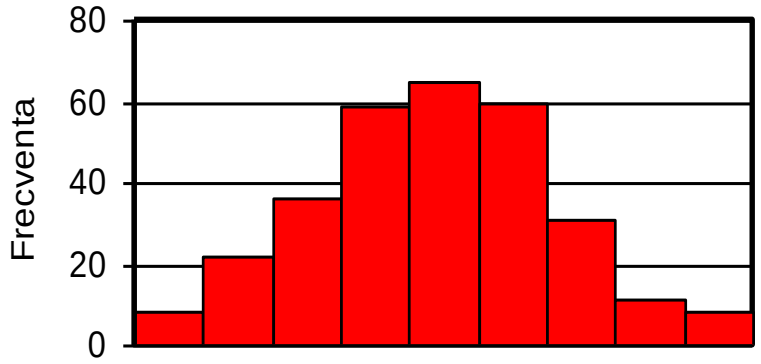
Tipuri de distribuţii şi compararea statistică

Forma distribuțiilor de frecvențe

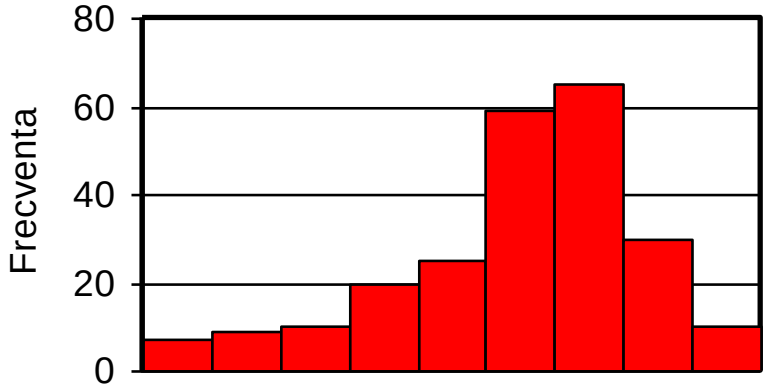
- Simetrice (normale)
- Deviate
  - pozitiv
  - negativ

Majoritatea distribuțiilor întâlnite în practica medicală sunt simetrice sau deviate pozitiv

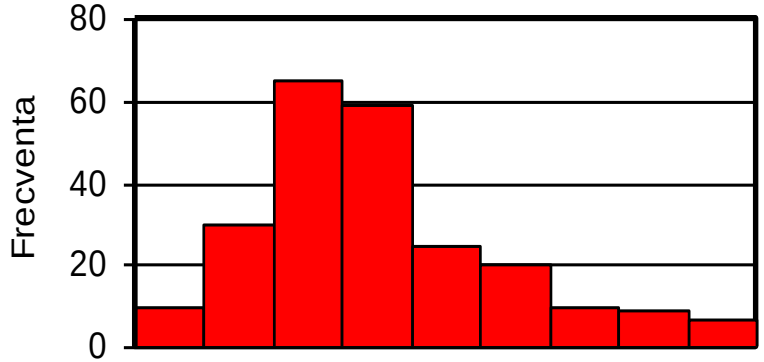
Distribuții de frecvențe



Normală

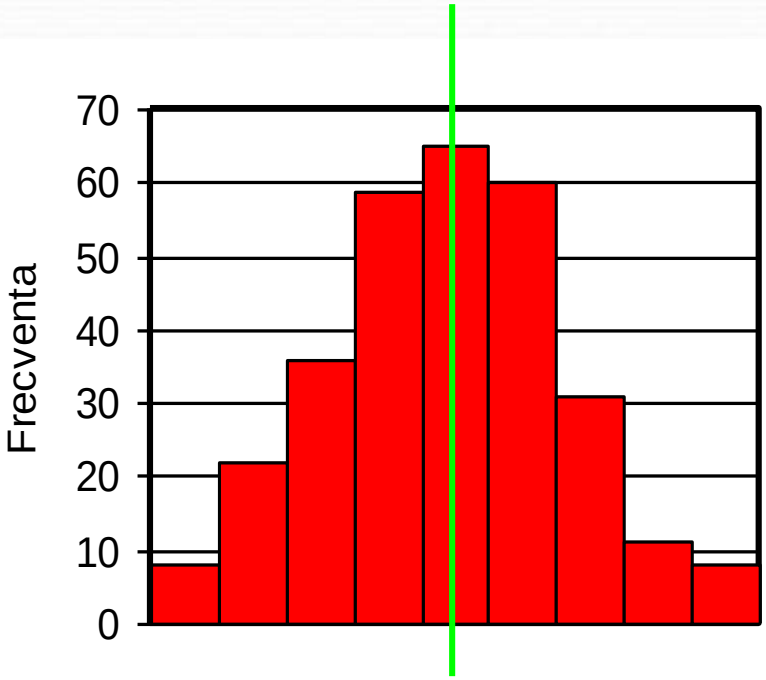


Deviată negativ

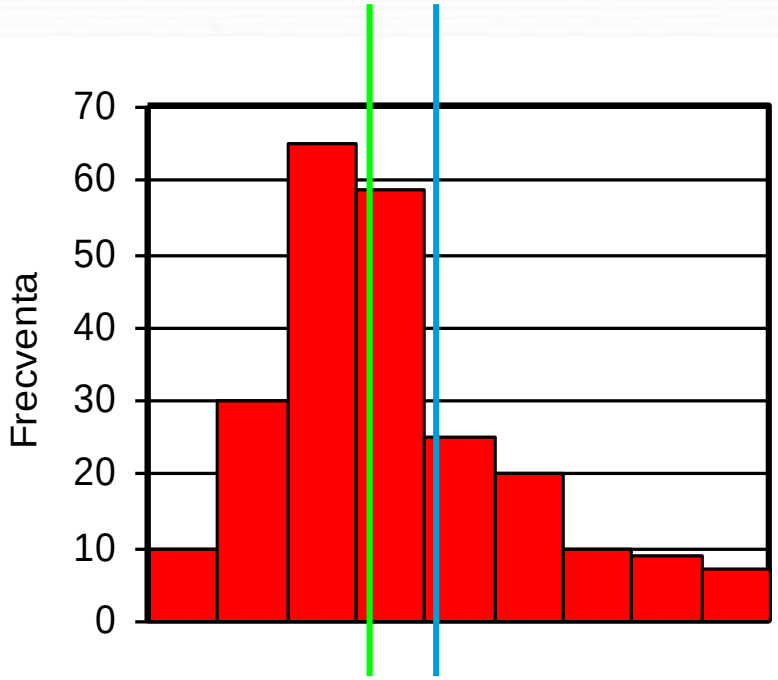


Deviată pozitiv

Localizarea Mediei şi Medianei



media = mediana



mediana ≠ media

Medie sau Mediană?

Când distribuția frecvențelor setului de date este:

- *simetrică* - media este în mod obişnuit cea mai potrivită unitate de măsură.
- *deviată* - mediana este în mod obişnuit cea mai potrivită unitate de măsură.

