PRELUCRAREA DATELOR DE SONDAJ SI INFERENTA STATISTICA

Testarea ipotezelor statistice

Stud. Master - AMP

ISAIC- MANIU ALEXANDRU
web www.amaniu.ase.ro
e-mail AL.ISAIC-MANIU@CSIE.ASE.RO

Cateva elemente recapitulative

- **Estimare**- operaţie de stabilire, în baza datelor unui eşantion, a valorilor parametrilor repartiţiei populaţiei din care a fost prelevat eşantionul
- ✓ Rezultatul, se poate exprima printr-o valoare unică (estimator punctual), sau printr-un interval(numit frecvent "interval de incredere ")
- ✓ Utilizand doar părţi din populaţie, rezultatele obtinute sunt acompaniate de anumite riscuri
- ❖ Spre deosebire de **statistica descriptivă,** inferenţa foloseşte procedee specifice bazate pe modele matematice (în esenţă, probabiliste) pentru analiza materialului statistic organizat de metodele descriptive

Medie aritmetică de sondaj –

(sample average, sample mean-value**)**

- Raportul dintre suma tuturor valorilor x observate în eşantionul considerat şi numărul total n al acestora:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i^*$$

- Observaţii

În cazul valorilor observate, aranjate în ordine crescătoare sau descrescătoare:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \sum_{i=1}^{n} f_i x_i$$

- în care:

n numărul total al valorilor observate;
n_i frecvenţa absolută corespunzătoare valorii x_i
f_i frecvenţa relativă corespunzătoare valorii x_i

Media caracteristii binare (alternative)

Când valorile variabilei aleatoare **X** sunt proporţia elementelor A sau respectiv a elementelor non A (de exemplu, proporţia cetăţenilor cu intenţie de a se prezenţa la vot şi absenteiştii), atunci valorile tipice respective tipice sunt:

media: **M**(X) **= p**

dispersia de sondaj: **Var²**(X) = **pqn**

în care:

n – volumul eşantionului;

p şi **q –** proporţiile respective ale elementelor A şi non A.

Dispersia de sondaj - (S^2)

- Momentul centrat de ordinul doi:

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

- Valoarea numerică a acestui indicator sintetic caracterizează împrăştierea repartiţiei statistice
- Dispersia de sondaj poate fi folosită ca estimaţie aproximativă a dispersiei din populaţia originară, considerându-se formula corectă:

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

- **Eroarea în estimarea** unui parametru (^ θ - θ) unde ^θ este rezultatul estimării, iar θ este parametrul estimat

Interval de încredere

(Confidence interval)

- **1. Interval de încredere unilateral** *(One-sided confidence interval; Intervalle de confiance unilatéral)*. Dacă Z este o funcţie a valorilor observate, iar **θ** este un parametru de estimat al populaţiei, astfel ca probabilitatea este cel puţin egală cu o probabilitate fixata de cercetator
- intervalul cuprins între cea mai mică valoare posibilă a lui θ şi Z (sau intervalul între Z şi cea mai mare valoare posibilă a lui θ) este intervalul unilateral de probabilitate **P** pentru θ .

2. Interval de încredere bilateral

- Dacă Z_1 şi Z_2 sunt două funcţii ale valorilor observate, iar θ este un parametru estimat al populaţiei, astfel ca probabilitatea este cel puţin egală cu $1-\alpha$, [unde $1-\alpha$ este un număr fixat, pozitiv şi mai mic decât 1], intervalul dintre Z_1 şi Z_2 este un interval de încredere bilateral de pentru θ
- Limitele Z_1 şi Z_2 ale intervalului de încredere sunt statistici care, în general, au valori diferite de la un eşantion la altul

Erori in verificarea ipotezelor statistice

❖Eroare de genul intai : ipoteza H se respinge, când ea este adevărată

❖Eroare de genul al doilea: ipoteza H se admite, când ea este falsă

Probabilităţile de a fi comise erori sunt:

probabilitatea erorii de genul întâi

✓**risc de genul I (α)** şi respectiv

probabilitatea erorii de genul al doilea-

✓**risc de genul II (β)**

Ipoteza nulă şi ipoteza alternativă

- Afirmaţii asupra unuia sau mai multor parametri, sau asupra unor repartiţii, care urmează a fi validate prin **teste statistice**.
- Decizia asupra ipotezei nule este luată pe baza unui test statistic.
- Testul este construit cu elemente aleatoare, iar decizia comportă un anumit risc de eroare.

$$H_0:(p_1 = p_2) \qquad H_1:(p_1 \neq p_2)$$

- Ipoteza nulă (H_0) se referă la afirmaţii supuse testării, în timp ce ipoteza alternativă (H_1) se referă la afirmaţii care vor fi acceptate dacă se respinge ipoteza nulă.

Test statistic (Statistical test)

- ◊ Procedura statistică prin care se decide dacă ipoteza nulă poate fi respinsă în favoarea ipotezei alternative sau nu.
- ◊ În general, un test preia apriori o anumită ipoteză, care trebuie verificată (de exemplu, ipoteza de independenţă a observaţiilor, ipoteza de normalitate ,ipoteza egalitatii unor medii etc.).

Test "*neparametric*" (Distribution-free test)

Testul în care funcţia de repartiţie a statisticii decizionale utilizate **nu depinde de funcţia de repartiţie a observaţiilor.** Sensul este cel dat de termenul englezesc. Denumirea *neparametric* a fost aleasă mai curând pentru uşurinţa exprimării. În română ar trebui să spunem *"test independent de repartiţia iniţială a …"*, dar fiind o formulare prea lungă s-a optat pentru neparametric.

