INFERENȚE ASUPRA PROPORȚIILOR

- Proporția, procentajul din populație şi probabilitatea asociată producerii unui eveniment dat implică toate *parametrul binomial p* – probabilitatea teoretică (în populație) de succes.
- Dacă X=B(*n*,*p*), atunci $\mu = np$, $\sigma = \text{sqrt}(np(1-p))$
- X fiind numărul de succese din *n* încercări, definim *p'* ca probabilitatea binomială observată (a eşantionului): *p'*=X/n.
- X este aproximativ normală pentru $n>20$ şi $np>5$, $n(1-p)>5$. Aceasta permite utilizarea unora dintre metodele anterioare pentru inferențe asupra lui *p*.

INFERENȚE ASUPRA LUI *p*

- O valoare observată a lui *p'* aparține unei distribuții de selecție care este: aproximativ normală (în condițiile de mai sus), are media $\mu_{p'} = np/n = p$ şi eroarea standard $\sigma_{p'} = \text{sqrt}(\, p(1-p) / n\,)$.
- Se poate aplica atunci (cu aproximație!) procedura z, cu:

$$z_{esantion} = \frac{p' - p}{\sqrt{p(1-p)/n}}, \quad \text{unde} \quad p' = x/n.$$

- *p* este valoarea din H_0.
- Exemplu. A spune că cel puțin 15% din studenți fumează. B vrea să verifice şi găseşte că dintr-un eşantion de 200 de studenți, 17 fumează. Pentru nivelul de semnificație $\alpha = 0{,}10$, se poate respinge ipoteza lui A?

TESTAREA IPOTEZELOR ASUPRA PROPORȚIILOR

Cu valoarea critică:

- $H_0 : p = 0,15$ (≥). $H_a : p < 0,15$.
- Pentru $\alpha = 0,10$ se găseşte $z^* = -z(0,10) = -1,28$.
- $p' = 17 / 200 = 0,085$.

$$z_{esantion} = \frac{p' - p}{\sqrt{p(1-p)/n}} = \frac{0,085 - 0,150}{\sqrt{0,15 \cdot 0,85 / 200}} = \frac{-0,065}{0,025} = -2,6$$

- Se respinge H_0 : eşantionul aduce dovezi că mai puțin de 15% dintre studenți fumează.

Cu probabilități:

- $P = P\{z < z^* / H_0\} = P\{z < -2,60 / H_0\} = 0,0047$.
- Pentru $\alpha = 0,10$, informația din eşantion este semnificativă. Se respinge H_0.

INTERVAL DE ÎNCREDERE PENTRU PROPORȚII

- Estimarea parametrului *p* – proporția succeselor în populație – se face pornind de la statistica *p'* =x/n – valoarea observată în eşantion.
- Intervalul de încredere este:

$$\left(\quad p' - z(\alpha/2) \cdot \sqrt{\frac{p'(1-p')}{n}}, \quad p' + z(\alpha/2) \cdot \sqrt{\frac{p'(1-p')}{n}} \quad \right)$$

- Se observă că eroarea standard, necunoscută (depinde de *p*), se înlocuieşte cu *p'*.
- În exemplul anterior, cea mai bună estimare punctuală a lui *p* este *p'* = 0,085, iar intervalul de încredere la nivelul α=0,10 este (z(0,05)=1,645): 0,085+-0,033 → (0,052; 0,118)

DIMENSIONAREA EŞANTIONULUI (1)

- Dacă se dă eroarea maximă admisă E pentru estimarea proporției, atunci numărul de indivizi n necesari în eşantion pentru a nu depăşi E, cu nivelul de încredere α cerut este:

$$n = [z(\alpha/2)]^2 \cdot p \cdot (1-p) / E^2 .$$

- *p* se înlocuieşte fie cu o estimare a proporției, fie cu 0,5 (maximizând astfel valoarea lui n de mai sus).
- Câte persoane trebuie incluse într-un eşantion pentru a estima cu eroare cel mult 2%, la un nivel de încredere 0,10, proporția celor ce intenționează să voteze?
- $n \geq (1,645)^2(0,5)(0,5)/(0,02)^2 = 1701,56$. Deci, n=1702.

