

Table of contents

- 1 Statistică
 - Introducere
 - Vocabular
- 2 Statistică descriptivă
 - Variabile
 - Reprezentări grafice
 - Măsuri ale tendinței centrale
 - Media
 - Mediana
 - Mòdul
 - Cvartile
 - Măsuri ale variabilității
 - Domeniul
 - Dispersia eşantionului și deviația standard a eşantionului
 - Mediana, cvartilele și domeniul intercvartilic
 - Valori aberante
- 3 Bibliografie

Istorie

- Rădăcinile cuvântului Statistică sunt latine: Status (latina veche) care înseamnă stat (politic), Statista (italiană) înseamnă politician.
- La mijlocul secolului al XVII-lea într-o Universitate Germană a fost folosit pentru prima oară cuvântul statistik cu sensul de știință politică a statelor: analiza datelor privind statele.
- În Marea Britanie la sfârșitul secolului XVIII termenul de statistică a fost introdus cu un înțeles similar: știința statelor (sau aritmetica politică).
- Utilizarea statisticii fără a o numi în mod expres datează de la începutul civilizației umane: forme incipiente de recensământ al populației, sistematizarea datelor geografice și economice etc.

Istorie

- Primul studiu statistic este considerat în general a fi cel care a pus bazele demografiei: în 1662 doi englezi au introdus metode statistice cum ar fi tabelele speranței de viață și probabilitățile de supraviețuire la diferite vârste.
- Abia în secolul XIX rezultatele din teoria probabilităților au început a fi folosite în raționamentul statistic.
- Bazele matematice ale statisticii s-au consolidat datorită rezultatelor profunde obținute în teoria probabilităților din secolul anterior.
- Începând cu secolul XX au fost dezvoltate noi metode și teorii, iar o influență asupra statisticii a avut-o și dezvoltarea informaticii.

Introducere

"Statistics has become the universal language of the sciences."

Elementary Statistics, R. Johnson, P. Kuby

○ *studiu statistic* tipic cuprinde

- unul sau mai multe experimente aleatoare din efectuarea cărora rezultă o serie de date.
- o metodă de extragere a informației din date și de interpretare a rezultatelor.

Modul în care informația este procesată și interpretată dă naștere la două ramuri ale statisticii ca știință:

- *Statistica descriptivă* - colectează, prezintă și descrie datele (de multe ori în formă grafică).
- *Statistica inferențială* - folosind datele deja colectate ia decizii relative la populația în cauză.

Introducere

Definiția 1

Statistica este știința colectării, descrierii, interpretării datelor și luării de decizii pe baza acestor date.

Cele două ramuri ale statisticii sunt și cei doi pași dintr-un studiu statistic:

- *statistica descriptivă* are rolul de a
 - sintetiza, aduna și reprezenta datele;
 - aranja informația, pregătind-o pentru luarea deciziilor;
- *statistica inferențială* are drept scop
 - luarea deciziilor pe baza datelor strânse;
 - estimarea parametrilor (cum sunt media, dispersia etc);
 - verificarea ipotezelor statistice.

Terminologia statistică

- Statistica își are propriul limbaj, dincolo de împărțirea în descriptivă și inferențială.
- Cel mai important concept în statistică este acela de *populație*: colecția completă (exhaustivă) a obiectelor care prezintă interes pentru cel care face studiul.
- Exemple de populații: mulțimea studenților din Iași, mulțimea românilor analfabeți, mulțimea dozelor de cola produse într-o luna într-o fabrică, mulțimea furtunilor tropicale din 2018.

Definiția 2

O *populație* este o mulțime de obiecte (numite și indivizi) ale căror proprietăți vor fi analizate.

- O populație poate fi finită (dacă poate fi teoretic listată) sau infinită (populația cutremurelor de pământ din zona Vrancea).

Terminologia statistică

- Din cauza dimensiunilor mari ale unei populații studiul statistic se concentrează asupra unei porțiuni mai mici a populației. Acesta este un *eșantion* care constă din indivizi selectați din populație.

Definiția 3

Un eșantion este o submulțime a populației. Dintr-un punct de vedere teoretic fiecare individ are aceleași șanse de a aparține eșantionului, și orice grup particular de indivizi este ales în mod independent pentru a face parte din eșantion. Dacă aceste condiții sunt îndeplinite atunci avem un eșantion aleator simplu.

