL PALADE”
DIN TÂRGU MUREŞ

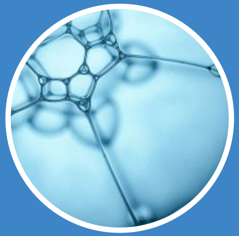
PROBABILITĂȚI ȘI STATISTICĂ ÎN SISTEME MEDICALE

Cursul 2, 23.09.2020

STATISTICA DESCRIPTIVĂ.

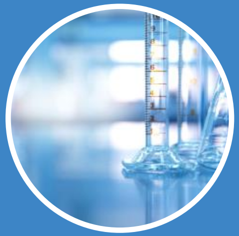
prof. univ. dr. habil Manuela Rozalia GABOR

2. STRUCTURA CURSULUI



2.1.Distribuții de frecvențe

2.2. Indicatorii tendinței centrale



2.3. Indicatori ai dispersiei

CONSIDERAȚII GENERALE

Principalii indicatori care caracterizează o serie de date (un set de valori ale unui eşantion statistic) sunt fie *indicatori ai tendinţei centrale*, fie ***indicatori ce caracterizează împrăştierea datelor în jurul unei valori medii*** (***indicatori ai dispersiei***).

Deoarece aceşti indicatori "descriu" din punct de vedere statistic distribuţia datelor studiate, permiţând chiar unele comparaţii ale acesteia cu distibuţia normală, modalitățile de utilizare ale acestor indicatori statistici fac obiectul unei ramuri a statisticii denumită **statistică descriptivă**.

O serie de date este alcătuită dintr-un şir de valori pe care le notăm :

$$x1, x2, \ldots, xn .$$

Indicatorii matematici mai importanţi ce caracterizează o serie de date sunt:

2.1. DISTRIBUȚII DE FRECVENȚE

Repartizarea datelor calitative şi a celor cantitative (numerice) dintro populaţie statistică sau un eşantion se poate efectua după frecvenţa de apariţie a caracteristicilor lor, obţinându-se structura colectivităţii. De

exemplu, o mulţime de date experimentale poate fi repartizată după “calitatea” efectelor observate: cu efect, fără efect sau cu efect gradat în

funcţie de doză.

Datele (măsurate pe întreaga populaţie statistică sau pe un eşantion) sunt de obicei organizate în aşa-numitele distribuţii de frecvenţă, pentru că o atare prezentare, reprezintă formatul cel mai convenabil de sinteză şi prezentare a acestora.

În cazul distribuţiilor de frecvenţă se afectuează o descriere calitativă sau cantitativă a observaţiilor (măsurătorilor) împreună cu numărul de apariţii ale unui anumit rezultat al măsurătorii respective – frecvenţa absolută.

Se mai poate utiliza şi frecvenţa relativă obţinută prin împărţirea

frecvenţei absolute la numărul total al observaţiilor.

De asemenea, pentru variabilele cantitative (numerice) continue este util să se construiască aşa-numitele ***intervale de variaţie***. Regulile generale de construire a acestor intervale sunt:

- numărul de intervale este bine să fie mai mic de 15
- limitele fiecărei clase să se potriveasca cu gradul de acurateţe (precizie, exactitate) al măsurătorilor
- sunt de preferat intervalele de lungime egală, deoarece sunt mai
- convenabile, facilitând prelucrările ulterioare
- intervalele trebuie sa fie mutual exclusive (capetele lor nu trebuie să se suprapună)

EXEMPLU

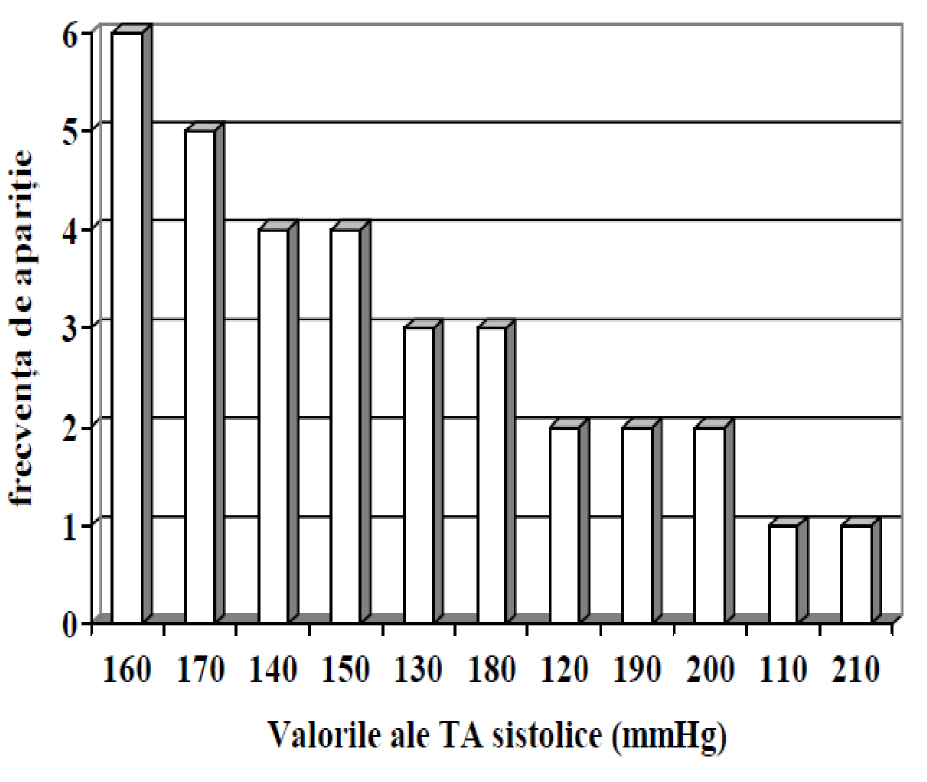
Pentru a fi mai expliciţi, să luăm un exemplu: Fie o serie de 33 de date numerice obţinute experimental (de exemplu valori ale tensiunii arteriale sistolice): 180, 120, 110, 200, 140, 210, 200, 190, 150, 170, 140, 130, 150, 170, 160, 120, 160, 140, 160, 170,

180, 160, 150, 130, 160, 180, 190, 160, 170, 170, 150, 150, 130.

Cu această serie se poate alcătui o *diagramă*, aşezând datele, în ordinea frecvenţei, pe o singură scară a graficului cartezian.

În felul acesta se obţine structura acestei colectivităţi şi se poate constata frecvenţa de apariţie a unor rezultate (de exemplu câte valori ale

tensiunii arteriale 110 se găsesc în respectiva colectivitate statistică).



Se obţine astfel o *distribuţie de frecvenţă* a colectivităţii respective.

Se pot obţine distribuţii de frecvenţă *homograde* (cum este cazul diagramelor), cu o singură scară de comparaţie în sistemul cartezian, sau

distribuţii heterograde.

Căutând şi ale modalităţi de caracterizare a colectivităţii, se poate stabili o distribuţie de frecvenţă *heterogradă*, pe două scări ale sistemului cartezian, înşiruind pe abscisă numerele, în ordine crescătoare sau descrescătoare şi notând, în acelaşi timp, *intervalele de clasă*, iar pe ordonată punând frecvenţele de apariţie (relative sau absolute). Se obţine astfel o *histogramă*. Prin unirea ordonatelor care trec prin mijlocul intervalelor de clasă se obţine *poligonul de frecvenţă*.

Practic, pentru a reprezenta grafic corect datele colectate în cadrul unei histograme, trebuie parcurşi următorii paşi:

i. Pentru şirul valorilor măsurate se va determina *numărul de intervale de grupare (de clasă)* – M (eventual *lungimea intervalului de grupare* - d), conform formulei lui Sturges:

$$M = 1+3,22*\log n$$

unde n = numărul măsurătorilor efectuate

Valoarea numărului de intervalule se va rotunji pentru a obţine un număr întreg. Pentru cazul nostru, numărul vom obţine

$$M=1+3,22*\log 33=5,889$$

adică vom avea 6 intervale.