DIMENSIONAREA EŞANTIONULUI (2)

- Exemplu. Furnizorul unei fabrici afirmă că doar 5% din piesele pe care le livrează spre asamblare au defecte. Să se determine mărimea unui eşantion care să permită estimarea proporţiei de piese defecte, cu o precizie de 0,02 şi la un nivel de încredere de 90%.
- Soluţie. $z(\alpha/2)=1,645$; $E=0,02$;
- $p=0,05$; $1-p=0,95$.
- În consecinţă:
- $n \geq (1,645)^2 \cdot (0,05) \cdot (0,95) / (0,02)^2 = 323,3$
- n=324. Aici însă, se dă valoarea lui *p*.

INFERENȚE ASUPRA DISPERSIEI

- Deseori, dispersia trebuie cunoscută / controlată. De exemplu, o companie de îmbuteliat băuturi trebuie să ştie cât de mult variază nivelul de umplere a sticlelor, chiar dacă media este cea corectă.
- Să presupunem că dispersia 0,0004 este acceptabilă, iar dacă trece de această valoare, se ajustează maşina de umplere.
- $H_0 : \sigma^2 = 0{,}0004\ (\leq)$; $H_a : \sigma^2 > 0{,}0004$.
- Statistica testului:

$$\chi^2 = (n-1)\, s^2 / \sigma^2,$$

unde s^2 este dispersia estimată nedeplasat din eşantion, iar σ^2, valoarea din H_0.

DISTRIBUȚIA χ^2 (1)

- Dacă se extrag eşantioane aleatoare de dimensiune n dintr-o populație normală de dispersie cunoscută σ^2, atunci variabila aleatoare $(n-1)s^2/\sigma^2$ are distribuție χ^2.
- Proprietăți ale distribuției χ^2:
  - Valorile χ^2 sunt pozitive;
  - Curba χ^2 este asimetrică, cu mòdul spre stânga;
  - Pentru df>2, media – aflată la dreapta mòdului – este chiar df (n-1 pentru inferențele prezentate);
  - Există câte o distribuție χ^2 pentru fiecare valoare df.

DISTRIBUȚIA χ^2 (2)

- $\chi^2 = \sum_{1..n} (\xi_k - \mu)^2 / \sigma^2$, ξ_k fiind variabile normale independente $N(\mu, \sigma)$.
- χ^2 are funcția de densitate de probabilitate (pentru $x \geq 0$) definită prin:

$$f_{\chi^2}(x) = \frac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{n/2} \cdot \Gamma(n/2)}$$

- Valorile critice se iau din tabele, sub forma $\chi^2(df; \alpha)$, α fiind aria de la dreapta valorii critice.

EXEMPLUL I

- În exemplul cu îmbutelierea: σ^2 admis este 0,0004. Dacă un eşantion de 28 de sticle dă o dispersie observată de 0,0010, se poate afirma, la nivelul de încredere 0,05, că procesul de îmbuteliere nu este sub control din punct de vedere al dispersiei?
- Regiunea critică se află sub partea dreaptă (>) a curbei de distribuție şi are o arie de 0,05.
- $\chi^2_{\text{critic}} = \chi^2(27;\ 0{,}05) = 40{,}1$.
- $\chi^2_{\text{eşantion}} = (n\text{-}1)\ s^2 / \sigma^2 = 27 \cdot 0{,}001/0{,}0004 = 67{,}5$.
- Concluzie: se respinge H_0 ($\chi^2_{\text{eşantion}}$ se află în regiunea critică).

EXEMPLUL II

- Un test este util dacă, în urma corectării, notele au o împrăştiere suficient de mare pentru a ierarhiza elevii, dar nu într-atât încât diferențele de note să fie prea mari.
- Se afirmă că un test cu punctaj total 100 este util dacă deviația standard este 12.
- La un test de 100 puncte dat la 28 de elevi, deviația standard observată este 10,5. Putem afirma cu nivel de încredere 95% că testul respectiv este "util"?
- H_0: $\sigma=12$; H_a: $\sigma\neq12$. H_a simetrică → două valori critice.
- $\chi^2_{critic1} = \chi^2(27; 0,975) = 14,6$;
- $\chi^2_{critic2} = \chi^2(27; 0,025) = 43,2$.
- $\chi^2_{eşantion} = (n-1)\cdot s^2 / \sigma^2 = 2976,75/144 = 20,6719$
- Decizie. H_0 nu se respinge: testul poate fi considerat "util".