- Când se alege o populație sau un eșantion pentru studiu, interesează o anumită trăsătură a indivizilor.
- Astfel de trăsături (attribute) pot fi: înălțimea, volumul, magnitudinea pe scara Richter, vârsta, presiunea sângelui, culoarea ochilor, suprafața etc.

Terminologia statistică

Definiția 4

O **variabilă** sau un **atribut** este o caracteristică a indivizilor din populație sau eșantion.

- După ce alegem un eșantion trebuie să măsurăm valorile unuia sau mai multor variabile asociate. Acestea sunt **datele**, ele pot fi numere reale, întregi, cuvinte, litere etc.

Definiția 5

Datele sunt valorile variabilei colectate de la fiecare individ din eșantion.

- O populație este descrisă numeric de **parametri** (medie, disperse, deviație standard); parametrii sunt în centrul unui studiu statistic.

Terminologia statistică

Definiția 6

Un **parametru** este o valoare numerică care privește întreaga populație.

- Dacă populația este foarte mare (ceea ce se întâmplă adesea) un parametru anume nu poate fi calculat.
- O soluție este de a calcula parametrul doar pentru un eșantion al populației. Aceasta este o **statistică**.
- Pentru orice parametru și fiecare eșantion există o statistică corespunzătoare.

Definiția 7

O **statistică** este un parametru calculat pentru un eșantion în locul întregii populații.

Folosirea terminologiei

- *Populație*: mulțimea studenților din primul an din Iași.
- *Eșantion*: studenții din primul an de la FII. (Acest eșantion nu este un eșantion simplu aleator.)
- *Variabilă/atribut*: dimensiunea vocabularului lor curent.
- *Date*: 4200, 3520, 1800, ... - dimensiunile vocabularului pentru fiecare student din primul an de la FII.
- *Parametru*: media dimensiunii vocabularului studenților din primul an din Iași
- *Statistică*: media dimensiunii vocabularului studenților din primul an de la FII.

Tipuri de variabile

- Clasificarea variabilelor împarte atributele în cantitative sau calitative. Astfel există
- Variabile care oferă o *informație calitativă*, cum ar fi culoarea ochilor studenților, genurile literare ale cărților dintr-o bibliotecă (ficțiune, știință, literatură motivațională etc), tipul de personalitate ale persoanelor dintr-o comunitate (sanguin, coleric, melancolic sau flegmatic), nivelul de satisfacție a clienților unui magazin etc.
- Variabile care dau o *informație cantitativă*; spre exemplu: înălțimea studenților, greutatea lor, suma de bani pe care un student o cheltuie pe cărți într-un an școlar ș. a.

Tipuri de variabile

Definiția 8

O **variabilă calitativă** (sau **categorică**) este o variabilă care descrie un individ dintr-o populație (conform unor categorii).

O **variabilă cantitativă** este o variabilă care măsoară sau numără ceva legat de un individ dintr-o populație.

- Variabilele calitative pot fi **nominale** sau **ordinale**.
- Variabilele nominale sunt: culoarea ochilor, tipul de personalitate, numele membrilor unei comunități etc.
- Exemple de variabile ordinale: nivelul de satisfacție a clienților, nivelul de educație (liceal, post liceal, universitar, post universitar, doctoral) etc.

Tipuri de variabile

Definiția 9

O **variabilă nominală** este o variabilă care numește sau descrie un individ dintr-o populație fără a putea asigura o ordine naturală acestor valori..

O **variabilă ordinală** este o variabilă ale cărei valori pot fi ordonate în mod natural.

- Variabilele cantitative pot fi **discrete** sau **continue**. Cele două tipuri pot fi distinse astfel: unele numără iar celelalte măsoară.

Tipuri de variabile

- O variabilă discretă de obicei numără: numărul de credite ale unui student, numărul de pagini ale unei cărți etc; câteodată o asemenea variabilă sumează puncte/note care nu pot fi continue.
- O variabilă continuă măsoară: volumul, înălțimea, viteza, presiunea etc.

Definiția 10

O **variabilă discretă** este o variabilă care are un număr finit sau infinit dar numărabil de valori; o astfel de variabilă poate avea valori corespunzând unor puncte izolate de pe un interval real.

O **variabilă continuă** este o variabilă care are un număr infinit și nenumărabil de valori; o astfel de variabilă poate avea, de obicei, orice valoare dintr-un interval real, incluzând orice valoare posibilă dintre orice două valori.