Eventual, lungimea intervalului de grupare va fi calculată cu formula:

$$d = \frac{X_{\max} - X_{\min}}{1+3,22*\log n}$$

unde Xmax, Xmin reprezintă valoarea maximă şi minimă măsurată. Pentru cazul nostru

$$d = \frac{210-110}{1+3,22*\log 33} = 16,97$$

Dată fiind precizia măsurătorilor efectuate cu tensiometrele clasice, putem lua în considerare un o lungime a intervalului de grupare de circa 20 mmHg.

Practic, pentru a reprezenta grafic corect datele colectate în cadrul unei histograme, trebuie parcurşi următorii paşi:

ii. În baza numărului calculat de intervale (M), respectiv a valorii lungimii intervalului de grupare (d), între valorile limită masurate (**Xmin – Xmax**), se vor stabili intervalele de grupare. Datele obţinute vor fi centralizate într-un tabel.

iii. Se determină *frecvenţa absolută* (ni), care reprezintă "numărul de apariţii" a datelor corespunzătoare fiecărui interval de grupare în parte. Şi aceste date se vor trece în tabelul sus-menţionat.

iv. Se poate determina şi *frecvenţa relativă* (fi), care reprezintă raportul frecventei absolute (ni) la numarul total de măsurători (n).

v. Se construieşte histograma, care reprezinta o diagrama în formă de "dreptunghiuri" având baza egala cu intervalul de grupare, iar înălţimea cu frecvenţa (absolută sau relativă)

vi. Se construieşte poligonul de frecvenţă, care se obţine prin unirea mijloacelor superioare ale histogramei prin segmente de dreaptă.

Tabelul cu intervale şi frecvenţa absolută, precum şi histograma rezultată în cazul celor 33 de valori ale tensiunii arteriale sistolice sunt prezentate mai jos:

Interval de grupare	*Frecvența absolută*
<110	1
111-130	5
131-150	8
151-170	11
171-190	5
>190	3

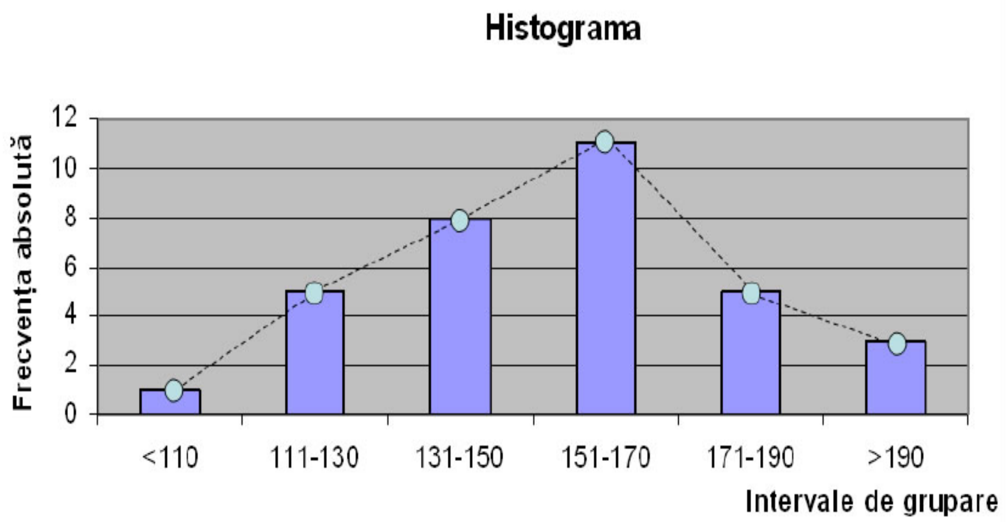
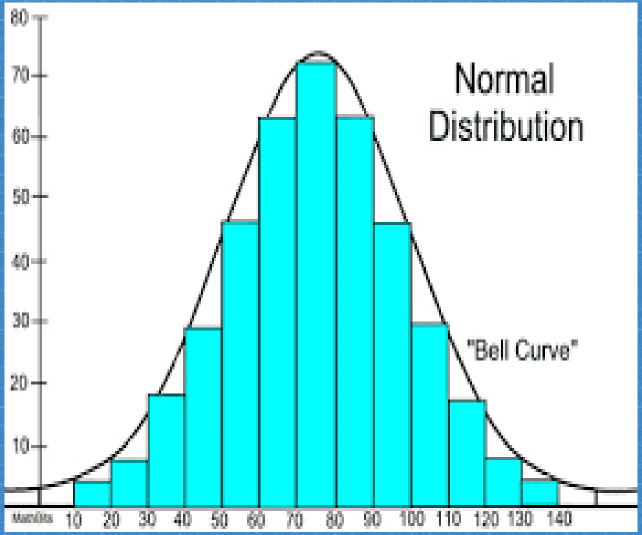


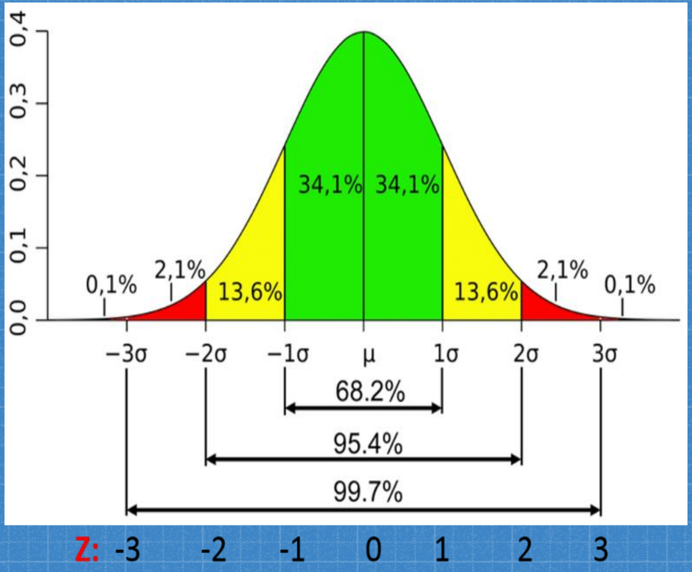
Fig. 2.2. Histograma

Distribuţia de frecvenţă permite determinarea şi a celorlalte caracteristici: ***tendinţa centrală*** (cu alte cuvinte, *media*, *mediana*, *forma distribuţiei*, *variabilitatea din interiorul ei*).

În figura de mai sus se poate constata că forma distribuţiei se apropie destul de mult de distribuţia normală Laplace-Gauss (distribuţia în formă de clopot). În acest caz, valoarea medie reprezintă în cele mai bune condiţii tendinţa centrală (pentru

cazul studiat, media = 158,7).

