STATISTICĂ INFERENȚIALĂ

- Statistica inferențială trage concluzii valabile pentru populație din datele unuia sau mai multor eşantioane, folosind calcule probabiliste. Fapte cunoscute → generalizare la populație.
- Fără suportul probabilităților, e posibil ca un efect să fie considerat sistematic, când de fapt el este aleator (de exemplu, k succese consecutive). Alteori, dimpotrivă, efecte sistematice pot trece neobservate.
- Exemple de inferență statistică: interval de încredere pentru estimarea valorii unui parametru; teste de semnificație pentru evaluarea unei aserțiuni (ipoteze). Aceste paradigme arată *ce s-ar întâmpla dacă metoda de inferență s-ar aplica de multe ori*.
- Metodele de inferență se bazează pe *distribuții de sondaj* (experimente: respectarea caracterului aleator!).
  - datele sunt privite ca provenind din eşantionare aleatoare.

ESTIMAREA PARAMETRILOR

- Estimarea parametrilor se face folosind statistici calculate din eşantioane.

- *Estimare punctuală*: parametrul este aproximat printr-o valoare.
- *Estimare prin interval*: o valoare inferioară şi una superioară, între care se află valoarea parametrului, cu o probabilitate dată.

REPARTIȚIA DE SONDAJ

- Fie o populație C formată din N obiecte, descrise de valorile unei caracteristici X: $a_1, a_2, \ldots, a_N$.
- **În C**, media şi dispersia caracteristicii X sunt:

$$M[X] = (1/N)\cdot\sum_i a_i = \mu;$$

$$D^2(X) = (1/N)\cdot\sum_i (a_i-\mu)^2 = M[(X-\mu)^2] = \sigma^2$$

- Estimarea de parametri (μ, σ^2 etc.) ai populației se face folosind eşantioane aleatoare de volum *n*.
- Pentru $X_i = \{x_i^1, \ldots, x_i^n\}$, fie $\overline{x}_i(n) = (x_i^1 + \ldots + x_i^n)/n$
  - Fiecare x_i^j este o valoare a unei v.a. cu aceeaşi repartiție ca şi X.
- $\{\overline{x}_1(n), \overline{x}_2(n), \ldots\}$ sunt valori succesive ale v.a. a mediilor de sondaj pentru e.a. de volum *n*.
- Repartiția unei astfel de v.a. se numeşte repartiție de sondaj.

REZULTATE PRIVIND ESTIMAREA

- Teoremă. Media şi dispersia mediei de sondaj sunt μ, respectiv σ^2/n.
- Nu este util să exprimăm media şi dispersia mediei de sondaj prin μ şi σ^2 , care sunt necunoscute.
- **Dispersie**: pentru eşantioane de dimensiune n, o aproximare a lui σ^2 este s^2 dată de

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{x_n})^2}{n-1}$$

ESTIMAREA MEDIEI

- **Medie**. Cu aproximarea anterioară pentru dispersie, din inegalitatea lui Cebâşev obţinem:

$$P\{\left|\overline{x_n} - \mu\right| < k \cdot D^2(\overline{x_n})\} \geq 1 - \frac{1}{k^2} \qquad sau$$

$$P\{\overline{x_n} - k \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x_n} + k \cdot \frac{s}{\sqrt{n}}\} \geq 1 - \frac{1}{k^2}$$

- Cu cât k este mai mare, cu atât probabilitatea este mai aproape de 1. Aşadar, o *aproximare a lui* μ este intervalul: $(\overline{x_n} - k \cdot \frac{s}{\sqrt{n}}, \quad \overline{x_n} + k \cdot \frac{s}{\sqrt{n}})$
- Teorema lui Leapunov. $\overline{x_n} = N(\mu, \frac{\sigma^2}{n})$

CARACTERISTICI ALE ESTIMATORILOR (1)

- Media populației poate fi estimată prin media eşantionului (sau mediana, mòdul, media de ordin k, media geometrică, media armonică a acestuia).
- Cum alegem un estimator?
- *Acurateţe*: statistica trebuie să indice valoarea corectă a parametrului.
- *Încredere*: valorile statisticii trebuie să fie cel mai frecvent aproape de valoarea parametrului.