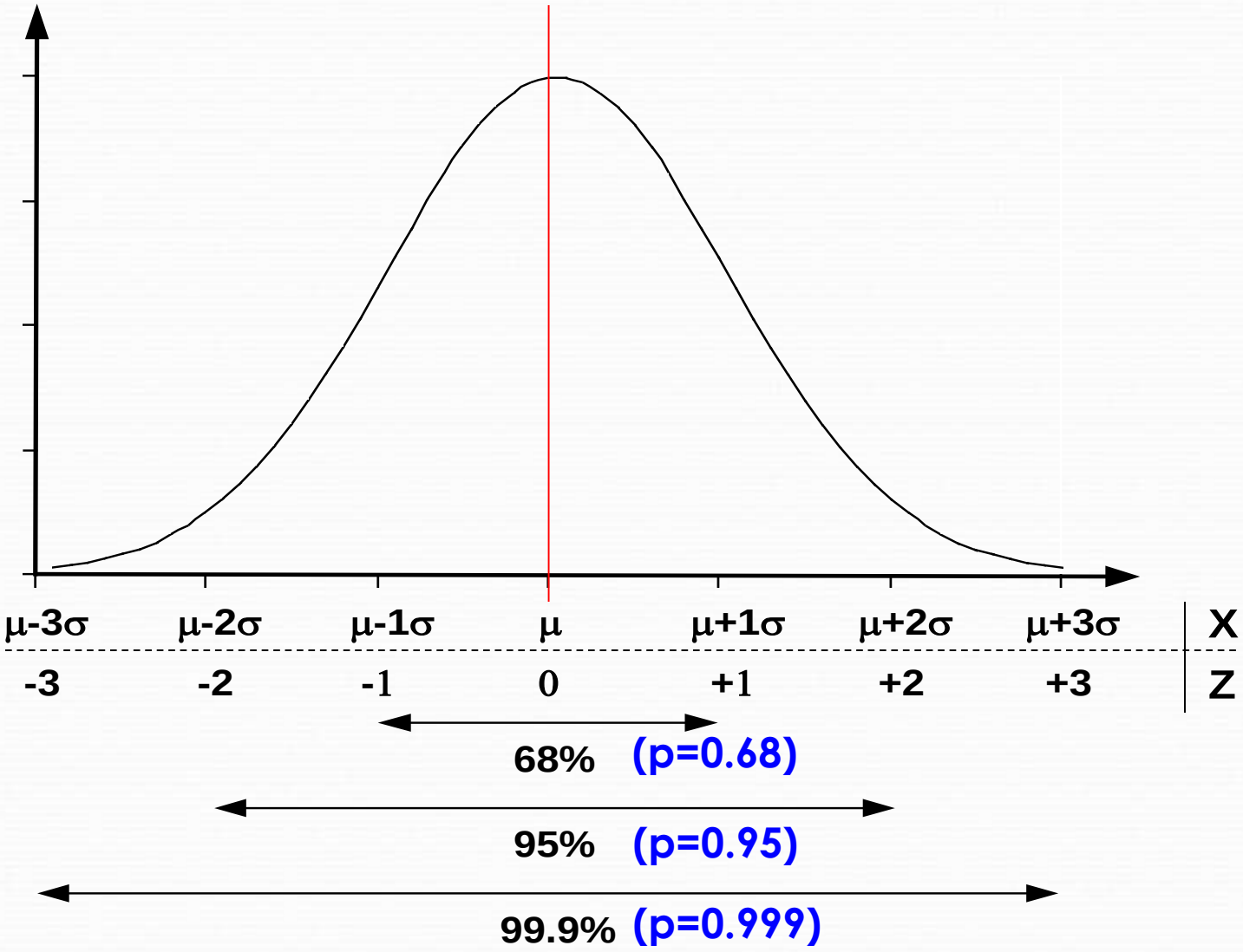
Distribuția normală(I)

- Numită şi distribuție Gaussiană este una din cele mai importante distribuții în statistică.
- Distribuțiile multor măsurători medicale aproximează distribuția normală ex. Tensiunea arterială, Colesterolul sanguin.
- Se defineşte printr-o medie şi o variantă şi se caracterizează ca fiind în formă de clopot şi simetrică în jurul mediei.
- Va fi mai plată dacă varianța este mai mare şi mai ascuțită dacă varianța este mai mică.

Distribuția normală(II)

- Probabilitatea ca o varabilă aleatoare Normală să ia o valoare între:
  - medie şi 1 deviație standard de ambele părți este 0.68,
  - medie şi 1.96 x deviația standard de ambele părți este 0.95
  - medie şi 2.58 x deviația standard de ambele părți este 0.99

Distribuția normală



Distribuția normală

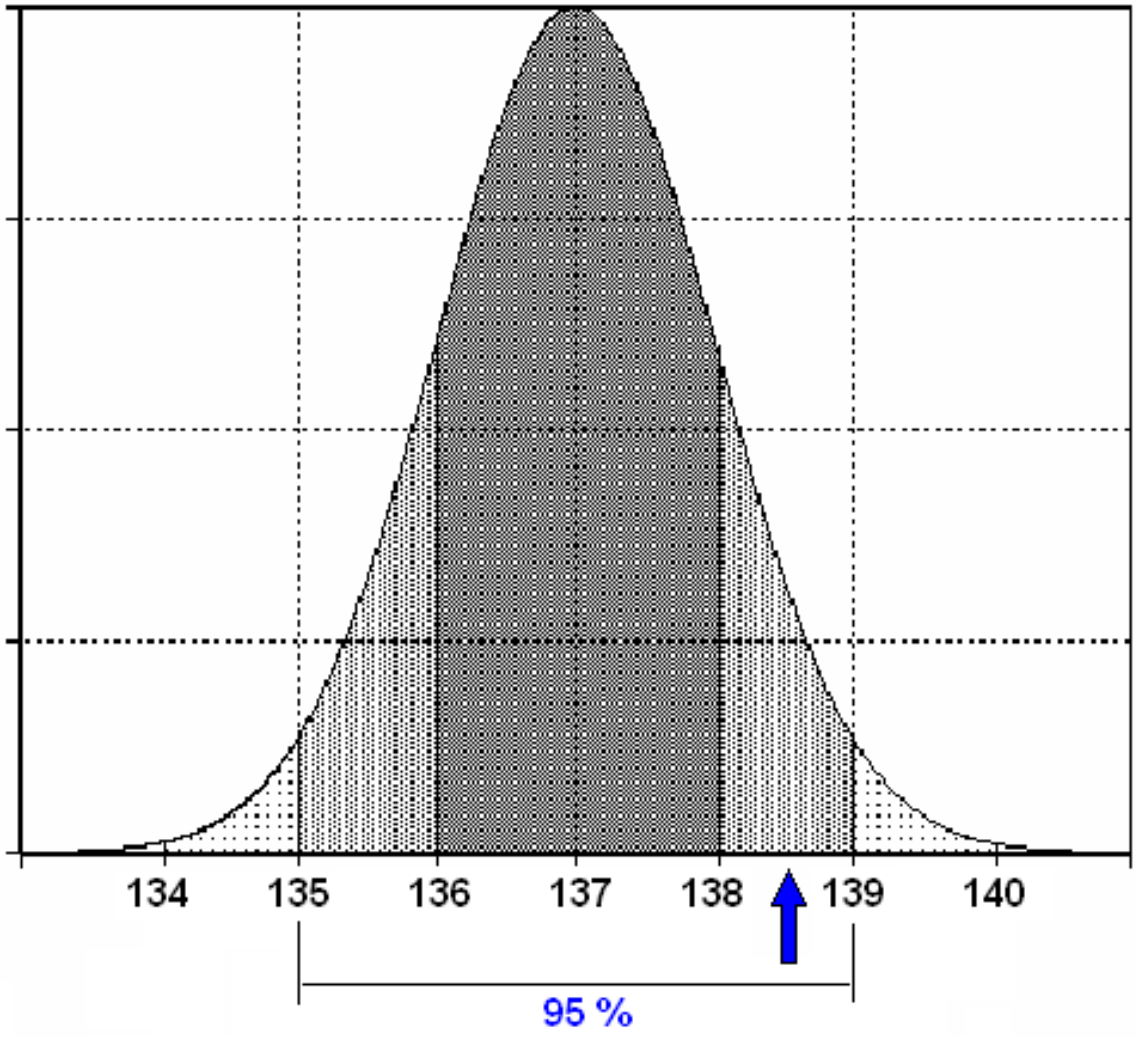
- Distribuția normală apare pentru date continue când media şi varianța populației sunt cunoscute.
- Când acestea nu sunt cunoscute şi când avem doar informații obținute pe eşantion despre medie şi varianță atunci folosim distribuția testului t Student. Pe măsură ce mărimea eşantionului creşte, el tinde spre o distribuție normală.
- De fapt, multe distribuții tind spre Normalitate, mai ales dacă se strâng numeroase eşantioane. De aceea, distribuția Normală a devenit o parte importantă a teoriei statistice.