INTERVAL DE ÎNCREDERE PENTRU DISPERSIE

- Capetele intervalului de încredere se obțin din cele două *valori critice*; pentru calculul intervalului de încredere, eşantionul furnizează doar n şi valoarea lui s.
- $\chi^2 = (n-1)\cdot s^2 / \sigma^2 \quad \rightarrow \quad \sigma^2 = (n-1)\cdot s^2 / \chi^2$.
- Dat nivelul α, se obțin valorile critice:
- $\chi^2(df;1-\alpha/2) < \chi^2(df; \alpha/2)$.
- Capetele intervalului sunt:

$(n-1)\cdot s^2 / \chi^2(df; \alpha/2)$; $(n-1)\cdot s^2 / \chi^2(df; 1-\alpha/2)$.

EXEMPLUL II – INTERVAL DE ÎNCREDERE

- Cu datele din exemplul II, intervalele de încredere la nivel α=0,05 pentru dispersia, respectiv deviația standard a populației sunt:
- Dispersie: extremele intervalului sunt date de $(27)(10,5)^2 / 43,2$, respectiv $(27)(10,5)^2 / 14,6$.
- Aşadar, cu 95% încredere estimăm dispersia populației ca fiind între 68,9 şi 203,9.
- Intervalul de încredere pentru deviația standard a populației este dat de radicalii valorilor de mai sus: (8,3; 14,3).

ALTE APLICAȚII ALE LUI χ^2

- Pentru variabile categoriale – tabele ale frecvențelor (eventual, pe intervale sau *clase*).
- Inferențe statistice pentru:

  1.- EXPERIMENTE MULTINOMIALE.

  2.- TESTE DE INDEPENDENȚĂ.

  3.- TESTE DE OMOGENITATE.
- Toate folosesc statistica χ^2: $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$
- O_i, E_i → frecvența i observată, frecvența i aşteptată.
- La eşantionări repetate şi pentru n (numărul total de observații) mare, statistica de mai sus are aproximativ distribuția cu aceeaşi notație.
- Ipotezele statistice sunt mai "libere" – nu se exprimă neapărat direct prin parametri.

INFERENȚE ASUPRA EXPERIMENTELOR MULTINOMIALE

- Să presupunem că testăm ipoteza H_0: "zarul este corect", cu α=0,05. Pentru a o testa, aruncăm zarul de 60 de ori. H_0 ar fi în mod ideal satisfăcută dacă fiecare față a zarului ar fi apărut exact de 10 ori (frecvența aşteptată).
- Observăm frecvențele (în ordinea, irelevantă, a numerelor de pe cele k=6 fețe ale zarului):
- 7, 12, 10, 12, 8, 11.
- Din calcule, rezultă: $\chi^2 = 2,2$.
- α=0,05, iar în cazul multinomial, df=k-1=6-1=5.
- χ^2 (5; 0,05) = 11,1 (cu regiunea critică la dreapta).
- Decizie: Nu se respinge H_0.

EXPERIMENTE MULTINOMIALE

- n repetări în condiții identice ale aceluiaşi experiment;
- rezultatul fiecărei repetări este exact unul din k rezultate posibile;
- fiecare rezultat posibil are ataşată o probabilitate prezumată fixă. $p_1 + p_2 + \ldots + p_k = 1$.
- experimentul dă frecvențele observate $O_1, O_2, \ldots, O_k$ ($O_1+O_2+\ldots+O_k=n$).
- $E_i = n \cdot p_i$ pentru statistica χ^2.

- Ipoteza nulă nu se exprimă neapărat prin parametri.
- Valoarea critică se obține din nivelul de semnificație α şi din numărul de grade de libertate df=k-1.
- Regiunea critică se află la dreapta.