CARACTERISTICI ALE ESTIMATORILOR (2)

Def.1: Statistica $\mathbf{t}_n$ (n - cardinalul eşantionului) este un *estimator nedeplasat* al parametrului θ dacă $M[\mathbf{t}_n] = \theta$.

Def.2: Statistica $\mathbf{t}_n$ este un *estimator consistent* pentru parametrul θ dacă

$\lim_{n\to\infty} P\{ |\mathbf{t}_n - \theta| < \varepsilon\} = 1$ (împrăştierea în jurul valorii parametrului să fie oricât de mică, prin n).

Def.3: Statistica $\mathbf{t}_n$ este un *estimator eficient* pentru parametrul θ dacă $\mathbf{t}_n$ dă valori concentrate mai aproape de valoarea lui θ decât valorile oricărei alte statistici.

- Media de sondaj este un estimator nedeplasat al mediei μ.
- Pentru populații normale, media aritmetică este estimator eficient.

ESTIMAȚII ALE DISPERSIEI

1.- Abaterea medie pătratică. $$s_n^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x_n})^2}{n}$$

- Temă. $M[s_n^2] = ((n-1) / n)\cdot\sigma^2$
- Indicație. Se calculează media v.a. care este pătratul mediei de sondaj.
- Deci, $M[s_n^2] \neq \sigma^2$, adică s_n^2 nu este un estimator nedeplasat al lui σ^2.

2.- Estimatorul s^2 de mai sus este un estimator nedeplasat al dispersiei populației:

- $M[s^2] = M[(n/(n-1)) \cdot s_n^2] = \sigma^2$.

3.- Amplitudinea estimează dispersia pentru eşantioane mici. Este un estimator *instabil (nerobust)*: valorile aberante produc distorsiuni.

INTERVALE DE ÎNCREDERE

- Intervalul de încredere constă dintr-*un interval* rezultat din eşantion şi *un nivel de încredere* (probabilitatea ca intervalul să acopere valoarea parametrului).
- Nivelul de încredere se specifică (de regulă, 0,90 sau mai mult). Se dă de obicei α, unde nivelul de încredere este 1-α (0,95 corespunde la α=0,05).

Definiţie. Un interval de încredere de nivel 1-α pentru parametrul θ este dat de două statistici U şi L astfel încât: $P\{ L \leq \theta \leq U \} = 1 - \alpha$.

- L şi U sunt variabile aleatoare, construite din statistici ale eşantionului: la eşantioane diferite, iau valori diferite.

INTERVAL DE ÎNCREDERE PENTRU MEDIE (1)

- Se dau: un e.a. de dimensiune n şi nivelul $1-\alpha$.
- Se cere: un interval de încredere pentru μ.
- Baza: cunoaştem distribuția mediei eșantionului, anume $N(\mu, \sigma^2/n)$.
- Căutăm numărul z^* pentru care distribuția normală acoperă probabilitatea (aria) $1-\alpha$ pe o distanță de z^* deviații standard de la medie spre stânga şi spre dreapta.
- z^* se găseşte în tabelele distribuției normale standard. $x=z^*$ delimitează, la dreapta sa, aria $\alpha/2$.

VALOARE CRITICĂ

- Exemplu. Dacă nivelul de încredere cerut este 90%, rezultă $\alpha = 0{,}1$; $\alpha / 2 = 0{,}05$. Pentru N(0,1), $x = z^*$ trebuie să lase la dreapta sa aria 0,05 iar la stânga 0,95.

  Din tabel rezultă că z^* aparține intervalului [1,64; 1,65]. Se interpolează $z^* = 1{,}645$ (deviații standard de la medie).