DIFERENTE SEMNIFICATIVE SI NESEMNIFICATIVE

a) Exemplu:

- BAIETI FETE
- n = 25 n = 25
- X = 137 cm X = 138.5
- s = 5 cm s = 5
- s_x = 1 cm s_x = 1
- (135, 139) …95%

g) Distributia mediilor esantioanelor



DIFERENTE SEMNIFICATIVE SI NESEMNIFICATIVE

a) Exemplu:

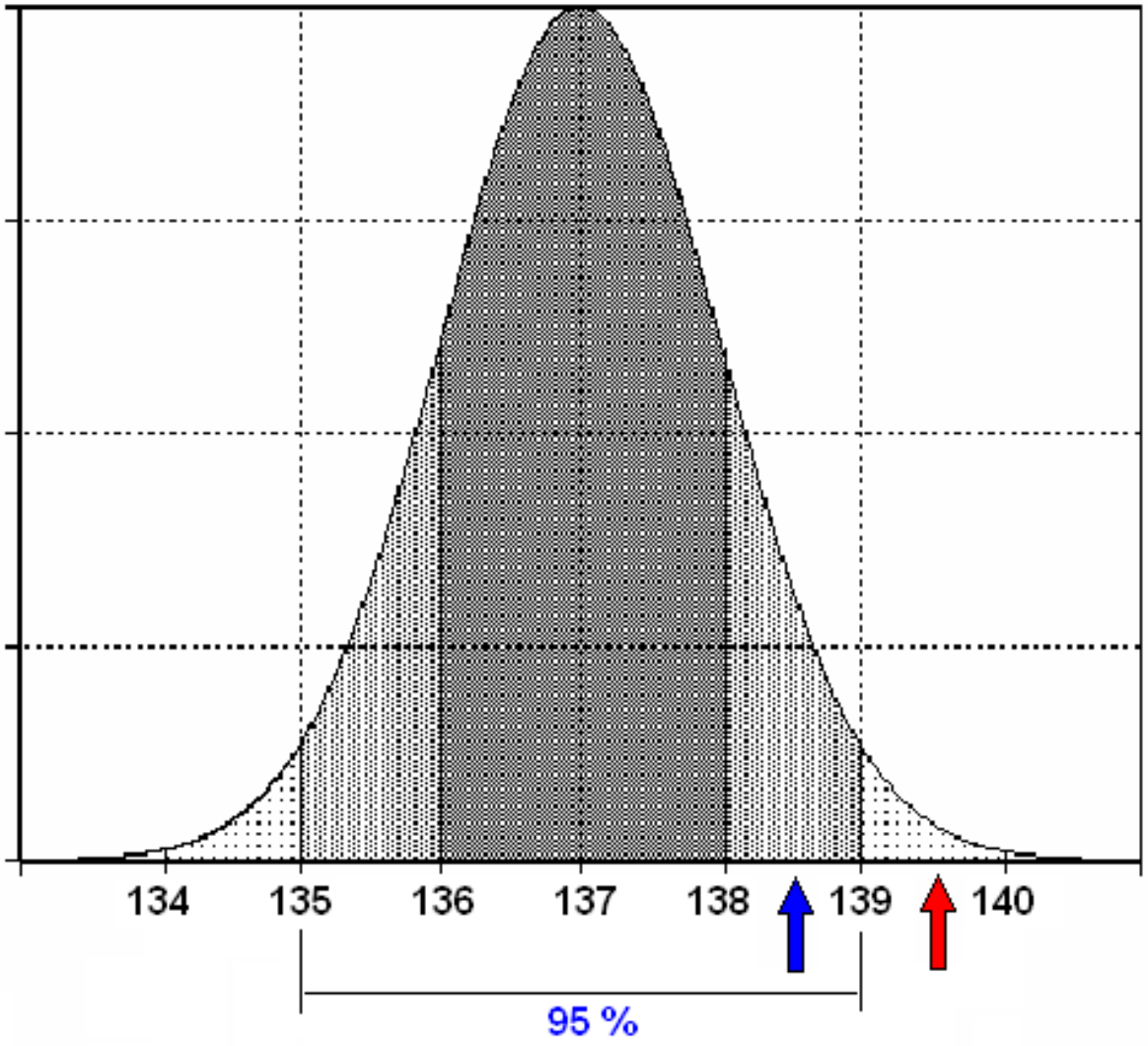
- BAIETI FETE
- n = 25 n = 25
- X = 137 cm X = 138.5
- s = 5 cm s = 5
- s_x = 1 cm s_x = 1
- (135, 139) …95% **dif. nesemnificative**

DIFERENTE SEMNIFICATIVE SI NESEMNIFICATIVE

a) Exemplu:

- BAIETI FETE
- n = 25 n = 25
- X = 137 cm X = 138.5 X = 139.5
- s = 5 cm s = 5
- s_x = 1 cm s_x = 1
- (135, 139) …95% dif. nesemn.

g) Distributia mediilor esantioanelor



DIFERENTE SEMNIFICATIVE SI NESEMNIFICATIVE

a) Exemplu:

- BAIETI FETE
- n = 25 n = 25
- X = 137 cm X = 138.5 X = 139.5
- s = 5 cm s = 5
- s_x = 1 cm s_x = 1
- (135, 139) ...95% d. nesemn. dif. semnificative

b) DEFINITII

- **DIFERENTE NESEMNIFICATIVE**
- **Au probabilitate mare sa apara din intamplare**
- **Cauza: Variabilitatea de esantionare**
- **Cele doua esantioane apartin aceleeasi populatii**

- **DIFERENTE SEMNIFICATIVE**
- **Au probabilitate mica sa apara din intamplare**
- **Trebuie atribuite altei cauze**

Teorema limită centrală (1)

- Dacă se extrag eșantioane de volum n dintr-o populație, atunci pentru valori "mari" ale lui n mediile de eșantion sunt distribuite (aproximativ) normal.