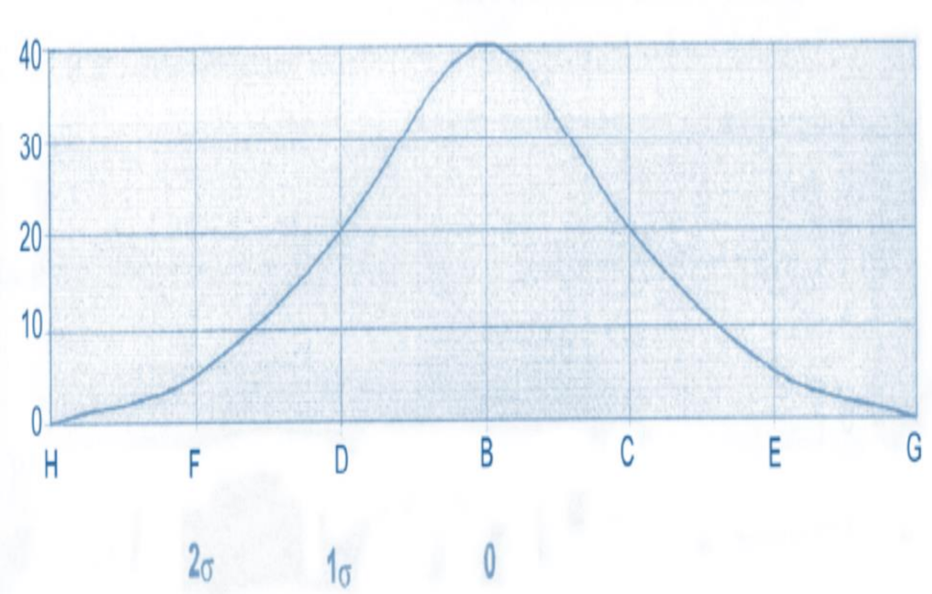




Curba de distribuție normală

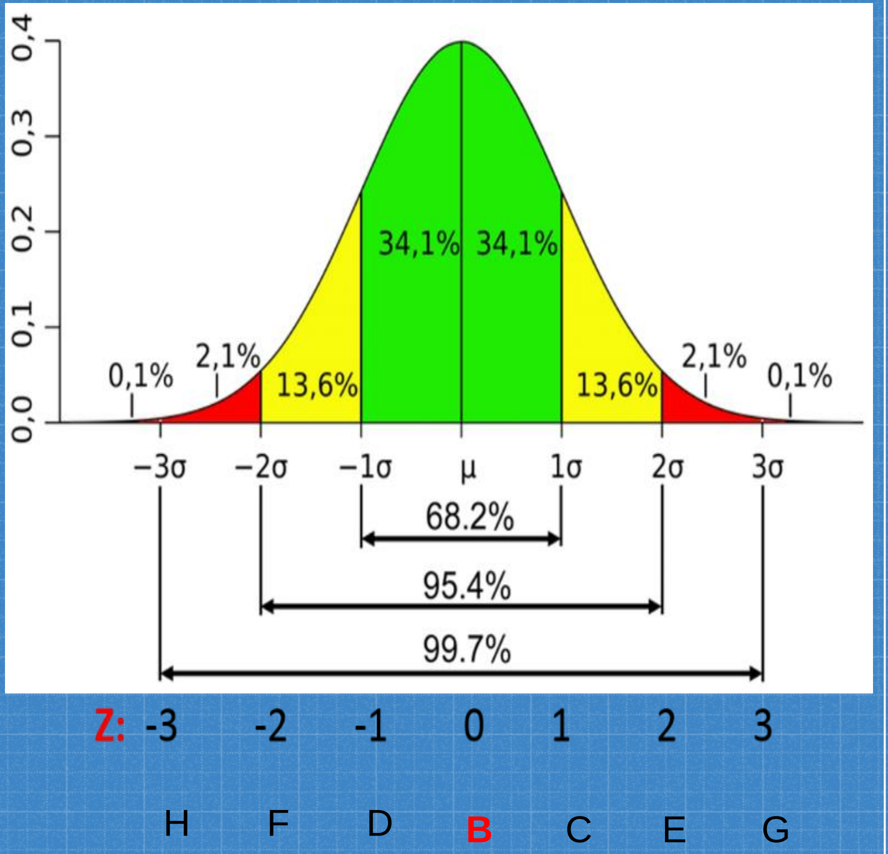
Van Vijngaarden (1926) a arătat pentru prima dată că variaţia rezultatelor biologice se datorează sensibilităţii individuale a animalelor (care generează, astfel, erorile întâmplătoare) şi că ele se supun legii de distribuţie normală a frecvenţei stabilită, în 1820, de Laplace şi Gauss

Curba de distribuţie normală a frecvenţei reprezintă frecvenţa cu care revine acelaşi rezultat în mai multe determinări succesive. Ea se poate obţine aşezând pe *abscisa* unui grafic diferenţele, obţinute în mai multe determinări, între media rezultatelor şi rezultatele individuale, iar pe ordonată frecvenţele de apariţie a răspunsului pentru fiecare diferenţă. Graficul are forma unui clopot.



Media rezultatelor individuale, care se repetă cel mai des (are cea mai mare frecvenţă de apariţie), este punctul cel mai înalt al curbei. Valoarea medie este notată pe abscisă cu 0, deoarece diferenţa sa faţă de medie este, evident, zero. De o parte şi de alta a punctului "culminant", se desfăşoară simetric frecvenţele corespunzătoare diferenţelor dintre media rezultatelor şi rezultatele individuale, care se găsesc pe abscisă; cele negative (mai mici decât media) în partea stângă, cele pozitive (deci mai mari decât media) în partea dreaptă a valorii medii.

Distanţa BD sau BC (0-1 sau 0+1) reprezintă convenţional o unitate denumită *abatere standard* şi notată cu σ (sigma). Perpendiculara pe valoarea medie este axul de simetrie al suprafeţei acoperită de curbă. Perpendicularele în punctele de pe abscisă care corespund valorii medii plus abaterea standard şi valorii medii minus abaterea standard, închid două treimi din suprafaţa acoperită de curbă (66%). Perpendicularele care corespund valorii medii plus sau minus 2σ închid circa 95% din suprafaţa acoperită de curbă. Suprafeţele terminale ocupă numai 5 % din suprafaţa totală.



CARACTERISTICILE DISTRIBUŢIILOR DE FRECVENŢĂ

Orice serie de date cantitative se poate descrie prin două elemente caracteristice:

1

INDICATORII TENDINŢEI CENTRALE

2

INDICATORI AI ÎMPRĂŞTIERII SAU DISPERSIEI DATELOR

Tendința centrală (distribuțiile de frecvență)

Media reprezintă tendinţa centrală a unei distribuţii (vom studia

mai târziu modul ei de calcul).

Modul reprezintă valoarea cea mai frecventă a unei distribuţii,

care se confundă, de fapt, cu vârful poligonului de frecvenţă

Mediana corespunde valorii care se găseşte la punctul care împarte

seria statisticii în două grupuri egale

O altă caracteristică a distribuţiilor, care uneori este foarte folositoare, este *forma curbei de frecvenţă*. Ea poate fi:

simetrică

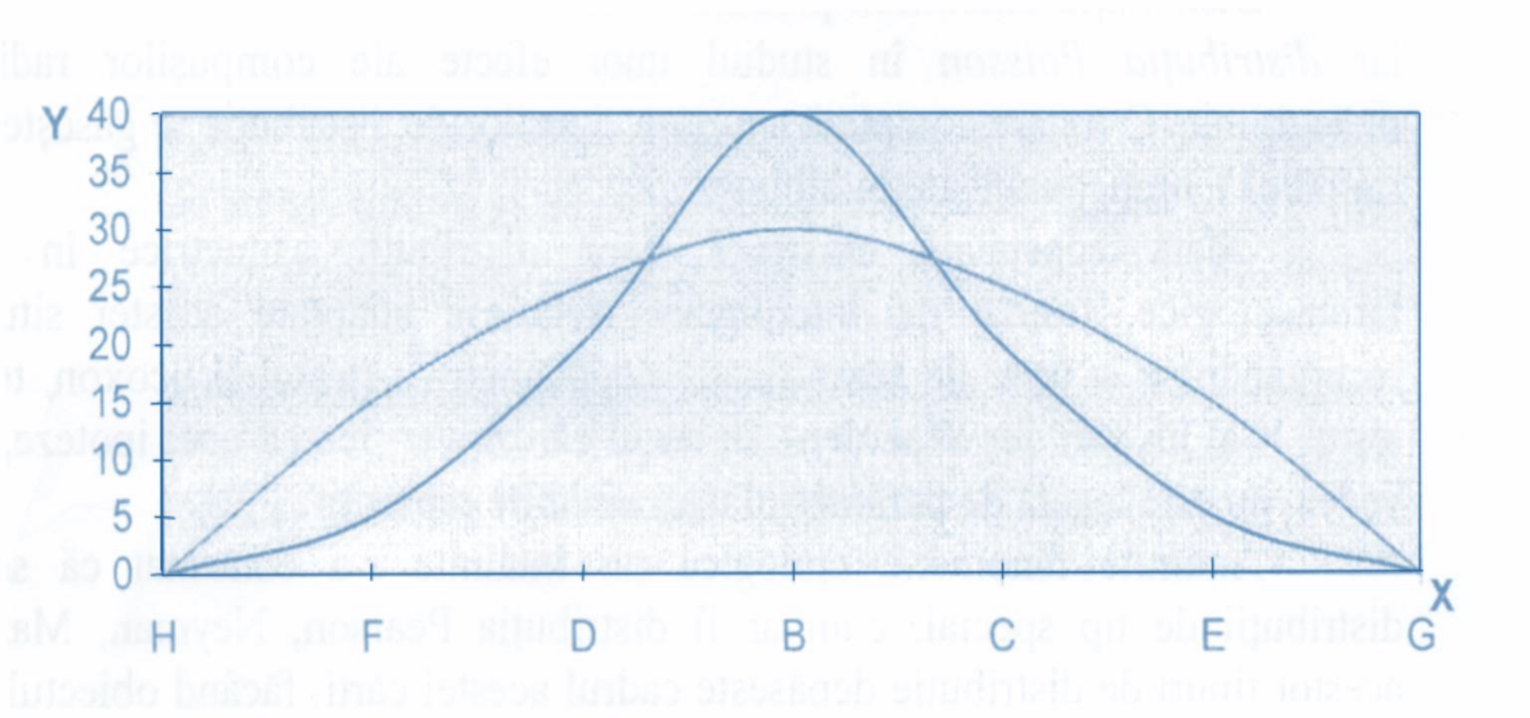
asmietrică

- **pozitivă**, caz în care „coada lungă” a distribuţiei este de partea valorilor pozitive.
- **negativă**, caz în care „coada lungă” a distribuţiei este de partea valorilor negative.

Două curbe cu aceeaşi medie, modul şi aceeaşi mediană se pot deosebi după bază şi înălţime: mai îngustă şi mai înaltă sau mai largă şi mai joasă. Întinderea bazei poate da o măsură a variabilităţii. Deschiderea este cu atât mai mare cu cât participarea factorilor întâmplători este mai mare (în figura următoare) se pot observa diferenţele între două curbe cu aceeaşi medie).

Calculul precis la împrăştierea rezultatelor se face cu ajutorul *abaterii standard.*

Două curbe cu aceeaşi medie şi modul, dar cu împrăştieri diferite ale rezultatelor



În cazul distribuţiilor simetrice şi unimodale există egalitate între media aritmetică, mod şi mediană (este vorba despre curba normală de distribuţie a frecvenţei Laplace-Gauss).

DISTRIBUȚII ANORMALE (NON –GAUSSIENE)

Se cunosc, în afara distribuţiei normale unimodale, şi distribuţii pLurimodale sau asimetrice.

Pentru a întelege mai bine importanţa tipului de distribuţie în cazul datelor provenite din cercetarea biomedicale trebuie spus că, de exemplu, existenţa unor distribuţii anormale pot arăta o lipsă de omogenitate a afectelor farmacodinamice ale unui medicament. Unele distribuţii pot lua forma literei U, unde importantă este valoarea minimă (de exemplu, în cazul acţiunii hipoglicemiante a unor substanţe active).

Există cazuri, însă, când fenomenele studiate se supun unei distribuţii de tip special. Cele mai des întâlnite distribuţii de acest tip au fost descrise de:

Bernouli (***distribuţia binomială***) prezintă interes mai ales în studiul fenomenelor eredităţii

Poisson (***distribuţia evenimentelor rare***) în studiul unor efecte ale compuşilor radioactivi şi în radiochimie

În cazul experimentării în domeniul farmacodinamic, rezultatele unei cercetări, odată reprezentate grafic dau distribuţii empirice sau experimentale. Compararea acestor distribuţii cu distribuţiile teoretice poate fi de un real folos pentru o interpretare justă a fenomenelor observate. Trebuie menţionat că majoritatea distribuţiilor obţinute în urma analizei datelor rezultate din cercetarea biomedicală se supun legilor normale de repartiţie a frecvenţei, de aceea calculele şi tehnicile de lucru prezentate în în continuare se referă, în mod special, la această ipoteză .

2.2. INDICATORII TENDINȚEI CENTRALE

Media aritmetică - notată de regulă cu $\bar{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$ *(III.1)*,

Mediana - este acea valoare din şirul de date care împarte în două părţi egale şirul ordonat de valori (atenţie, şirul este ordonat crescător), situându-se la mijlocul seriei statistice. Dacă numărul de valori n este un număr impar, atunci mediana este valoarea

$M_e = x_k$ *(III.2)*, , unde $k = \dfrac{n}{2} + 1$.

Dacă n este par, deci avem un număr par de valori, mediana este definită ca fiind $M_e = \dfrac{x_k + x_{k+1}}{2}$ *(III.3)*, unde $k = n/2$.

Modul - constituie valoarea care apare cel mai des, deci valoarea cu numărul cel mai mare de apariţii.