Reprezentări grafice

- O primă formă de explorare a datelor este utilizarea reprezentărilor grafice care pot revela un comportament sistematic (un șablon) al variabilei.
- Tipul de reprezentare grafică depinde în mod normal de tipul variabilei.
- Pentru date calitative reprezentările grafice folosite sunt pie charts și bar graphs.
- Pentru datele cantitative scopul reprezentărilor grafice este de a afla forma distribuției variabilei.

Reprezentări grafice - date calitative

- Datele calitative sunt mai întâi transformate în frecvențe.
- *Frecvența* unei observații (o valoare a unei variabile) este numărul de repetări ale acelei observații în eșantion.
- *Frecvența relativă* a unei observații este raportul dintre frecvența observației respective și numărul total de observații (dimensiunea eșantionului).
- *Distribuția frecvențelor* unei variabile calitative este familia tuturor perechilor formate din observație și frecvența sa corespunzătoare.

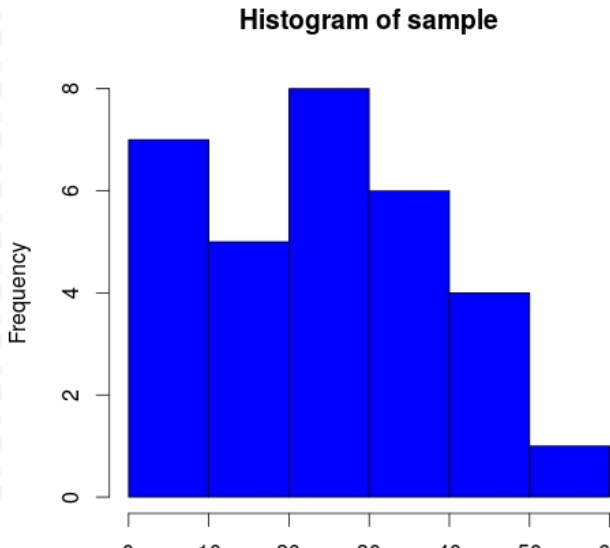
Reprezentări grafice - date cantitative

- Pentru date cantitative putem utiliza frecvențele și frecvențele relative sau *gruparea* datelor pentru a regăsi distribuția frecvențelor:
 - datele sunt *grupate* în clase (sau *bins*) care sunt uzual intervale cu aceeași lungime; clasele nu trebuie să se acopere.
 - o regulă pentru determinarea lungimii claselor: $1 + \log n / \log 2$ unde n este dimensiunea eșantionului.
 - apoi datele sunt *sortate* pe clase: se determină numărul observațiilor din fiecare clasă - acestea sunt frecvențele.
 - suma frecvențelor este dimensiunea eșantionului (n).
 - frecvențele relative se pot afla împărțind frecvențele la n .

Reprezentări grafice - date cantitative

- Cea mai utilizată metodă de reprezentare grafică a datelor cantitative este *histogram*.
- O altă metodă la îndemână pentru eşantioanele relativ mici este *stem-and-leaf*.

Reprezentări grafice - histograma



Datele

- Când privim reprezentarea grafică a datelor din eșantion ne putem pune următoarele întrebări.
- Care sunt valorile centrale/medii?
- Cât de mult sunt împrăștiate aceste date în jurul valorilor medii?
- Care este forma distribuției?
- Există valori care nu se potrivesc cu imaginea generală a distribuției?

Tendența centrală

- *Tendența centrală* sau *centrul distribuției* este centrul (abstract) al datelor. Toate măsurile tendinței centrale sunt legate într-un fel sau altul de noțiunea de medie.
- Diferite moduri de a defini tendința centrală:
 - Punctul care ține distribuția în echilibru.
 - Numărul care minimizează suma tuturor deviațiilor absolute.
 - Numărul care minimizează suma tuturor deviațiilor la pătrat.
 - Cea mai frecventă valoare.

Media (aritmetică)

- Să presupunem că valorile din eșantion sunt x_1, x_2, \dots, x_n .

Definiția 11

Media de selecție sau media eșantionului este media aritmetică a tuturor datelor din eșantion:

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

- Formula mediei pentru întreaga populație este în esență identică.
- *Media populației* se notează cu μ .

Media, eşantionul și populația

- În limbajul teoriei probabilităților media populației este media unei variabile aleatoare, X , ale cărei valori sunt cele ale indivizilor din populație; deci $\mathbb{E}[X] = \mu$.
- Media de selecție este o statistică care estimează media populației.
- Să presupunem că X_1, X_2, \dots, X_n sunt variabilele din spatele fiecărui individ al eşantionului, iar x_i este doar o valoare a variabilei X_i .
- Atunci X_i este o variabilă aleatoare cu aceeași distribuție ca a lui X . Mai mult, variabilele $(X_i)_{1 \leq i \leq n}$ sunt independente în ansamblu.
- Aceste observații conduc la faptul că media de selecție poate fi văzută ca o variabilă aleatoare, iar media aritmetică calculată pentru un eşantion este una dintre posibilele valori ale ei (fiecare eşantion dă o altă valoare a mediei de selecție).