- Definiție. *Valoarea critică* pentru nivelul de încredere 1-α este numărul z^* pentru care dreapta $x = z^*$ delimitează sub curba de densitate normală standard aria $\alpha / 2$.

$$P\{-z^* \le \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \le +z^*\} = 1 - \alpha \qquad \textit{sau}$$

$$P\{\bar{x} - z^* \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z^* \cdot \frac{\sigma}{\sqrt{n}}\} = 1 - \alpha$$

INTERVAL DE ÎNCREDERE PENTRU MEDIE (2)

- **Se selectează un e.a. de dimensiune n dintr-o populație de medie necunoscută μ şi deviație standard cunoscută σ.**
- **Un interval de încredere de nivel 1-α pentru μ este:**

$$\left(\bar{x} - z^* \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \cdot \frac{\sigma}{\sqrt{n}}\right)$$

unde z^* este valoarea critică "superioară $\alpha/2$" pentru N(0,1).

- **Acest interval este exact pentru populații cu distribuție normală şi aproximativ (n>>) pentru alte populații.**

INTERVAL DE ÎNCREDERE PENTRU MEDIE (3)

Nivel încredere	$p=\alpha/2$	z^*
90%	0,05	1,645
95%	0,025	1,960
99%	0,005	2,576

- Lungimea intervalului de încredere este $\mathbf{2}\,\mathbf{z}^* \cdot \frac{\sigma}{\sqrt{\mathbf{n}}}$
- Dacă se cere de la început o anumită lungime w a intervalului, atunci se alege $n = (2z^* \sigma/w)^2$.
- Ceea ce uneori este practic imposibil.

UN EXEMPLU

- Se analizează mostre dintr-un produs farmaceutic pentru a stabili concentrația de substanță activă. Rezultatele măsurătorilor repetate ale aceleiaşi mostre urmează o distribuție normală; media μ a distribuției este chiar concentrația reală a mostrei. Deviația standard a procesului de măsurare este σ = 0,0068 grame pe litru. Se fac trei măsurători ale unei mostre şi se raportează media lor. Dacă cele trei măsurători ale unei mostre au fost 0,8403; 0,8363; 0,8447, să se construiască un interval de încredere la nivelul 99% pentru concentrația reală μ.

  $\bar{x}_3 = 0{,}8404; \quad \alpha = 0{,}01; \quad \alpha/2 = 0{,}005.$

- Rezultă din tabelul N(0,1): z^*=2,576. $z^* \cdot \sigma/\sqrt{n} = 0{,}0101$

Intervalul de încredere este (0,8404-0,0101; 0,8404+0,0101).

- 0,0101 este *eroarea marginală*. $\mu \in$ **(0,8303; 0,8505)**

DEPENDENȚA DE n

- În exemplul anterior, dacă n ar fi fost 1, pentru acelaşi nivel de încredere 99% şi cu măsurătoarea unică egală cu 0,8404, atunci intervalul de încredere ar fi devenit (0,8229; 0,8579) - adică (0,8404-0,1750; 0,8404+ 0,1750).
- Pentru n mai mic se obține un interval mai mare, adică o precizie mai mică: eroarea marginală scade când n creşte.
- Intervalul de încredere poate fi văzut ca:

$$estimarea\_mediei \ +- \ z^{*} \cdot \sigma\_estimării$$

TESTE DE SEMNIFICAȚIE

- Evaluarea statistică a valorii de adevăr a unei aserțiuni (ipoteze), pe baza doar a datelor existente.
- Studiu descriptiv şi studiu inferențial.

EXEMPLU

- O companie producătoare de brânzeturi ia lapte de la mai mulți producători. Există bănuiala că unii producători adaugă apă în lapte pentru a-şi creşte profiturile. Temperatura de înghețare a laptelui variază **normal** cu media μ= -0,545ºC şi deviația standard σ=0,008ºC. Apa în lapte afectează această variație normală, crescând temperatura de înghețare. Se măsoară temperatura de îngheț la cinci loturi succesive de lapte de la acelaşi producător, media obținută fiind $\bar{x}_5 = -0{,}538$. Este aceasta o dovadă că producătorul respectiv adaugă apă în lapte?
- Ipoteza de lucru.Media producătorului este $\mu_p = \mu$= -0,545ºC
- Care este probabilitatea ca pe un eşantion de 5, $\bar{x}_5 = -0{,}538$?
- Soluție. Cu lapte natural, probabilitatea este 0,025.
- Concluzie. 1/40: există dovezi că producătorul adaugă apă.