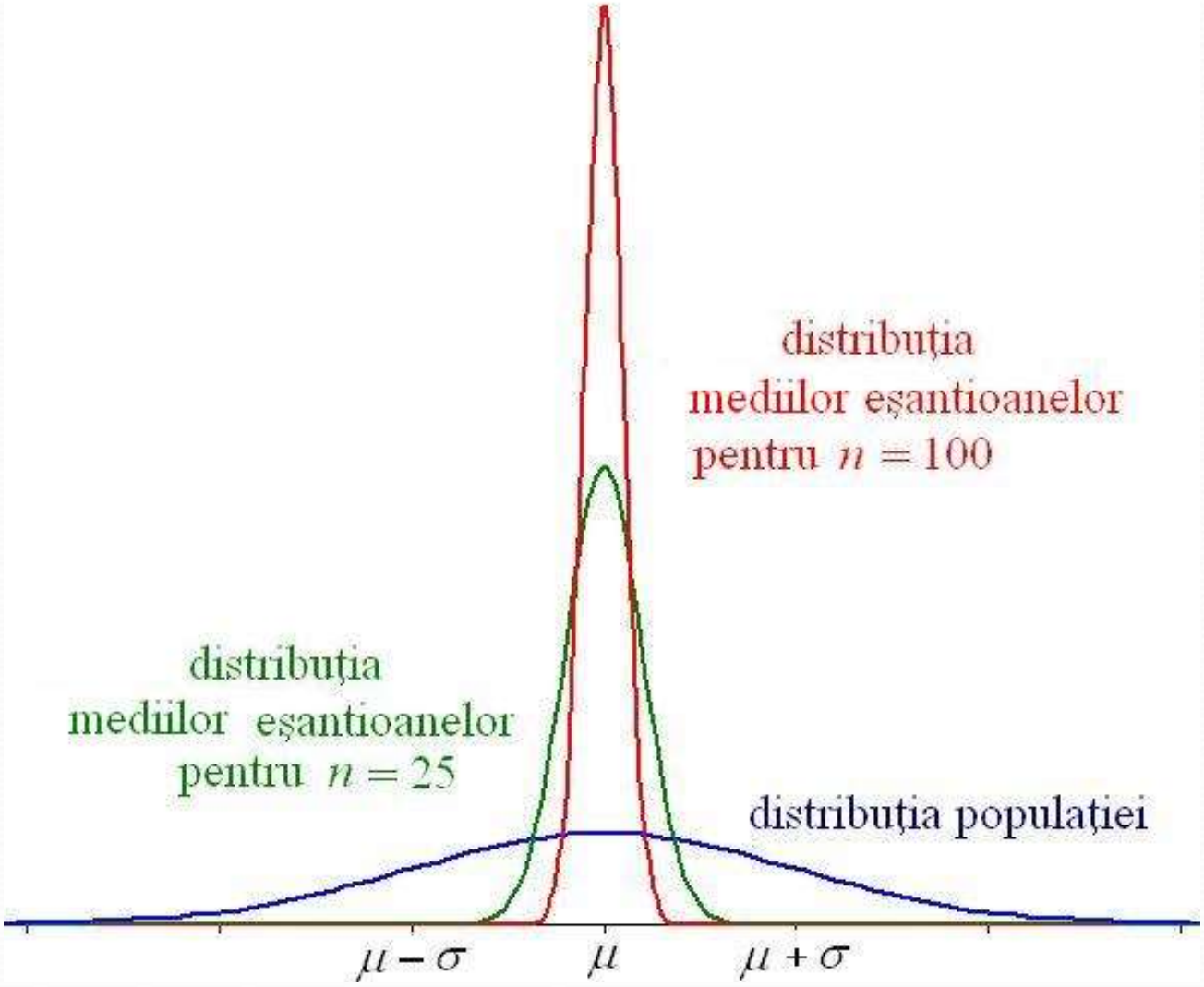
- În caz că X are o distribuție normală $N(\mu, \sigma^2)$, atunci

  $$M = (X_1 + X_2 + \ldots + X_n)/n$$

  are o distribuție normală $N(\mu, \sigma^2/n)$.

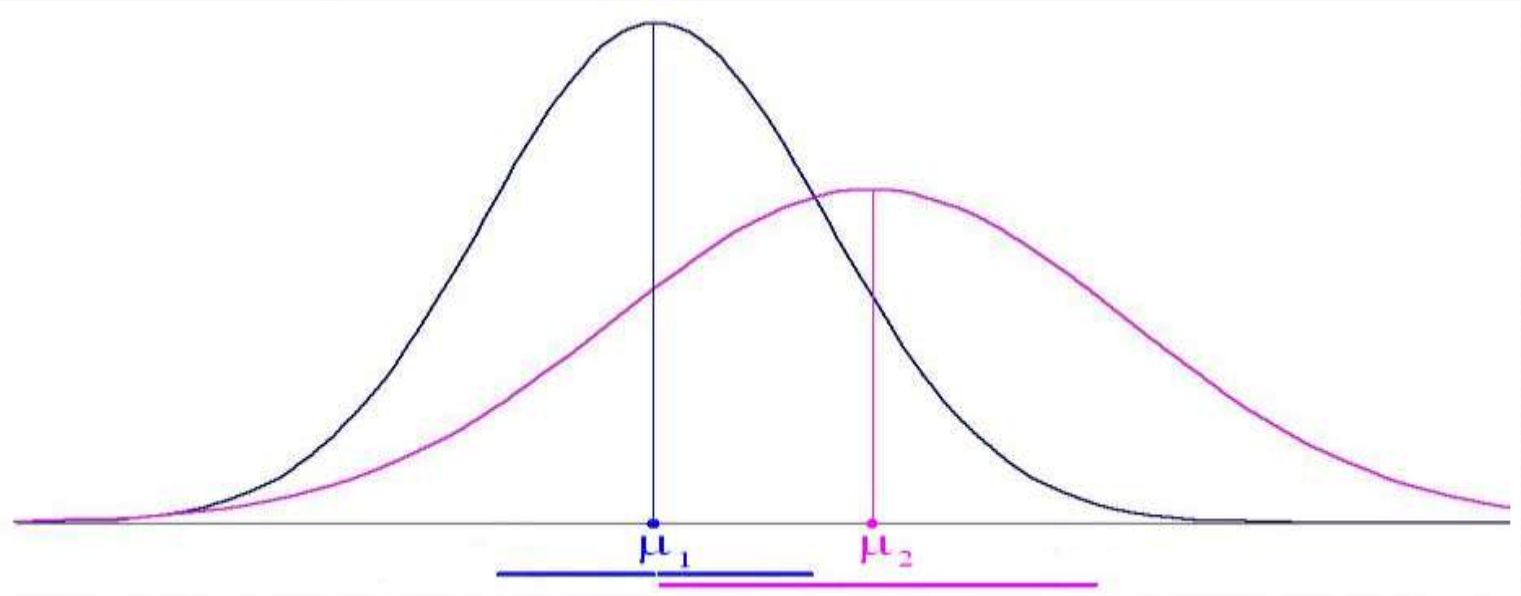
- Iar dacă variabila aleatoare X este distribuita aproximativ normal, atunci M va fi distribuita normal chiar și pentru valori "mici" ale lui n.

Teorema limită centrală (2)



Compararea populaţiilor (1)

- Atunci când comparăm între ele două populaţii distribuite normal, comparaţia se poate face la nivelul mediilor μ_1 şi μ_2 şi/sau la nivelul varianţelor σ_1^2 şi σ_2^2.



Diferă populaţiile între ele?

Testarea statistică de semnificaţie (1)

- Este o metodă de stabilire a gradului de plauzibilitate ("adevărului") al unei afirmaţii
- Explicaţii necesare:

  testare = procedură standard de examinare, aplicată nediscriminatoriu tuturor indivizilor din eşantion

  statistică = formulă de calcul cu datele provenite din eşantion

  semnificaţie = înţeles "precis"

  1 este "semnificativ" diferit de 0 ?

  Da, la nivelul km. Nu, la nivelul μm. Răspuns subiectiv.

Testarea statistică de semnificaţie (2)

- Robert A. Fisher (1925):

  validitatea unei ipoteze stiintifice este stabilita pe baza unui singur test, cu opţiunea de a nu emite o judecată definitivă atunci când rezultatul testului nu este “suficient de limpede”.

Sunt posibile doar două opţiuni:

fie vom “respinge ipoteza nulă”,

fie vom amâna decizia (nu sunt suficiente date pentru a trage vreo concluzie).

Testarea statistică de semnificaţie (3)

- Ipotezele ştiinţifice se refera la populatii teoretice, care au de obicei un număr infinit de indivizi şi sunt reprezentate de distribuţii continue.
- Ipoteza ştiinţifică este înlocuită printr-o ipoteză statistică, exprimată prin intermediul parametrului acelei populaţii (cum ar fi proporţia, media ...)
- Valoarea parametrului este estimată prin exploatarea datelor obţinute dintr-un eşantion extras din populaţie, apoi este comparată cu o valoare "aşteptată".
- Discrepanţa dintre cele două va influenţa "credinţa" noastră în validitatea ipotezei ştiinţifice.

Testarea statistică de semnificaţie (4)

- Ideea testării statistice de semnificaţie este simplă: ipoteza statistică va servi ca alternativa la o alta ipoteza – așa-numita “ipoteză nula” – care este luata în considerare doar pentru a fi respinsă.

Prin acceptarea inițială a adevărului ipotezei nule vor rezulta anumite consecințe logice/statistice, iar acestea vor fi confruntate cu datele observate.

Orice dovadă aflată în contradicție cu ipoteza nulă va servi ca justificare a alternativei.