Calcul online medie aritmetică

https://www.calculat.org/ro/medie/medie-aritmetica.html

Calcul online medie aritmetică ponderată

https://www.calculat.org/ro/medie/medie-ponderata.html

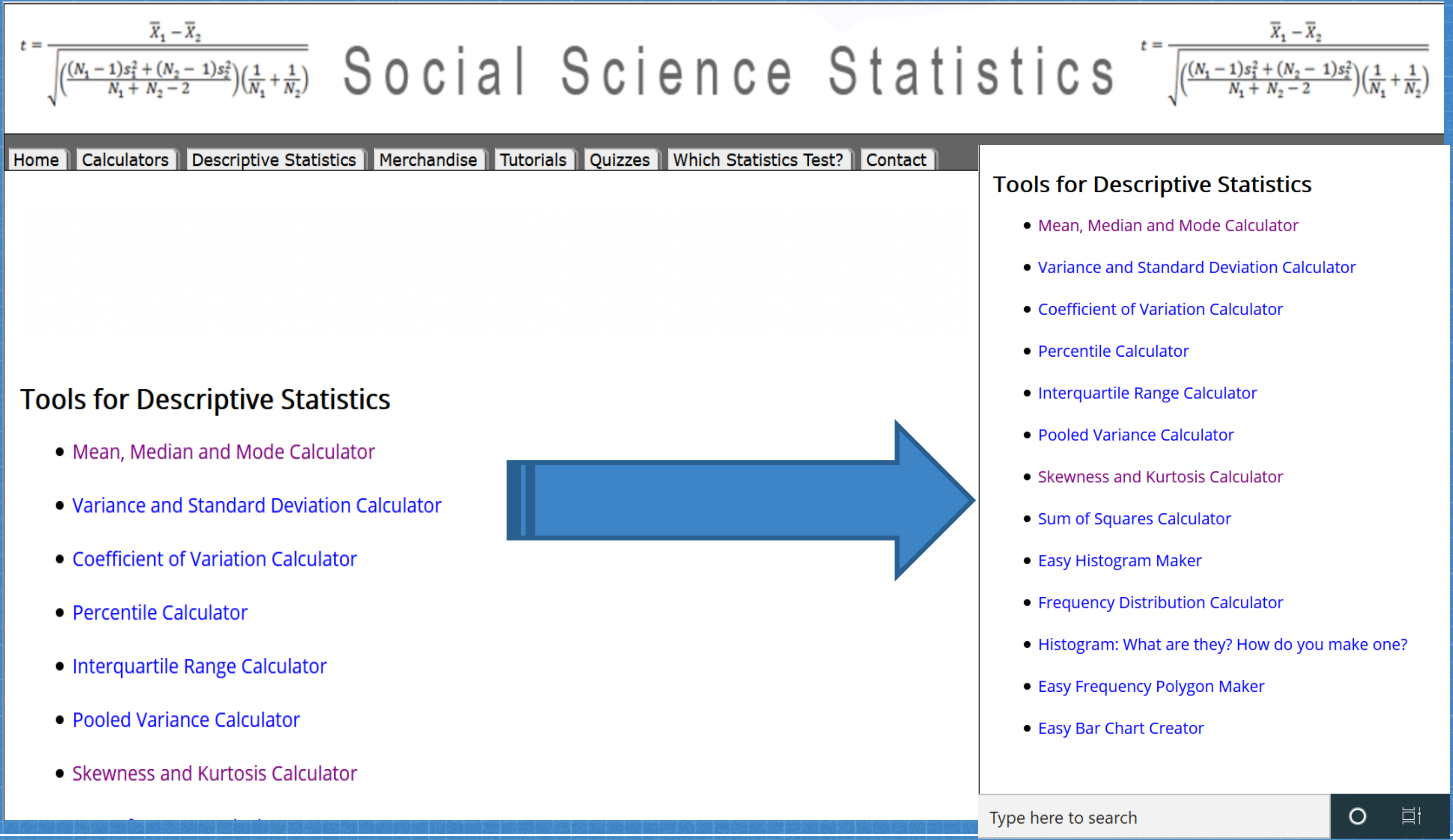
Calculator statistici

https://www.calculator.net/statistics-calculator.html

https://www.calculator.net/standard-deviation-calculator.html

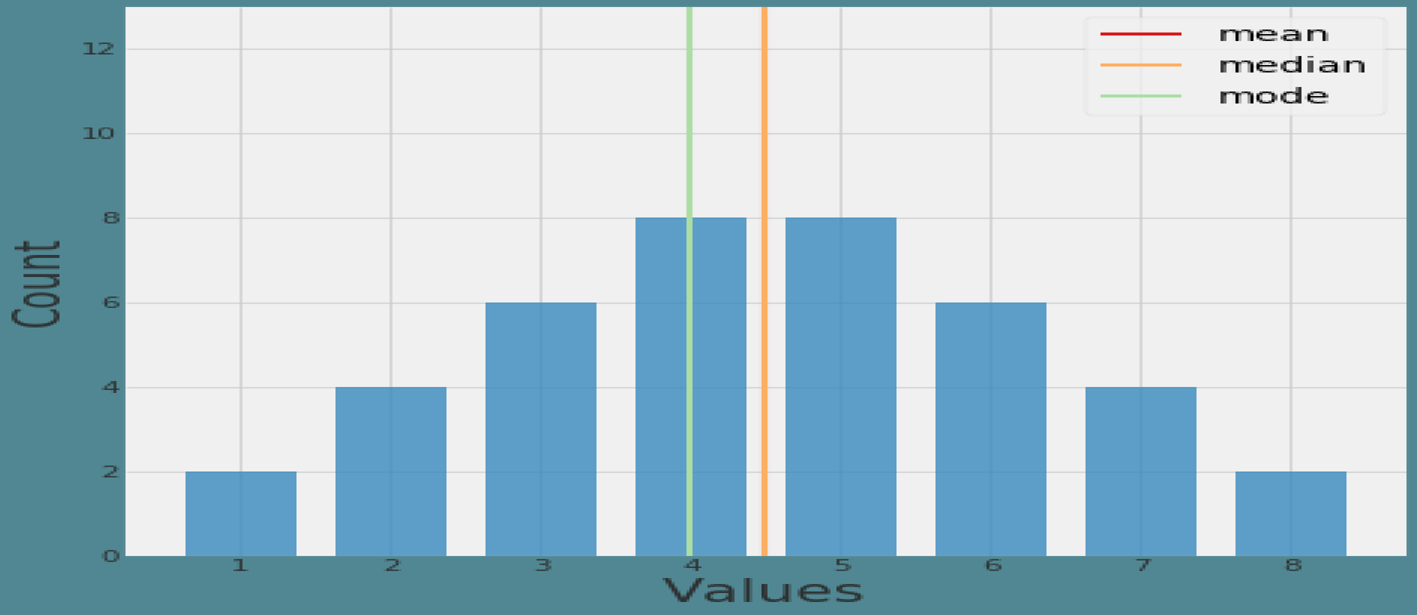
https://www.socscistatistics.com/descriptive/

https://www.socscistatistics.com/descriptive/



EXEMPLE (https://www.probabilisticworld.com/mean-mode-median/)

(SIMULATION)

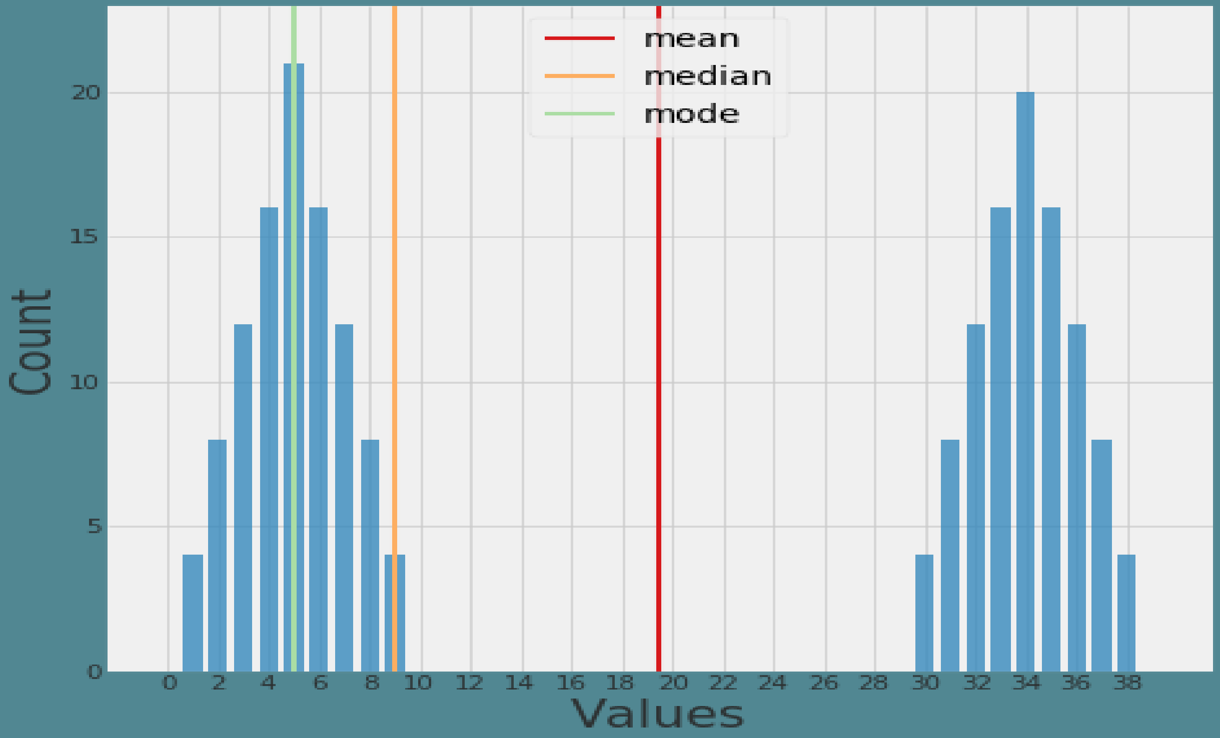


EXEMPLE (https://www.probabilisticworld.com/mean-mode-median/)