Media, eşantionul și populația

- Dacă media de selecție este o variabilă aleatoare, îi putem calcula media:

$$\begin{aligned}\mathbb{E}[\bar{x}_n] &= \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \\ &= \frac{1}{n}\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \\ &= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]}{n} = \mu.\end{aligned}$$

- Media mediei de selecție este media populației.
- O astfel de statistică se numește *estimator nedeplasat* al parametru-lui corespunzător.
- Media de selecție este un estimator nedeplasat pentru media populației.

Media pentru date grupate

- Formula din definiția anterioară este valabilă pentru date negrupate. În acest caz toate datele din eșantion contribuie direct la calculul mediei de selecție.
- Pentru date grupate se folosește o formulă cu ponderi:

$$M = \frac{m_i * f_i}{\sum_i f_i}$$

unde m_i este mijlocul intervalului clasei i , iar f_i este numărul de observații care aparțin clasei i .

- În această formulă observațiile nu contribuie direct la calculul mediei; cu toate acestea este o formulă preferată în cazul datelor grupate pentru eșantioane mari fiind mai ușor de calculat.

Proprietăți ale mediei

- Ne întoarcem acum la definiția inițială (pentru date negrupate) a mediei de selecție.
- Variații mici în suma de la numărător nu modifică prea mult media. Spunem ca media este stabilă la variații mici ale datelor.
- *Valorile aberante* sau *extreme* pot avea o influență mare asupra mediei; introducând o valoare foarte mare sau foarte mică media se poate schimba foarte mult.
- Media este o funcție *liniară* (la fel ca media unei variabile aleatoare).
- *Deviatiile* de la medie sunt $(x_i - \bar{x}_n)$; suma lor este zero:

$$\sum_i (x_i - \bar{x}_n) = 0.$$

Proprietăți ale mediei

- (*Definiția variațională*) Se poate arăta că media este numărul M care minimizează suma deviațiilor la pătrat:

$$\sum_i (x_i - M)^2.$$

- Există și alte tipuri de medie în afară de cea aritmetică (A): media geometrică (G), media armonică (H).

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}, H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}.$$

- Să presupunem că o mașină parcurge distanța dintre două orașe de patru ori cu vitezele 80km/h, 90km/h, 60km/h, and 120km/h, respectiv. Care a fost viteza sa medie?
- Folosim media aritmetică obținem 87.5 km/h; dar media adecvată aici este cea armonică: 82.3km/h.

Mediana

- Mediana este o *statistică ordonată*; calculul unei astfel de statistici presupune ordonarea crescătoare a datelor din eșantion.

Definiția 12

Mediana (Me) este valoarea din mijloc când datele din eșantion sunt sortate.

- Mediana împarte datele din eșantion în două jumătăți: o jumătate conține datele mai mari sau egale decât mediana, iar cealaltă jumătate le conține pe cele mai mici sau egale.
- Valoarea medianei este o observație sau media a doua observații (pentru eșantioane de dimensiune pară).
- Ca statistică mediana este mult mai puțin influențată de existența valorilor aberante.

Mòdul

Definiția 13

Mòdul este observația cea mai frecventă din eșantion.

- Pentru date grupate se alege mai întâi clasa cu cea mai mare frecvență *clasa modală*. Fie i indexul acestei clase, a_i marginea stângă a intervalului corespunzător și L lungimea comună a intervalelor.
- Atunci mòdul poate fi calculat folosind formula

$$mod = a_i + \frac{L * (f_i - f_{i-1})}{(f_i - f_{i-1}) + (f_i - f_{i+1})}$$

- *Antimòdul* este cea mai puțin frecventă observație.

Compararea diferitelor măsuri

- Mai stabile la valorile aberante sunt mediana și mōdul.
- Media încorporează toate valorile și nu poate fi calculată în cazul datelor grupate pentru distribuții deschise (primul, sau ultimul interval deschis).
- Mediana și mōdul nu sunt funcții liniare.
- Mōdul este calculat mai ales pentru date grupate.
- Pentru distribuții asimetrice mōdul oferă cea mai reală imagine asupra tendinței centrale.