Testarea statistică de semnificaţie (5)

- Ipoteza alternativă trebuie să exprime o inegalitate sau o discrepanță (în niciun caz o egalitate/coincidență)
- Ipoteza nulă exprimă, în aceiași termeni ca și alternativa, o egalitate.
- Ipoteza alternativă este cea pe care o dorim confirmată ca adevărată
- Notații tradiționale: (H_1) sau (H_a) alternativa
  (H_o) ipoteza nulă

IPOTEZE STATISTICE

- **a) IPOTEZA NULA (DE ZERO)**
  - **$H_0 : X_1 = X_2$**
  - **Nu sunt diferente semnificative intre cele doua valori (esantioane)**
  - **Semnul "=" nu este în sens matematic, ci statistic**
- **b) IPOTEZE ALTERNATIVE**
  - **$H_1 : X_1 \neq X_2$ (bilaterala)**

    $X_1 > X_2$, $X_1 < X_2$ (unilaterala)

- **PRAG DE SEMNIFICATIE**
  - **a) DEFINITIE:**
    - **valoarea probabilitatii sub care incepem sa consideram diferentele ca fiind semnificative**
  - **b) VALOARE UZUALA:**
    - **α = 0.05 = 5 %**
  - **c) NIVEL DE INCREDERE (confidenta)**
    - **1 - α = 0.95 = 95 %**
- **1.4. COEFICIENTUL p**

  p = probabilitatea ca diferentele observate sa fi aparut din intamplare

1.5. DECIZIA

- ***Daca p > 0.05 => Diferente Nesemnificative, (N) , H_0 este acceptata***

- ***Daca p < 0.05 => Diferente Semnificative, (S), H_0 este respinsa***
- ***Daca p < 0.01 => Diferente Foarte Semnificative, (F), H_0 respinsa***
- ***Daca p < 0.001 => Diferente Extrem de Semnificative, (E), H_0 respinsa***

Testarea statistică de semnificaţie (6)

- Exemple de ipoteze ştiinţifice:

  Vârsta “foarte înaintată” este un predictor semnificativ al maladiei Alzheimer

  Medicamentul A ajută pacienţii să se însănătoşească mai bine decât medicamentul B

  Pacienţii îşi revin în urma unui tratament standard

- Exemplu “negativ”:

  Efectele medicamentului E asupra pacienţilor bărbaţi şi femei sunt similare.

Testarea statistică de semnificaţie (7)

- Ipotezele statistice corespunzătoare:

  Incidența maladiei Alzheimer este mai mare la persoanele de vârstă foarte înaintată (prin comparaţie cu persoanele de vârstă înaintată)

  Proporția pacienților însănătoșiți dintre cei trataţi cu medicamentul A este mai mare decât proporția corespunzătoare pentru medicamentul B

  Starea medie de sanatate a pacientilor, în urma unui tratament standard, este mai bună decât înaintea începerii tratamentului

Testarea statistică de semnificaţie (8)

- Ipotezele nule corespunzătoare:

  Incidența maladiei Alzheimer la persoanele de vârstă foarte înaintată este aceeași cu cea la persoanele de vârstă înaintată

  Proporția pacienților însănătoșiți dintre cei tratați cu medicamentul A este egala cu cea corespunzatoare pentru medicamentul B

  Starea medie de sănătate a pacienților, în urma unui tratament standard, nu suferă nicio schimbare

Conform lui R. A. Fisher, ipoteza nulă este "ridicată" – ca un complement al ipotezei alternative – doar pentru a fi respinsă, iar prin respingerea ei vom accepta ca „adevărată" ipoteza științifică inițială

T. S. S. – abordarea clasica (1)

O testare statistica a semnificatiei se efectuează în cinci pași consecutivi:

- Pasul 1: Specificăm ipoteza alternativă, apoi ipoteza nulă.
- Pasul 2: Alegem statistica adaptată situației concrete.
- Pasul 3: Alegem nivelul de semnificație, și pe baza sa calculăm pragul de separare (între valorile "acceptabile" și cele considerate ca "inacceptabile").
- Pasul 4: Calculăm valoarea statisticii, folosind efectiv datele din eșantion (ales aleator).
- Pasul 5: Decidem, prin compararea valorii calculate cu pragul dat de nivelul de semnificație, dacă să respingem sau nu ipoteza nulă.

T. S. S. – abordarea clasica (2)

- Discuția în jurul testării statistice de semnificație începe cu ultimul pas.
- În acesta un decidentul (*d-voastră*) va trebui
  fie să respingă ipoteza nulă H_o
  (și prin urmare să accepte ipoteza alternativă H_a),
  fie să nu respingă pe H_o.
- În realitate H_o este fie adevarata, fie falsa – dar decidentul nu cunoaște situația reală.
- Patru posibilități:

T. S. S. – abordarea clasica (3)

Realitatea (necunoscută)

Decizia

	H_0 este falsa	H_0 este adevarata
Respingem H_0	Corectă!	Eronată (eroare de tipul I)
Nu respingem H_0	Eronată (eroare de tipul al II-lea)	Corectă!

T. S. S. – abordarea clasica (4)

- În testarea statistică de semnificaţie importanţa maxima o are eroarea de tipul I. Probabilitatea ei, numarul

$$\alpha = \text{Prob}(\text{decizie eronatä} \mid H_0 \text{ este adeväratä})$$

este **<u>nivelul de semnificatie</u>** a carui valoare a fost aleasa anterior (la Pasul 3).

Fiecare decident doreşte să păstreze nivelul de semnificaţie cât mai mic posibil – întrucât este de fapt probabilitatea de a face o eroare!

Astfel că valori cum este 0,05 sunt destul de des întâlnite, iar în ştiinţele medicale se recomandă alegerea unor valori mai mici, de exemplu 0,001.

Exemplu (1)

(H_a) Ritmul cardiac mediu al pacienţilor hipertensivi scade în urma administrării medicamentului,

(H_o) Ritmul cardiac mediu al pacienţilor hipertensivi, în urma administrării medicamentului, nu suferă nici o schimbare.

Exprimăm formal ipotezele de mai sus astfel:

(H_a) $\mu_a < \mu_b$

(H_o) $\mu_a = \mu_b$

unde μ_b, respectiv μ_a reprezintă ritmul cardiac mediu înainte, respectiv după administrarea medicamentului.

Exemplu (2)

Datele pe care le obţinem apar în mod natural împerechiate; mai precis, pentru fiecare pacient măsurăm ritmul cardiac înainte (x_b) şi după (x_a) administrarea medicamentului.

Evident, am putea calcula diferenţa $d = x_b - x_a$ şi am putea considera că medicamentul este:

eficace pentru pacientul nostru dacă $d > 0$,

ineficace dacă $d = 0$ (nu se constată schimbare)

şi dăunător dacă $d < 0$.

Exemplu (3)

Să notăm cu δ diferența medie; atunci testarea statistică de semnificație de mai sus este înlocuită prin

(H_a) $\delta > 0$

(H_o) $\delta = 0$

Dacă presupunem că ritmul cardiac al pacienților hipertensivi, și înainte, și după administrarea medicamentului, este distribuit normal, atunci rezultă că diferențele *d* sunt și ele distribuite normal.

Statistica adaptată situației este $t = \dfrac{m}{s/\sqrt{n}}$

în care *m* este media eșantionului diferențelor, *s* este abaterea standard, *n* este volumul eșantionului.