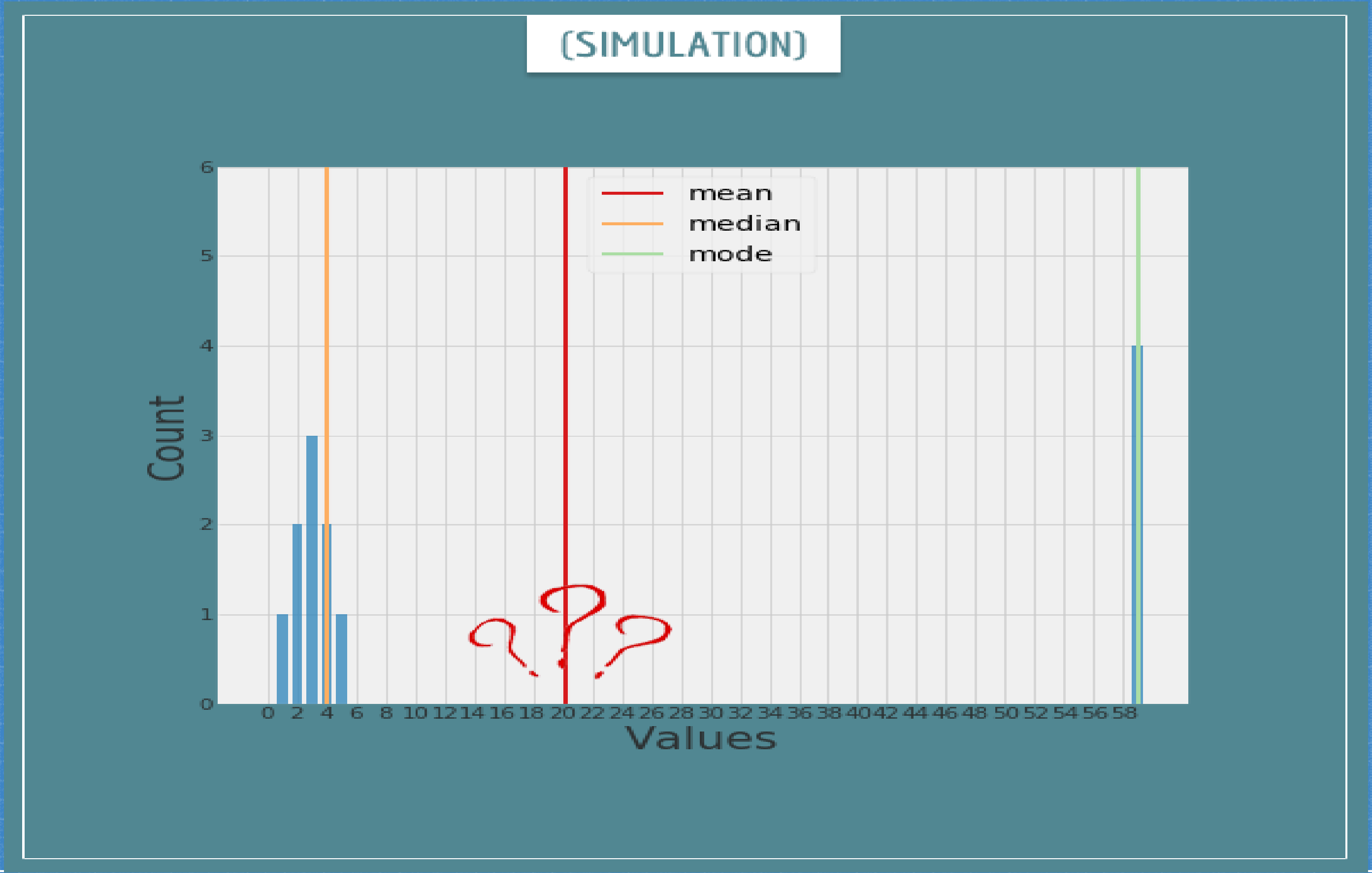
(SIMULATION)

Exemplu

(SIMULATION)



Reprezentativitatea mediei (https://www.probabilisticworld.com/mean-mode-median/)



Simple Learning Pro (https://www.youtube.com/watch?v=mk8tOD0t8M0)

- **Mode, Median, Mean, Range, and Standard Deviation (1.3)**

2.3. INDICATORI AI DISPERSIEI (ÎMPRĂȘTIERII) DATELOR ÎN JURUL VALORII MEDII

Varianța notată $s^2{}_x$ este un indicator de împrăştiere a datelor. Formula de calcul este: $$s_x^2 = \frac{\sum x_i^2 - (\bar{x})^2}{n-1} \quad (III.4),.$$

Abaterea standard sau deviația standard reprezintă rădăcina pătrată din varianță (dispersie) : $$s_x = \pm\sqrt{s_x^2} \quad (III.5),$$

Coeficientul de variație se calculează ca un raport procentual între abaterea standard şi valoarea medie a şirului de valori.

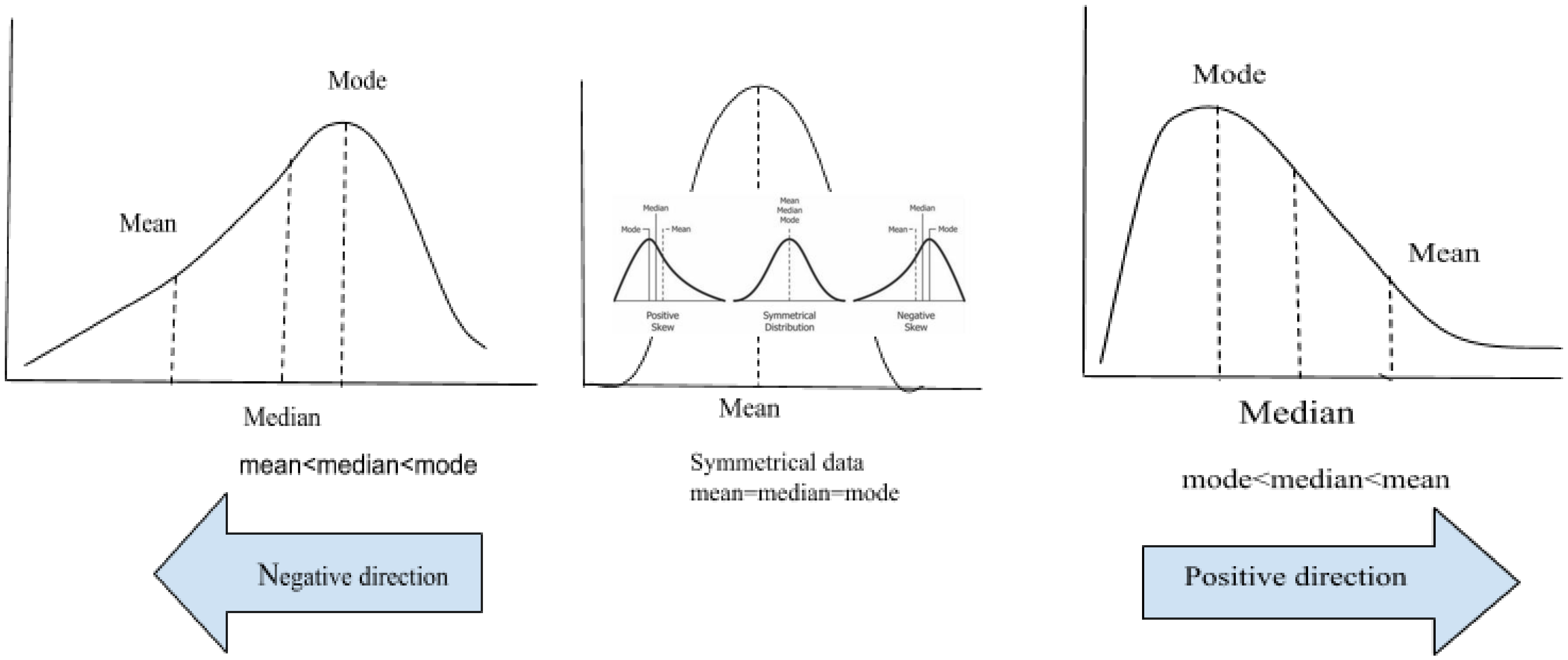
$$CV\% = \frac{s_x}{\bar{x}} \cdot 100 \quad (III.6),$$

Asimetria (skewness) caracterizează gradul de “asimetrie” a unei distribuții în jurul valorii medii, comparativ cu distribuția normală. Valori pozitive ale asimetriei indică o distribuție de frecvență ce prezintă o “coadă” (în engleză tail) asimetrică în zona valorilor “pozitive” ale distribuției (valori mai mari decît media). Similar, valori negative ale asimetriei indică o distribuție de frecvență ce prezintă o “coadă” (în engleză tail) asimetrică în zona valorilor “negative” ale distribuției (valori mai mici decît media).