Compararea diferitelor măsuri

- Dacă eșantionul conține date foarte mari sau foarte mici mediana este măsura preferată mediei - stabilitatea o face mai reprezentativă.
- Pentru distribuții simetrice cele trei măsuri sunt aproape egale.
- Forma distribuției poate fi legată de relația dintre medie și mediană; forma poate fi
- asimetrică spre stânga dacă $\bar{x}_n < Me$;
- simetrică dacă $\bar{x}_n = Me$;
- asimetrică spre dreapta dacă $\bar{x}_n > Me$;

Cvartile

- Relativ la măsurile tendinței centrale există *măsurile de poziție* care sunt statistici ordonate ca și mediana.

Definiția 14

Cvartilele sunt valori care împart domeniul (ordonat al) observațiilor în patru segmente egale.

- Prima cvartilă, Q_1 , este o valoare astfel în cât 25% dintre observații sunt cel mult egale cu Q_1 și cel mult 75% sunt mai mari sau egale.
- A treia cvartilă, Q_3 , este o valoare astfel în cât 75% dintre observații sunt cel mult egale cu Q_3 și cel mult 25% sunt mai mari sau egale.

Cvartile

- A doua cvartilă, Q_2 , este o valoare astfel în cât 50% dintre observații sunt cel mult egale cu Q_2 și cel mult 50% sunt mai mari sau egale. Din acest motiv a doua cvartilă este egală cu mediana: $Me = Q_2$.
- Cvartilele au proprietăți similare cu cele ale medianei. Cea mai importantă fiind aceea că sunt stabile în prezența valorilor aberante.
- Statistici ordonate similare sunt: *decilele*, *percentilele* etc. Toate aceste statistici împart datele ordonate în subeșantioane egale.
- De exemplu există nouă decile care împart datele sortate în zece părți egale, fiecare parte reprezentând 10% din eșantion.

Măsurile ale variabilității

- După determinarea "centrului" datelor studiul statistic continuă cu analiza *împrăstierii* sau a *variabilității* datelor
- Valorile din eșantion pot să difere mult între ele și față de valoarea "centrală".
- Măsura în care valoare "centrală"/medie este reprezentativă pentru întreg eșantionul depinde de variabilitatea (sau dispersia) datelor.
- Eșantionul are variabilitate mare dacă există valori foarte mari sau foarte mici față de valoarea medie.
- Deoarece avem două metode mai importante de a măsura tendința centrală (media și mediana) vom avea două metode de a măsura împrăstierea.

Domeniul

Definiția 15

Domeniul este diferența dintre cea mai mică și cea mai mare valoare din eșantion.

$$\text{range} = \text{max} - \text{min}.$$

- Deoarece definiția aceasta se bazează doar pe valorile extreme, dacă minimul sau maximum este foarte mare respectiv foarte mic, domeniul nu este reprezentativ pentru variabilitatea datelor.
- Se observă că valorile aberante au o influență directă asupra domeniului.

Dispersia eșantionului

Începem cu măsurile variabilității legate de medie.

- Deviațiile față de medie sunt $(x_i - \bar{x}_n)$.
- O deviație $(x_i - \bar{x}_n)$ este pozitivă (negativă) când x_i este mai mare (mai mică) decât media de selecție.
- Pentru a descrie o valoare medie a deviațiilor s-ar putea utiliza media aritmetică a acestor deviații. Dar pentru căsuma acestor deviații este zero, o astfel de medie este nulă.
- Putem îndepărta acest efect ridicând la pătrat deviațiile și utilizând o medie pătratică în locul uneia aritmetice.

Dispersia eșantionului

Definiția 16

Dispersia eșantionului, s^2 , n fiind dimensiunea eșantionului, este:

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

- Dispersia eșantionului este nenegativă și este zero dacă și numai dacă valorile sunt toate identice.
- Dispersia eșantionului este statistica asociată *dispersiei populației*, notată cu σ^2 .