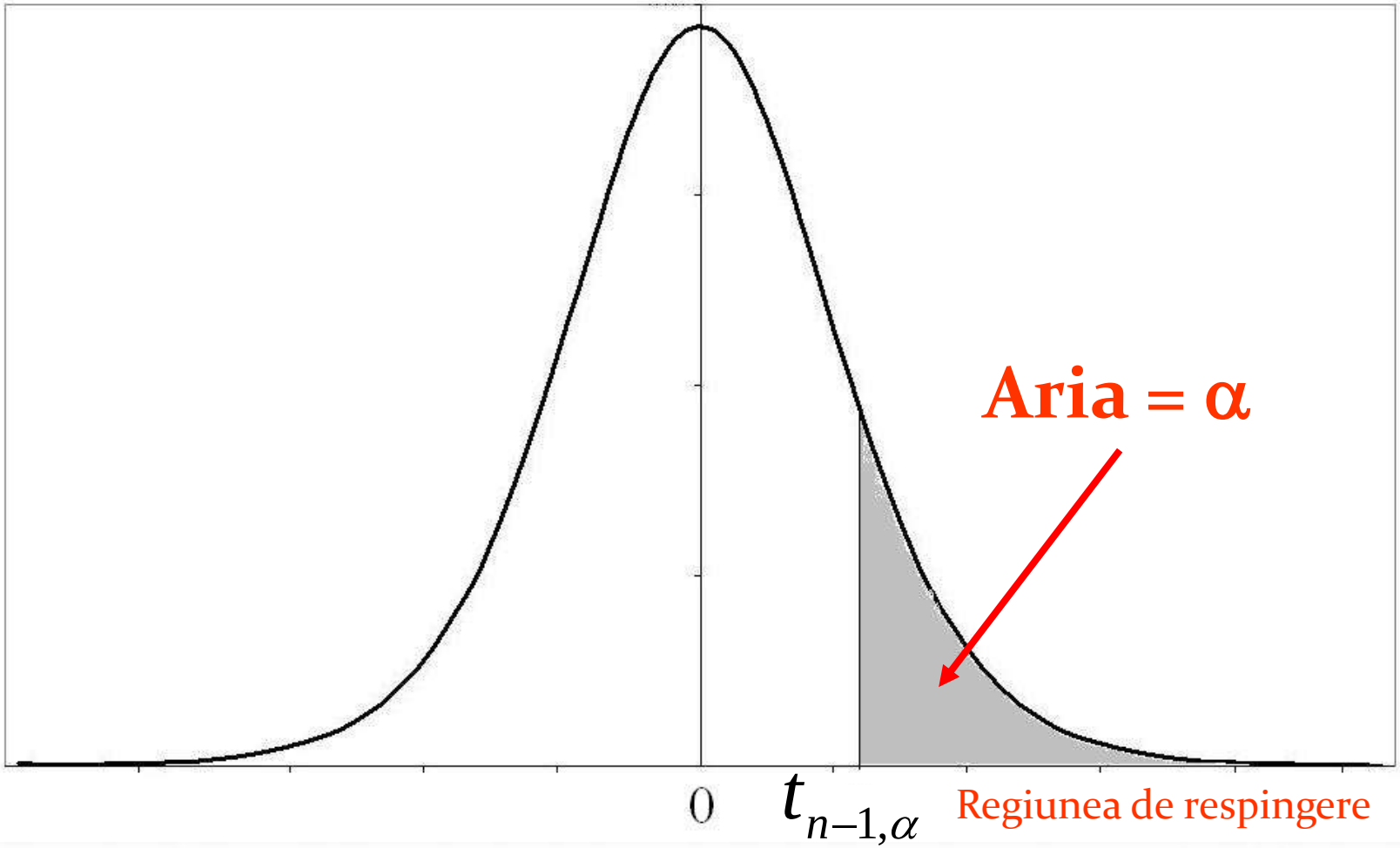
Exemplu (4)

Știm că valorile t sunt distribuite Student, cu $n-1$ grade de libertate.

Marea majoritate a lor se plasează în jurul lui 0. Medicamentul fiind hipertensiv, ne așteptăm ca diferențele d să fie pozitive, deci valorile lui m (și implicit ale lui t) să fie pozitive.

Alegem nivelul de semnificație $\alpha = 0.10$. Ca rezultat vom obține o valoare prag $t_{n-1,\alpha}$

Exemplu (5)



Exemplu (6)

Pentru $n = 8$ avem $t_{n-1,\alpha} = t_{7,0.1} = 1.8946$

Folosim acum datele obținute dintr-un eșantion (8 indivizi):

Înainte (b/m)	După (b/m)	Diferența
66	58	8
69	65	4
…	…	…
73	66	7

Calculăm $m = 3.375$, $s = 4.406$, $t = 2.1667$.

Exemplu (7)

Regiunea de respingere pentru $\alpha = 0.1$

$t_{7,0.1} = 1.8946$ $t = 2.1667$

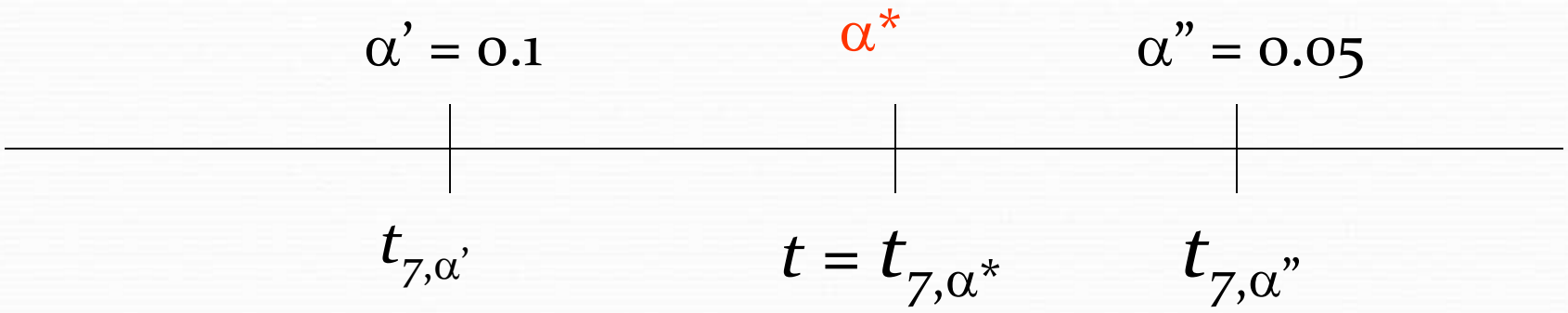
Concluzie: respingem ipoteza nulă!

Regiunea de respingere pentru $\alpha = 0.05$

$t = 2.1667$ $t_{7,0.05} = 2.3646$

Concluzie: nu putem respinge ipoteza nulă!

Exemplu (8)



Nivelul de semnificație particular α^* este cunoscut ca <u>valoarea p</u> a ipotezei alternative.

Este cel mai mic nivel de semnificație care ne permite să acceptăm ca adevărată ipoteza alternativă – prin respingerea ipotezei nule –bazându-ne doar pe datele din eșantionul ales.

Este riscul de a accepta ca adevarata ipoteza alternativa – ceea ce dorim de fapt – atunci când ipoteza nula este cea adevărată.

T. S. S. – abordarea moderna

O testare statistica de semnificatie se efectuează astăzi în doi paşi:

- Pasul 1: Specificăm ipoteza alternativă, apoi ipoteza nulă.
- Pasul 2: Calculăm valoarea p a ipotezei alternative, apoi interpretăm aceasta ca risc.

Pentru calculul valorii p se utilizează în Excel funcţia

TTEST(domeniu1, domeniu2, lateralitate, tip)

tip = 1 înseamnă test pereche

χ^2 ca test statistic de semnificaţie (1)

- Compară două variabile aleatoare (cu câte 2 valori fiecare) și le evaluează independența statistică.
- De exemplu, variabila *X* este “Medicamentul”

  iar valorile sale sunt “MedA”, “MedB”;

  pe de altă parte, variabila *Y* este “Starea pacientului”,

  cu valorile “vindecat”, “nevindecat”.
- Independența statistică a lui *X* și *Y* corespunde dependenței liniare a liniilor (sau a coloanelor) din tabelul extins de contingență.