EXEMPLUL III

- Studenții doresc o cât mai mare libertate în alegerea cursurilor. Şapte cursuri similare, predate de cadre didactice diferite, au fost alese de 119 studenți astfel (ordinea este aleatoare):
- 18, 12, 25, 23, 8, 19, 14. Indică datele preferințe pentru anumiți profesori?
- H_0 : "distribuție fără preferințe".
- $p_i = 1/7;$ $\alpha = 0,05;$ $\chi^2(6; 0,05) = 12,6.$
- $\chi^2_{esantion} = (18-17)^2 / 17 + (12-17)^2 / 17 + (25-17)^2 / 17 + (23-17)^2 / 17 + (8-17)^2 / 17 + (19-17)^2 / 17 + (14-17)^2 / 17 = 220 / 17 = 12,9411.$
- Decizie. Se respinge H_0!

TABELE DE CONTINGENȚĂ (1)

- Aranjament de date pe linii şi coloane – două variabile, pentru care se testează (in)dependența sau omogenitatea.

1.- Independența. 300 de studenți, clasificați pe sexe, au fost întrebați în ce domeniu al "artelor liberale" preferă să-şi aleagă cursurile.

Sex	Mat.-Şt.	Şt. Soc.	Şt.Um.	Total
F	35	72	71	178
M	37	41	44	122
Total	72	113	115	300

TABELE DE CONTINGENȚĂ (2)

- H_0 : alegerea cursurilor este independentă de sex.
- H_a : alegerea cursurilor este dependentă de sex.
- Valoarea critică. Numărul de grade de libertate este numărul de celule ce pot fi completate fără restricții dacă se dau totalurile: două în acest caz. În general: (nr_linii – 1)·(nr_coloane – 1). $\chi^2(2; 0,05) = 6,00$.
- Regiunea critică este la dreapta:
- $\chi^2_{esantion} > \chi^2_{critic} \rightarrow$ se respinge H_0
- Probabilitățile $p_{i,j}$ ataşate fiecărei celule: proporționale cu totalurile marginale (ce se întâmplă în general este valabil şi pentru fiecare sub-populație). De exemplu, băieți alegând fiecare domeniu ar trebui să fie: (72/300)·122; (113/300)·122; (113/ 300)·122.
- $p_{i,j} = total\_linie_i \cdot total\_coloană_j / n$

TABELE DE CONTINGENȚĂ (3)

Sex	Mat.-Şt.	Şt.Soc.	Şt.Uman.	Total
F	35 (42,72)	72 (67,05)	71 (68,23)	178
B	37 (29,28)	41 (45,95)	44 (46,77)	122
Total	72	113	115	300

- $\chi^2_{esantion} = (35 - 42,72)^2 / 42,72 + (72 - 67,05)^2 / 67,05 + (71 - 68,23)^2 / 68,23 + (37 - 29,28)^2 / 29,28 + (41 - 45,95)^2 / 45,95 + (44 - 46,77)^2 / 46,77 = 1,395 + 0,365 + 0,112 + 2,035 + 0,533 + 0,164 =$ **4,604 < 6,00!**
- Decizie. Nu se poate respinge H_0 !

TABELE DE CONTINGENȚĂ (4)

- **2. Omogenitate.** Experimentatorul controlează una din cele două variabile pentru a obține totaluri date.
- Exemplu. Se proiectează un sondaj de opinie asupra părerilor despre o lege (pentru / împotrivă), intervievând persoane din mediile urban, suburban şi rural. Proporțiile sunt date (fie ele 2/5, 1/5, 2/5). Opiniile asupra legii diferă în cele trei medii?
- Fie α=0,05. Să presupunem că au fost intervievați 500 de subiecți, cu răspunsurile date în tabel.
- H_0 : proporția celor ce sunt pentru legea respectivă este aceeaşi în mediile urban, suburban, rural.
- H_a : în cel puțin un mediu proporția este alta.

OMOGENITATE - TABELUL

MEDIUL	PENTRU	CONTRA	TOTAL
URBAN	143 (101,6)	57 (98,4)	200
SUBURBAN	13 (50,8)	87 (49,2)	100
RURAL	98 (101,6)	102 (98,4)	200
TOTAL	254	246	500

- df = (3-1)(2-1) = 2. χ^2_{critic} (2; 0,05) = 6,00.
- $\chi^2_{esantion} = (143\text{-}101{,}6)^2 / 101{,}6 + \ldots = 91{,}72$.
- Decizie: Se respinge H_0: proporțiile diferă.