TESTAREA IPOTEZELOR, CA TIP DE RAȚIONAMENT

- Teste de semnificație (Laplace 1820; Edgeworth 1885).
- "Semnificativ": "**pare** a corespunde unei diferențe reale".
- "Datele sunt departe de ce s-ar întâmpla dacă H_0 ar fi adevărată" este tipul de argument ce duce la respingerea ipotezei H_0.
- Se caută în date prezența unui anumit efect (corelația "mare" din cazul loteriei, creşterea temperaturii de îngheț în cazul laptelui).

1.- Se presupune că efectul nu este prezent;

2.- Se verifică în date tăria dovezilor că ipoteza de la pasul 1.- este falsă;

3.- Dacă se găsesc dovezi puternice la pasul 2.-, atunci se acceptă ipoteza că efectul există.

4.- În caz contrar, se afirmă că "dovezile nu sunt suficient de puternice pentru a respinge ipoteza absenței efectului".

IPOTEZA NULĂ

- Ipoteza care se verifică (că efectul nu este prezent, că nu există nici o diferență, nici o corelație etc.) este ipoteza nulă H_0 *(status quo-ul, "prezumția de nevinovăție").*
- H_0 este o afirmație referitoare la o populație, exprimată prin unul sau mai mulți parametri (în exemplul al doilea, H_0 a fost "μ_p = -0,545ºC").
- Un test de semnificație evaluează cât de puternice sunt, în date, dovezile împotriva ipotezei nule.
- De fapt, când se aplică un test de semnificație, se crede sau se speră că o altă afirmație şi nu H_0 este adevărată. Aceasta este …

IPOTEZA ALTERNATIVĂ

- H_a este ipoteza alternativă (ipoteza de cercetare).
- În exemplul cu laptele, H_a a fost "$\mu_p > -0,545^oC$".
- În exemplul cu loteria, H_a a fost "$\rho \neq 0$", H_0 fiind "$\rho = 0$".
- Ca şi H_0, H_a se referă tot la populație în ansamblu şi, deci, se exprimă tot prin parametri ai acesteia.
- Dificultate: H_a să se exprime simetric sau nu?
- Primul exemplu are H_a simetrică, al doilea are H_a asimetrică.
- Dacă nu e evident altceva, H_a se alege simetrică.

STATISTICA UTILIZATĂ

- Orice test de semnificație folosește valoarea unei *statistici* calculată din date (eșantion). Prin comparație, această valoare dă argumentul pentru respingerea sau nu a ipotezei nule.
- De obicei, statistica folosită estimează parametrul ce apare în ipotezele nulă și alternativă.
- E de așteptat ca valori ale statisticii apropiate de cea din H_0 să ducă la ne-respingerea lui H_0.
- Valori ale statisticii depărtate de cea din H_0 oferă dovezile împotriva ipotezei nule (H_a arată ce sens trebuie să aibă abaterea de la H_0).
- În exemple: r (H_0: $\rho=0$; H_a simetrică – contează $|r| >>0$),

  respectiv $\overline{x_3}$ (H_0: $\mu_p = \mu = -0.545°C$; H_a asimetrică – numai >)

VALORI P

- Ipoteza alternativă este cu atât mai probabilă cu cât faptul dedus din date este mai puțin probabil în condițiile ipotezei nule.
- $P\{\mu_p \geq -0{,}538\ °C / H_0\}$
- Definiție. *Probabilitatea* – calculată considerând $H_0^{x_5}$ adevărată – ca statistica din test să ia o valoare *cel puțin la fel de extremă* ("de depărtată de H_0") ca aceea din date se numeşte valoarea P (probabilitatea critică) a testului.
- Cu cât valoarea P este mai mică, cu atât mai puternică este dovada că H_0 este falsă.