χ^2 ca test statistic de semnificaţie (2)

- Ipoteza alternativă este că există o asociere între variabile.
- Ipoteza nulă este că nu.
- În Excel, calculul valorii p se face cu funcţia

CHITEST(observate, aşteptate)

- Folosirea funcţiei este precedată de calculul valorilor "aşteptate" într-un tabel de contingenţă "paralel" cu cel ce conţine datele iniţiale.

Testul hi-pătrat: exemplu (1)

- Datele în tabelul de contingență

	Medicament A	Medicament B (Placebo)
Vindecați	**55**	**40**
Nevindecați	**25**	**35**

- Ipoteza alternativă: există o asociere între variabilele "Starea de sănătate" și "Tip medicament"

- Ipoteza nulă: nu există asociere (variabilele sunt statistic independente)

Testul hi-pătrat: exemplu (2)

- Tabelul extins cu totaluri:

	Medic. A	Placebo	Total rânduri
Vindecați	55	40	95
Nevindecați	25	35	60
Total coloane	80	75	155

Testul hi-pătrat: exemplu (3)

- Tabelul "paralel" (calculat în situaţia ipotezei nule adevărate) cu datele obţinute din totaluri:

	Medicament A	Medicament B (Placebo)
Vindecaţi	49.03	45.97
Nevindecaţi	30.98	29.03

$$49.03 = \frac{total\_rând \times total\_coloana}{total\_general} = \frac{95 \times 80}{155}$$

Testul hi-pătrat: exemplu (4)

- Statistica:

$$X^2 = \sum_{celule} \frac{(observ - astept)^2}{astept}$$

are o distribuţie aproximativ χ^2 ((nr.rânduri-1)(nr.coloane-1))

- Valoarea p calculată cu funcţia CHITEST:

0.0489

Concluzie: asocierea este “semnificativă” (clasic)

sau

Există o asociere semnificativă între “Tipul medicamentului” şi “Starea de sănătate”

Testul hi-pătrat: corecţia Yates

Testul dă rezultate “corecte” în caz că toate valorile aşteptate sunt > 5. Altfel, valoarea p este incorectă.

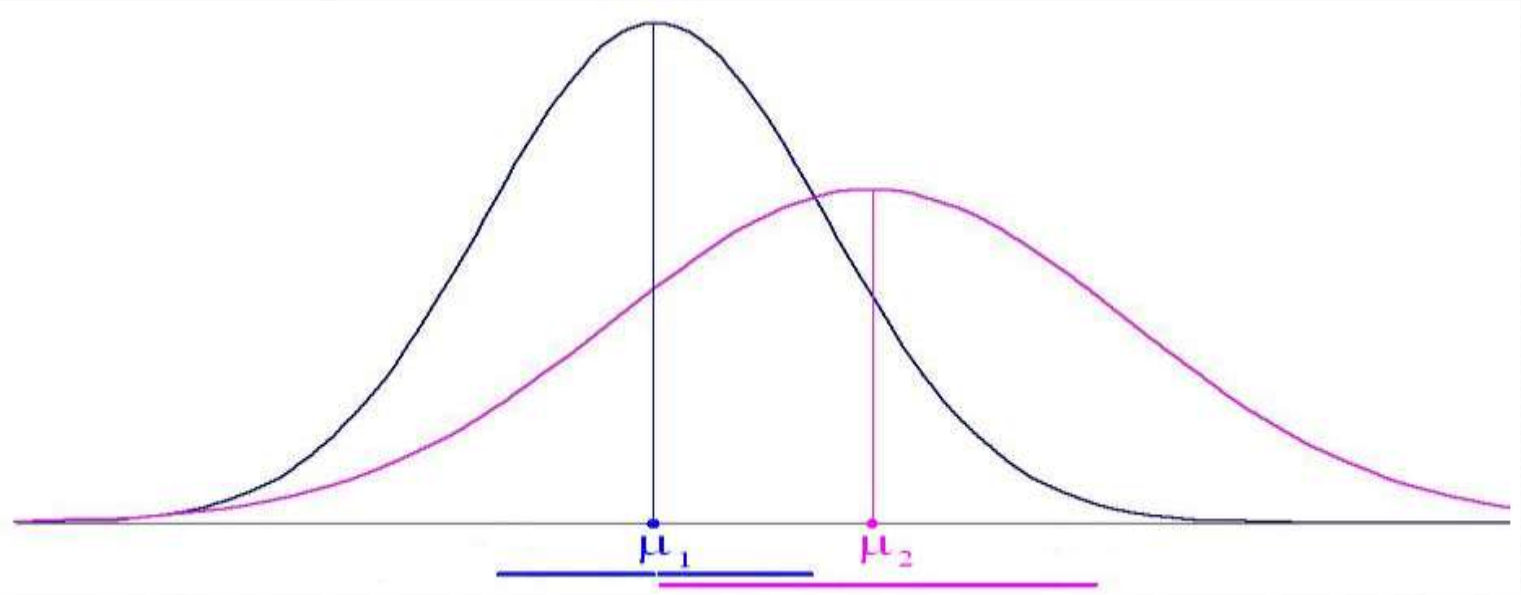
Varianta cu corecţia Yates:

$$X^2 = \sum_{celule} \frac{(observ - astept - 0.5)^2}{astept}$$

produce valori p uşor mai mari, dar mai corecte.

Compararea populaţiilor (1)

- Atunci când comparăm între ele două populaţii distribuite normal, comparaţia se poate face la nivelul mediilor μ_1 şi μ_2 şi/sau la nivelul varianţelor σ_1^2 şi σ_2^2.



Diferă populaţiile între ele?

Compararea populaţiilor (2)

- Nu cunoaştem nici mediile μ_1 şi μ_2 şi nici varianţele σ_1^2 şi σ_2^2. Ele sunt estimate, pe baza a două eşantioane de volume n_1 şi respectiv n_2, prin mediile de eşantion m_1 şi m_2
- Să presupunem că $m_2 > m_1$. Are sens să considerăm ipoteza alternativă

$$(H_a)\ \mu_2 > \mu_1$$

şi să-i calculăm valoarea p.

Reamintim că valoarea p este riscul de a accepta ipoteza alternativă atunci când de fapt ipoteza nulă este adevărată.

Valoarea p o putem calcula imediat, în Excel, cu TTEST.

Compararea populaţiilor (3)

- Ne-am pus problema comparării a două populaţii luând în considerare mediile (sau proporţiile, eventual varianţele lor). Am luat în considerare parametrii care determină populaţiile: mediile (μ), proporţiile (π), varianţele (σ^2).
- Multe dintre metodele de comparaţie se bazează pe ipoteza "fundamentală" că anumite variabile sunt distribuite (aproximativ) normal. Acestea sunt cunoscute sub numele de teste parametrice.
- Există însă situaţii în care:

  fie nu cunoaştem deloc felul în care sunt distribuite variabilele,

  fie normalitatea distribuţiei este încălcată flagrant.

În asemenea situaţii, pentru compararea populaţiilor este posibil să folosim teste care nu presupun nimic despre tipul de distribuţie, cu alte cuvinte **teste neparametrice**.

Compararea populaţiilor (4)

- În **testele de rang** valorile numerice ale variabilelor – obţinute din eşantion – sunt înlocuite prin rangurile lor.
- Să prezentăm, ca exemplu, testul Wilcoxon.
- Ipoteza alternativă de la care plecăm, într-o exprimare generală, este următoarea:

  (H_a) distribuţia valorilor variabilei numerice (care ne interesează) este asimetrică în raport cu o.

Îi vom opune ipoteza nulă:

(H_o) distribuţia valorilor variabilei numerice este simetrică în raport cu o.