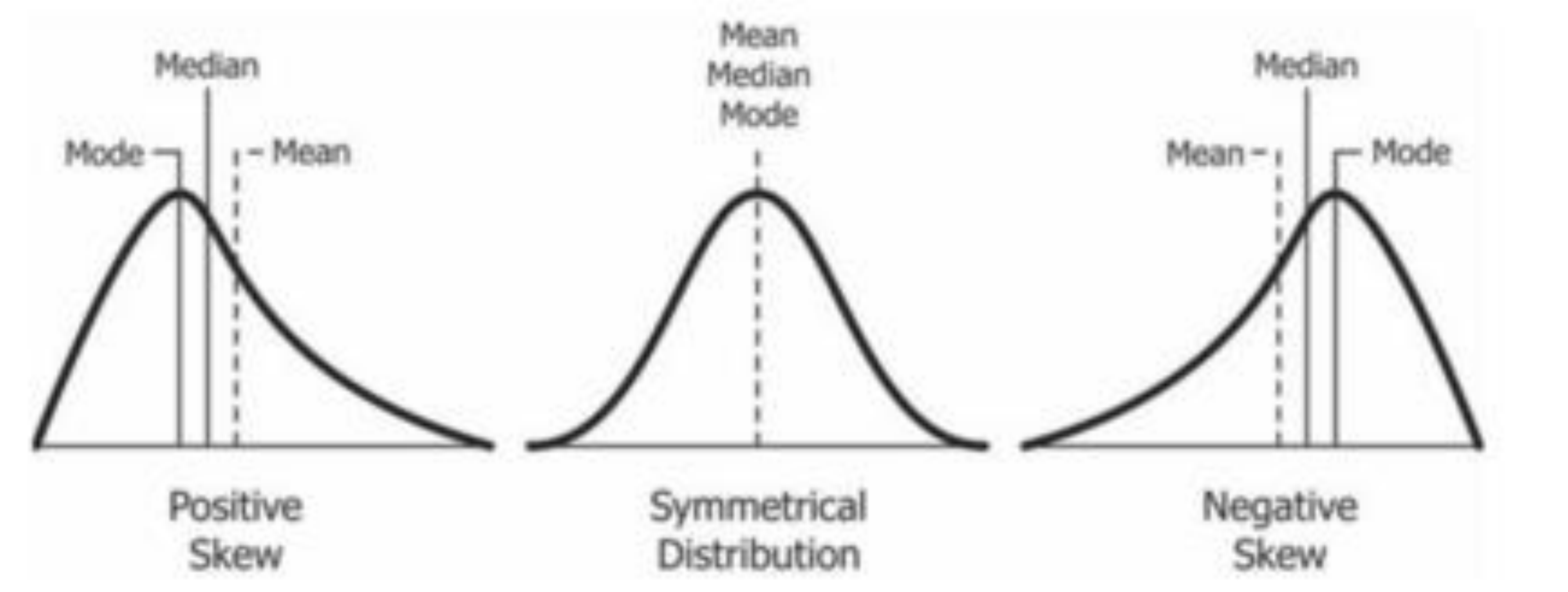
$$Skewness = \frac{n}{(n-1)\cdot(n-2)} \cdot \sum \left(\frac{x_i - \bar{x}}{S} \right)^3 \quad (III.7),$$

, unde S = abaterea standard.

EXEMPLE

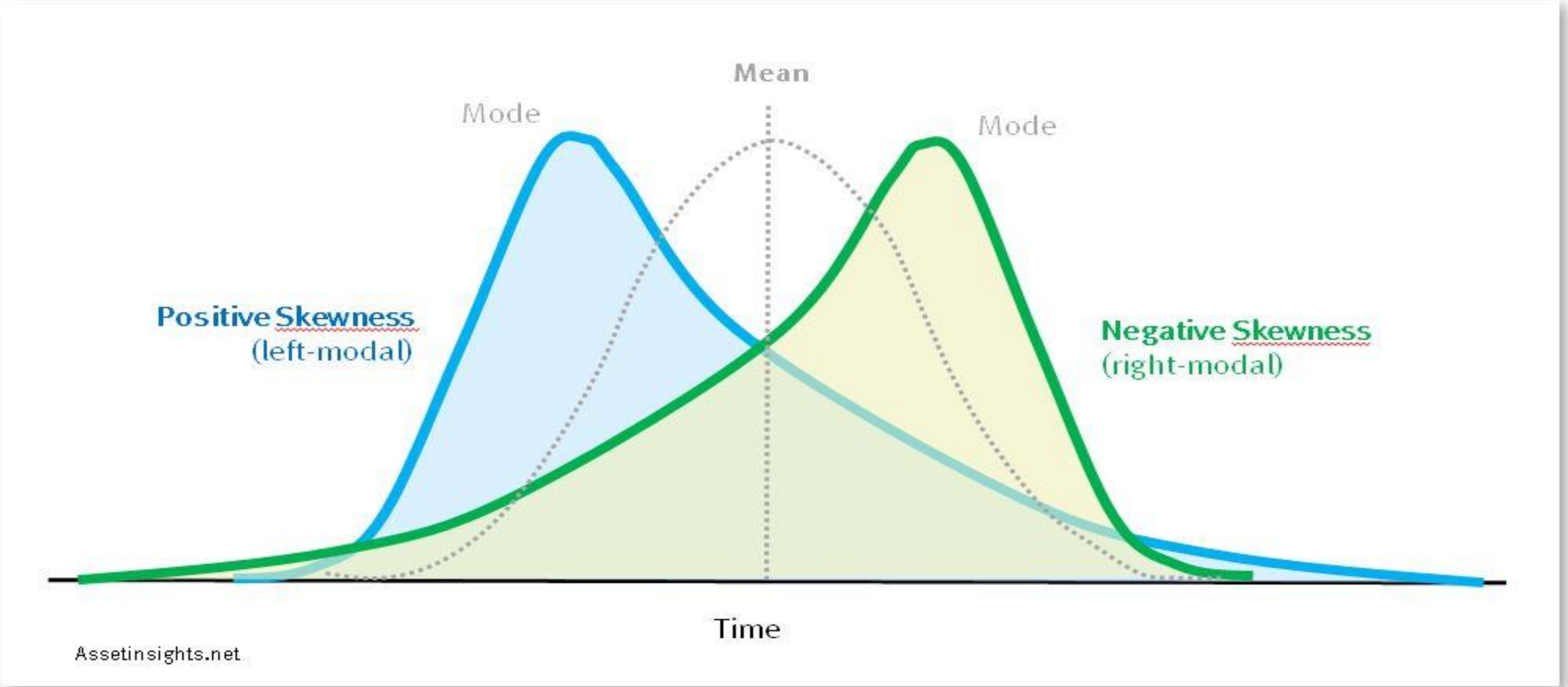

Mode
Mean
Median
mean<median<mode
Negative direction
Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew
Mean
Symmetrical data
mean=median=mode
Mode
Mean
Median
mode<median<mean
Positive direction