EXEMPLUL II (2)

- Din populația normală de măsurători, de medie μ_p și σ = 0,008°C, se "extrage" un eșantion de 5 măsurători, rezultând $\overline{x_5}$=-0,538°C
- $H_0 : \mu_p = -0{,}545°C$; $H_a : \mu_p > -0{,}545°C$.
- $P\{\overline{x_5} \geq -0{,}538°C / \mu_p = -0{,}545°C\} = ?$
- Cum $\overline{x_5}$ are distribuție $N(\mu_p , \sigma/\text{sqrt}(5))$:

$$P\{\overline{x_5} \geq -0{,}545\} = P\{\frac{\overline{x_5} - (-0{,}545)}{0{,}008/\sqrt{5}} \geq \frac{-0{,}538 - (-0{,}545)}{0{,}008/\sqrt{5}}\} =$$

$$P\{Z \geq 1{,}96\} = 1 - 0{,}9750 = 0{,}025$$

SEMNIFICAȚIE STATISTICĂ

- Se poate decide *a priori* ce prag pentru valoarea P va separa acceptarea ipotezei nule de respingerea acesteia.
- Această valoare-limită se numeşte nivel de semnificație şi se notează cu α.
- Exemplu. α=0,05 înseamnă: se acceptă H_0 dacă, presupunând-o adevărată, datele existente nu ar apărea mai rar decât în 1 din 20 selecții (P≥0,05)
- Definiție. *Datele sunt statistic semnificative la nivel* α dacă se obține o valoare P mai mică sau egală decât α. Atunci se respinge H_0.

SCHEMA UNUI TEST DE SEMNIFICAȚIE

- **I.- Se formulează H_0 şi H_a. H_a este ceea ce se acceptă dacă se respinge H_0.**
- **II.- (opțional) Se stabileşte nivelul de semnificație α - cât de tari să fie dovezile pentru a fi acceptate?**
- **III.- Se calculează, printr-o statistică pe care se bazează testul, cât de mult se potrivesc datele cu ipoteza H_0.**
- **IV.- Se calculează probabilitatea P ca, H_0 fiind adevărată, valoarea statisticii să fie totuşi atât de împotriva lui H_0 pe cât a rezultat din date.**
- **V.- Dacă $P \leq \alpha$, atunci rezultatul testului este semnificativ la nivel α şi ipoteza nulă se respinge.**
- **Dacă $P > \alpha$, atunci testul nu este semnificativ şi ipoteza nulă nu se poate respinge.**
  - Ceea ce **nu dovedeşte că ipoteza H_0 este adevărată.**

TIPURI DE ERORI

STAREA REALĂ A H_a necunoscută CONCLUZIE TEST	H_a **ADEVĂRATĂ**	H_a **FALSĂ**
RESPINGEREA IPOTEZEI NULE **(“rezultat semnificativ”)**	**DECIZIE CORECTĂ**	**EROARE DE TIP I**
NU SE RESPINGE IPOTEZA NULĂ **(“rezultat nesemnificativ”)**	**EROARE DE TIP II**	**DECIZIE CORECTĂ**

PUTEREA STATISTICĂ A UNUI TEST

- La stabilirea nivelului de semnificație, tendința de a evita un tip de eroare duce la creşterea şansei de a face celălalt tip de eroare (0,05 şi 0,01 echilibrează).
- Puterea statistică a unui test este probabilitatea ca testul să dea rezultat semnificativ dacă ipoteza alternativă este adevărată (cu alte cuvinte: probabilitatea de a nu face erori de tip II).
- Nivelul de semnificație α este probabilitatea de a face erori de tip I.
- Stabilirea puterii statistice poate ajuta la determinarea dimensiunii eşantioanelor.