Testul Wilcoxon (1)

- Conform teoriei generale, vom încerca să "deducem" consecințe logice ale acceptării adevărului ipotezei nule, apoi să evaluăm dacă datele provenite din eșantion sunt "compatibile" cu consecințele.
- Să începem prin a analiza datele numerice $x_1, x_2, \ldots, x_n$ provenite dintr-un eșantion de volum n. Unele valori vor fi pozitive, altele vor fi negative, câteva chiar egale cu 0. Să presupunem că m dintre ele sunt nenule.
- Vom ordona crescător valorile nenule, luate în modul, apoi le vom înlocui cu rangurile lor:

$$|x_{(1)}| \leq |x_{(2)}| \leq \ldots \leq |x_{(m)}|$$

Testul Wilcoxon (2)

- Să notăm cu T_+ suma rangurilor valorilor pozitive, și cu T_- suma rangurilor valorilor negative. Dacă acceptăm ideea că ipoteza nulă este adevărată, atunci T_+ și T_- n-ar trebui să difere prea mult între ele. Pe de alta parte, suma lor ar trebui să fie egală cu suma tuturor rangurilor. Ar trebui să ne așteptăm ca atât T_+ cât și T_- să fie apropiate de $\frac{m(m+1)}{4}$.

- Cu cât T_+ diferă mai mult de $\frac{m(m+1)}{4}$, cu atât ipoteza nulă devine mai implauzibilă și drept urmare vom fi înclinați să acordăm credit alternativei (H_a).

Testul Wilcoxon (3)

- Calculul valorii p a ipotezei alternative se bazează pe faptul că statistica

$$\frac{T_+ - m(m+1)/4}{\sqrt{m(m+1)(2m+1)/24}}$$

este distribuită (cel puțin pentru valori "mari" ale lui *n*) aproximativ normal standard.

Testul Mann-Whitney este în esență echivalent cu testul Wilcoxon.

Testul Wilcoxon (4)

- Exemplu:

	A	B	C
1	valori	semn	rang
2	1.9	+	4.5
3	1.9	+	4.5
4	3.5	+	7
5	0.5	+	2
6	-3.6	-	8
7	-5.2	-	9
8	0.6	+	3
9	-2.1	-	6
10	-0.4	-	1

- Comanda

MEANS valori semn

din **EpiInfo** da

valoarea p = 0.0139

Analiza varianţei - ANOVA (1)

- Deseori se pune problema comparării prin medii
  a mai mult de două populaţii, sau
  a unei populaţii stratificate în mai mult de două straturi.

În asemenea situaţii se poate aplica o generalizare a testului t pentru două populaţii, cunoscută sub numele de **analiza varianţei** sau testul ANOVA.

- Prima aplicare a analizei varianţei s-a făcut într-o situaţie în care se analizau recoltele obţinute în urma tratării solului cu diferite feluri de îngrăşăminte. Se păstrează, tradiţional, unele dintre notaţiile/noţiunile folosite atunci (cum este "media tratamentului").

Analiza varianţei - ANOVA (2)

- Pentru a explica modul în care se efectuează analiza varianţei, să luăm în considerare mai multe populaţii, fiecare populaţie având o medie şi o varianţă proprie (evident, necunoscute).
- Extragem, din fiecare populaţie, câte un eşantion, conform schemei următoare:

Analiza varianţei - ANOVA (3)

Populaţia 1		Populaţia k		Populaţia K
media μ_1		media μ_k		media μ_K
varianţa σ_1^2		varianţa σ_k^2		varianţa σ_K^2
Eşantion de volum n_1		Eşantion de volum n_k		Eşantion de volum n_K
media de eşantion m_1		media de eşantion m_k		media de eşantion m_K
varianţa de eşantion s_1^2		varianţa de eşantion s_k^2		varianţa de eşantion s_K^2

Analiza varianţei - ANOVA (4)

Analiza varianţei se efectuează pentru o ipoteză nulă

(H_o): nu există diferenţe între mediile populaţiilor

care va trebui respinsă, pentru a se confirma ipoteza alternativă

(H_a): cel puţin două dintre mediile diferă între ele (adică cel puţin două dintre populaţii diferă prin medii).

Ca de obicei în problemele de testare de ipoteze, admitem pentru moment că ipoteza nulă ar fi adevarata, şi deducem consecinţe logice ale ei.

Analiza varianţei - ANOVA (5)

Exemplu: acţiunea unui medicament asupra indivizilor din patru categorii de vârstă, timp de 60 de zile, exprimată în scăderea procentuală a nivelului colesterolului:

Sub 20 ani	**20-39 ani**	**40-59 ani**	**Peste 60 ani**
15	22	17	13
17	25	22	8
31	20	28	19
7	36	15	16
19	22	10	22
20	12	2	media = 15.60
media = 18.17	9	8	
	41	media = 14.57	
	17		
	media = 22.67		

Analiza varianţei - ANOVA (6)

Avem $N = 27$, $K = 4$.

Rezultatele oferite de ***EpiInfo*** sunt urmatoarele:

ANOVA, a Parametric Test for Inequality of Population Means
(For normally distributed data only)

Variation	SS	df	MS	F Statistic
Between	305.4376	3	101.8125	1.3414
Within	1745.7576	25	75.9021	
Total	2051.1852	26		

Sum of Squares

P-Value = 0.2822

Mean of Squares

Analiza varianţei - ANOVA (7)

Valoarea p fiind 0.2822, respingerea ipotezei nule este improprie (chiar dacă discrepanța între medii ni s-ar părea suficient de mare).

Nu dispunem de suficiente date pentru a trage concluzia că scăderea procentuală a nivelului colesterolului depinde de categoria de vârstă.

(Dar nici nu putem trage concluzia că nu depinde de categoria de vârstă!)

TESTE STATISTICE

Teste uzuale

- ***2.1.a. O MEDIE EXPERIMENTALA (n>30) CU O VALOARE TEORETICĂ (media populaţiei)***
- **TESTUL Z**
- $H_o : X_E = \mu_o$
- condiţii: σ cunoscut, distribuţie normala
- **TEST DE SEMNIFICATIE, PARAMETRIC**

- ***2.1.b. O MEDIE EXPERIMENTALA (n<=30) CU O VALOARE TEORETICĂ (media populaţiei)***
- TESTUL t (nepereche, “pooled”)
- $H_o : X_E = \mu_o$
- Condiţii: distribuţie normală
- TEST DE SEMNIFICATIE, PARAMETRIC

- ***2.2.a. DOUĂ MEDII EXPERIMENTALE, SERII MARI (n>30)***
- TESTUL z
- $H_o : X_1 = X_2$
- condiţii: varianţele cunoscute, distribuţii normale
- TEST DE SEMNIFICATIE, PARAMETRIC

- ***2.2.b. DOUĂ MEDII EXPERIMENTALE, SERII INDEPENDENTE, MICI (n<=30) (indivizi diferiţi)***
- TESTUL t nepereche (“pooled”)
- $H_o : X_1 = X_2$
- Condiţii: distribuţii normale; 2 variante de lucru:
  - Varianţe egale ($s_1 = s_2$)
  - Varianţe diferite
- TEST DE SEMNIFICATIE, PARAMETRIC

- ***2.2.c. DOUĂ MEDII EXPERIMENTALE, SERII PERECHE , MICI (n<=30)***
  (aceeaşi indivizi, două condiţii diferite)
- TESTUL t pereche (“paired”, “matched”)
- $H_o : X_1 = X_2$
- Condiţii: distribuţii normale
- TEST DE SEMNIFICATIE, PARAMETRIC

2.2.d. DOUĂ SERII CU DISTRIBUTIE NECUNOSCUTA SAU NEGAUSSIANA

- TESTUL MANN – WHITNEY (‘u’)
- $H_o : X_1 = X_2$
- TEST DE SEMNIFICATIE, NEPARAMETRIC

- ***2.3. RANGURI - DOUA SERII***
- ***a) SERII INDEPENDENTE (nepereche)***
- TESTUL WILCOXON - 'RANK SUM' (suma rangurilor)
- ***b) SERII DEPENDENTE (serii pereche)***
- TEST WILCOXON - 'SIGN - RANK' (semn-rang)
- $H_0 : Me_1 = Me_2$
- TEST DE SEMNIFICATIE, NEPARAMETRIC, PENTRU VARIABILE ORDINALE