Dispersia eșantionului

- Motivul pentru care se utilizează $(n - 1)$ ca numitor în definiția dispersiei eșantionului este acela că astfel se obține un estimator nedepășat.
- Media dispersiei eșantionului (văzută ca o variabilă aleatoare) este

$$\begin{aligned} \mathbb{E}[s^2] &= \mathbb{E} \left[\frac{\sum_{i=1}^n (X_i - \bar{x}_n)^2}{n - 1} \right] = \mathbb{E} \left[\frac{\sum_{i=1}^n \left(nX_i - \sum_{j=1}^n X_j \right)^2}{n^2(n - 1)} \right] = \\ &= \frac{\sum_{i=1}^n \mathbb{E} \left[n^2 X_i^2 - 2nX_i \left(\sum_{j=1}^n X_j \right) + \left(\sum_{j=1}^n X_j \right)^2 \right]}{n^2(n - 1)} \end{aligned}$$

Dispersia eșantionului

$$\begin{aligned}
 &= \frac{\sum_{i=1}^n \mathbb{E} [n^2 X_i^2] - 2n \sum_{i=1}^n \mathbb{E} \left[\sum_{j=1}^n X_i X_j \right] + n \mathbb{E} \left[\sum_{j=1}^n X_j^2 + 2 \sum_{i < j} X_i X_j \right]}{n^2(n-1)} \\
 &= \frac{n^2 \sum_{i=1}^n \mathbb{E} [X_i^2] - 2n \sum_{i=1}^n \mathbb{E} [X_i^2] - 2n \sum_{i=1}^n \sum_{j \neq i} \mathbb{E} [X_i X_j]}{n^2(n-1)} + \\
 &\quad + \frac{n \sum_{j=1}^n \mathbb{E} [X_j^2] + 2n \sum_{i < j} \mathbb{E} [X_i X_j]}{n^2(n-1)} =
 \end{aligned}$$

Dispersia eșantionului

$$\begin{aligned}
 & n(n-1) \sum_{i=1}^n \mathbb{E}[X_i^2] - 2n \sum_{i < j} \mathbb{E}[X_i X_j] \\
 &= \frac{n^2(n-1) \sum_{i=1}^n \mathbb{E}[X_i^2] - 2n \sum_{i < j} \mathbb{E}[X_i] \mathbb{E}[X_j]}{n^2(n-1)} \\
 &= \frac{\sum_{i=1}^n \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2}{n} = \sigma^2.
 \end{aligned}$$

Deviația standard a eșantionului

- O formulă mai simplă (exercițiu) pentru dispersia eșantionului este

$$s^2 = \frac{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}{n(n-1)}.$$

Definiția 17

Deviația standard a eșantionului, s , este rădăcina pătrată a dispersiei eșantionului.

- Deviația standard a eșantionului este un estimator deplasat al deviației standard a populației, σ .
- Se poate arăta că media deviației standard a eșantionului este mai mică decât cea a populației, $\mathbb{E}[s] < \sigma$.

Sumarul celor cinci numere

Continuăm cu măsurile ale împrăștierii legate de mediană. Mai întâi sumarul celor cinci numere.

Definiția 18

Sumarul celor cinci numere este compus din

- 1 *min*, cea mai mică valoare din eșantion;
- 2 Q_1 , prima cvartilă;
- 3 Me , mediana;
- 4 Q_3 , a treia cvartilă;
- 5 *max*, cea mai mare valoare din eșantion.

Domeniul intercvartilic

- O metodă grafică de a reprezenta sumarul celor cinci numere: *box-and-whiskers*.

Definiția 19

Cvartila medie este valoarea de mijloc dintre prima și cea de-a treia cvartilă:

$$midq = \frac{Q_1 + Q_3}{2}.$$

Domeniu intercvartilic este diferența dintre prima și cea de-a treia cvartilă:

$$IQR = Q_3 - Q_1.$$





Valori aberante

- *Valorile aberante* sunt acele valori din eșantion care pot fi considerate prea mici sau prea mari față de "tabloul" general al eșantionului.
- Evident, valorile aberante sunt legate de variabilitatea datelor. În mod obișnuit aceste valori vin din erori de măsură, dar pot avea și cauze naturale.
- Câteodată aceste valori aberante (dacă sunt datorate măsurilor) pot fi eliminate din eșantion înainte de orice altă analiză statistică.
- Vom avea două reguli de determinare a valorilor aberante, deoarece și variabilitatea datelor se măsoară în două feluri.

Valori aberante

- Prima regulă este legată de medie. Pot fi considerate valori aberante acele valori ale eșantionului care nu aparțin intervalului $(\bar{x}_n - 2s, \bar{x}_n + 2s)$.
- A doua regulă se numește regula $1.5 * IQR$ și spune că o valoare este aberantă dacă nu aparține intervalului $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$.

Bibliography

-  Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.
-  Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.
-  Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
-  Spiegel, M. R., L. J. Stephens, *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw Hill, 3rd edition, 1999.