EXEMPLE

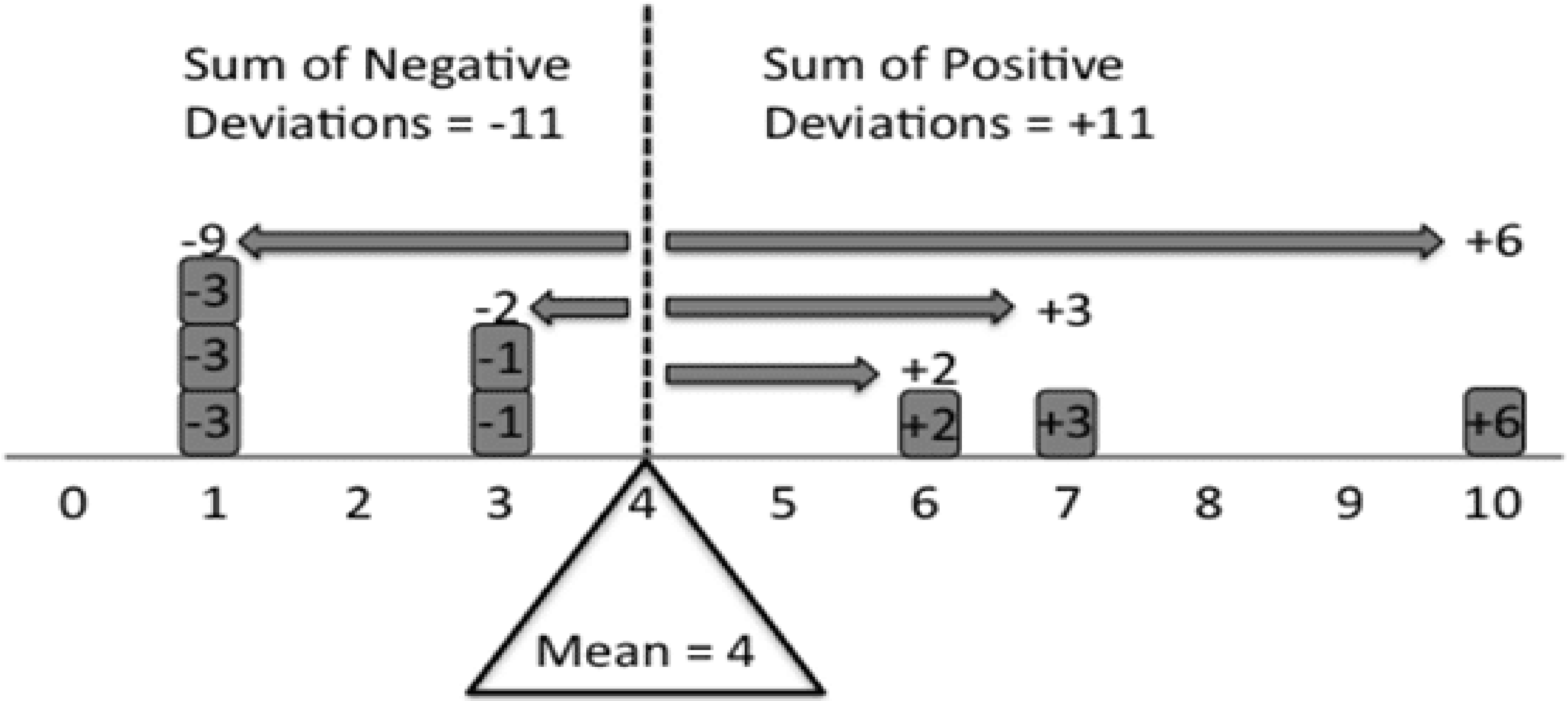


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

EXEMPLE



Mean
Mode
Mode
Positive Skewness
(left-modal)
Negative Skewness
(right-modal)
Time
Assetinsights.net


Sum of Negative
Deviations = -11
Sum of Positive
Deviations = +11
-9
+6
-2
+3
+2
-3
-3
-3
-1
-1
+2
+3
+6
0
1
2
3
4
5
6
7
8
9
10
Mean = 4

Aplatizarea (kurtosis) caracterizează gradul de “aplatizare” a unei distribuții, comparativ cu distribuția normală. Valorile pozitive ale acestui indicator indică o distribuție cu un “vârf” mai înalt decât distribuția normală. Similar, valori negative ale kurtosisului indică o curbă relativ aplatizată, comparativ cu distribuția normală.

$$Kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n+3)} \cdot \sum \left(\frac{x_i - \bar{x}}{S} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

(III.8),, unde S = abaterea standard.

Amplitudinea - este diferenţa dintre valoarea maximă şi cea minimă

$$A = A_{max} - A_{min}$$

Amplitudinea relativă - notată *A%* este raportul dintre amplitudinea absolută şi media aritmetică a seriei de date. Atunci când avem foarte multe date se recomandă includerea lor în clase egale ca mărime, ceea ce uşurează mult prelucrările statistice ulterioare. Spre exemplu sortăm pacienţii pe grupe de vârstă: 21-24 de ani, 25-30 ani, etc… În acest caz apare noţiunea de **frecvenţă a clasei**.

Indicatori statistici pentru serii de date cu apariţii frecvente ale aceleiaşi valori

Dacă datele pe care le studiem conţin valori care se repetă des, se obişnuieşte să se grupeze datele care au aceeaşi valoare . Numărul de apariţii ale unei valori anume se numeşte frecvenţa de apariţie şi se notează cu f_i. Presupunem că în urma măsurătorilor am obţinut şirul de valori: x_1 cu frecvenţa f_1, x_2 cu frecvenţa f_2, … x_n cu frecvenţa f_n

Indicatorii statistici se calculează conform noilor formule:

Media aritmetică

$$\overline{x} = \frac{\sum\limits_{i=1,n} x_i \cdot f_i}{\sum\limits_{i=1,n} f_i} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_n \cdot f_n}{f_1 + f_2 + \ldots + f_n} \quad (III.10)$$

Mediana – este x_k *(III.11,)*unde $k = \dfrac{\sum\limits_{i=1,n} f_i + 1}{2}$

Dispersia (varianța) : $s_x^2 = \dfrac{\sum\limits_{i=1,n} (x_i - \overline{x})^2 \cdot f_i}{\sum\limits_{i=1,n} f_i - 1}$ *(III.12)*

Calcul online medie aritmetică

https://www.calculat.org/ro/medie/medie-aritmetica.html

Calcul online medie aritmetică ponderată

https://www.calculat.org/ro/medie/medie-ponderata.html

Calculator statistici

https://www.calculator.net/statistics-calculator.html

https://www.calculator.net/standard-deviation-calculator.html

https://www.socscistatistics.com/descriptive/

EROARE STANDARD

Este cunoscut faptul că determinările biologice sunt supuse influenţei a două tipuri de erori: cele care influenţează precizia determinării şi cele care influenţează exactitatea determinării. Pentru a afla exactitatea cu care s-au făcut o serie de determinări, trebuie să se calculeze abaterea medie a valorilor medii obţinute sau, altfel spus, media erorilor ce se pot comite într-o determinare. Această abatere a fost denumită eroare standard, notată cu E. Calcularea ei se face cu ajutorul formulei:

$$E^2 = \frac{\Sigma(x_1 - \bar{x})^2}{n(n-1)}$$

Ştiind că, în cazul distribuţiei normale gaussiene, împrăştierea în jurul mediei colectivităţii a unei medii de şantion este n ori mai mică decât împrăştierea rezultatelor individuale, eroarea standard este dată şi de formula:

$$E = \frac{S}{\sqrt{n}}$$

Ea reprezintă formula clasică a erorii standard. Rezultatele experimentărilor biologice trebuie să fie însoţite întotdeauna de eroarea standard sau de abaterea standard, utilizându-se formulări de tipul M ± S sau M ± E, pentru a permite o justă interpretare a lor.

COEFICIENT DE VARIAȚIE

Coeficientul de variaţie se calculează ca un raport procentual între abaterea standard şi valoarea medie a şirului de valori.

$$CV\% = \frac{s_x}{\bar{x}} \cdot 100$$

De remarcat că valoarea coeficientului de variaţie nu are unitate de măsură, se exprimă procentual. Acest fapt permite folosirea indicatorului la compararea a două sau mai multe serii de date, indiferent de ordinul de mărime al variabilelor (variantelor) şi de unităţile de măsură folosite. Se poate considera că un coeficient de variaţie sub 10% indică o dispersie mică (o împrăştiere) a datelor, adică seria este omogenă. Un coeficient între 10% şi 30% indică dispersie mijlocie, iar peste 30% indică dispersie mare. *Dacă dispersia este mare, media nu este un indicator reprezentativ.*