TESTE PENTRU EGALITATEA MEDIILOR

Testul U

- Test utilizat pentru verificarea ipotezelor referitoare la mediile populaţiilor normale când se cunosc dispersiile teoretice.
- Testul U are forme diferite, în funcţie de ipotezele statistice formulate:
- Ex. :a) se verifică ipoteza $H_0: m = m_0$, testul U are expresia:

$$U = \frac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}}$$

Test U

◊ Test utilizat pentru verificarea ipotezelor referitoare la mediile populaţiilor normale când se cunosc dispersiile teoretice.

b) Când se verifică ipoteza egalităţii a două medii corespunzând la două populaţii normale care au aceeaşi dispersie teoretică σ^2, testul **U** are expresia:

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

c) Când se verifică ipoteza egalităţii a două medii, corespunzând la două populaţii normale care au dispersiile teoretice cunoscute, însă neegale, testul U are expresia:

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

În expresiile de mai sus, $\bar{x}_1$ şi $\bar{x}_2$ sunt mediile aritmetice de sondaj, iar n_1 şi n_2 sunt volumele eşantioanelor prelevate din populaţia originară.

Regiunea critică a testului U este $|U| > u(\alpha)$, în care $u(\alpha)$ se citeşte din tabelul repartiţiei normale normate $\mathbf{N}(0,1)$, astfel încât:

$$P[|U| > u(\alpha)] = 1 - \frac{1}{\sqrt{2\pi}} \int_{-u(\alpha)}^{u(\alpha)} e^{-\frac{z^2}{2}} dz = \alpha$$

Testul HI - patrat

- Regiunea critică a testului χ^2 pentru verificarea ipotezei $p_1 = p_2 = \ldots = p_m$ se construieşte pe baza indicatorului statistic de forma:

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i}$$

- care pentru $n \to \infty$ are repartiţia χ^2 cu k – 1 grade de libertate

Test F – Snedecor

- Testul statistic în care, pentru validarea ipotezei nule, statistica utilizată presupune existenţa repartiţiei F.
- Testul este utilizat pentru verificarea ipotezei egalităţii dispersiilor de sondaj obţinute în două eşantioane independente

Statistica testului F este definit prin relaţia:

$$\mathbf{F} = \frac{\mathbf{s}_1^2}{\mathbf{s}_2^2}$$

- astfel încât $s_1^2 \geq s_2^2$ în care s_1^2 şi s_2^2 sunt dispersiile de sondaj ale celor două eşantioane.

Repartiție t – (t – Student's distribution)

Repartiția de probabilitate a unei variabile aleatoare continue, care are funcția de densitate de probabilitate exprimată prin:

$$f(t;\nu)=\frac{1}{\sqrt{\pi\nu}}\left(\frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\Gamma\left(\frac{\nu}{2}\right)}\right)\left(\frac{1}{\left[1+\frac{t^2}{\nu}\right]^{\frac{\nu+1}{2}}}\right)$$

unde $-\infty < t < +\infty$ de parametru $\nu = 1,2,...$ iar Γ este funcția gamma.

Observație

Raportul dintre două variabile aleatoare independente, numărătorul fiind o variabilă normală normată și numitorul fiind rădăcina pătrată pozitivă a raportului dintre o variabilă aleatoare χ^2 și numărul său de grade de libertate ν, este o repartiție Student cu ν grade de libertate.Dacă variabila aleatoare X are o repartiție Student cu $\nu > 2$ grade de libertate, atunci:

$$M(X) = 0;\ \mathrm{Var}(X) = \frac{\nu}{\nu - 2}$$

■ Testul t – STUDENT

- ◊ Testul statistic în care, pentru validarea ipotezei nule, statistica utilizată presupune existența repartiției t (Student).
- ◊ Testul este aplicat, de exemplu, la următoarele probleme:

a. când **se verifică ipoteza** H_0: $m = m_0$, indicatorul t are expresia:

$$t=\frac{\bar{x}_1 - m_0}{\frac{s}{\sqrt{n}}}$$

cu $\nu = n-1$ grade de libertate, n fiind volumul eșantionului.

b. când **se verifică ipoteza egalității a două medii** corespunzând la două populații normale care au aceeași dispersie teoretică (necunoscută), indicatorul t are expresia:

$$t=\frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

cu $\nu = n_1 + n_2 - 2$ grade de libertate.

c. când se verifică **ipoteza egalității a două medii de sondaj** corespunzând la două populații normale care au dispersiile teoretice neegale, indicatorul t are expresia:

$$t=\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

această expresie are un număr de ν grade de libertate, rezultat din formula:

$$\nu=\frac{1}{\frac{c^2}{n_1-1}+\frac{(1-c^2)}{n_2-1}};\ c=\frac{\frac{s_1^2}{\nu_1}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Regiunea critică a testului t, în toate cazurile menționate, este:

$$|t| > t_\nu(\alpha)$$

◊ Pentru $\nu > 30$, testul t poate fi înlocuit cu testul U.

Egalitatea a două proporții

$$Z_c = \frac{|p_1 - p_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

CANDIDAT „A"
CURS : $p_1 = 52\%$; $n_1 = 1750$
CSOP : $p_2 = 45\%$; $n_2 = 1500$

$$Z_c = \frac{|0{,}45 - 0{,}52|}{\sqrt{\frac{0{,}25}{1500} + \frac{0{,}25}{1750}}} = 3{,}89$$

$Z_c > Z_{tab.} \Rightarrow$ Diferența
• EFECTIV VOTUL A COST 36,8%. !