EŞANTIOANE DEPENDENTE ŞI INDEPENDENTE

- De interes când se fac ipoteze asupra a două sau mai multe populații.
- O *sursă* este o persoană, un obiect etc. care produce o dată elementară.
- *Eşantionarea dependentă* se face atunci când se foloseşte aceeaşi mulțime de surse pentru ambele (toate) populații(le), selectarea unui element într-un eşantion impunând selectarea unui *anumit* element în al doilea (probabilitățile de selecție sunt dependente – v. exemplul cursului de franceză, la testul t pentru perechi).
- *Eşantionarea independentă* – când se folosesc mulțimi de surse fără legătură între ele (testarea cauciucurilor pe maşini diferite, nu pe aceleaşi maşini).

INFERENȚE ASUPRA A DOUĂ POPULAȚII

1. Inferențe asupra diferenței dintre două medii independente (dispersii cunoscute sau eşantioane mari): distribuția normală.
2. Inferențe asupra a două dispersii: distribuția F.
3. Inferențe asupra diferenței dintre două medii independente (dispersii necunoscute şi eşantioane mici): distribuția Student. Cazuri: dispersii egale; dispersii diferite.
4. Inferențe asupra diferenței dintre două medii dependente (controlul factorilor netestați): distribuția Student.
5. Inferențe asupra proporțiilor (distribuția normală).

DIFERENȚA DINTRE DOUĂ MEDII INDEPENDENTE

- Inferențe asupra diferenței parametrilor $\mu_1 - \mu_2$ se fac pe baza diferenței statisticilor, $\overline{x}_1 - \overline{x}_2$.
- Dacă se extrag eşantioane independente de dimensiuni n_1 şi n_2 din populații mari de medii necunoscute μ_1 şi μ_2 şi dispersii cunoscute σ_1^2 , respectiv σ_2^2, atunci distribuția de selecție a variabilei X= $\overline{x}_1 - \overline{x}_2$
  - este aproximativ normală;
  - are media $\mu = \mu_1 - \mu_2$ şi dispersia $\sigma^2 = \sigma_1^2 / n_1 + \sigma_2^2 / n_2$
- Se foloseşte statistica $z = \dfrac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$

EXEMPLUL I

- Se extrag eşantioane de câte 40 de indivizi din două populații diferite. Se obțin mediile de eşantion x_{med_1}=2,03 şi x_{med_2}=2,21. Se presupun cunoscute deviațiile standard $\sigma_1=\sigma_2= 0,6$. La nivel $\alpha=0,05$ se testează ipotezele:
- $H_0 : \mu_1 = \mu_2$ (>); $H_a : \mu_1 < \mu_2$ (sau $\mu_1 - \mu_2 < 0$).
- $z_{critic} = - z(0,05) = -1,645$.
- $z_{esantion} = -0,18/0,134 = - 1,343$
- Decizie: Nu se respinge H_0.
- Se pot construi intervale de încredere pentru $\mu_1 - \mu_2$
- În exemplul de mai sus, acesta este (-0,44; 0,08).

DISPERSII NECUNOSCUTE ŞI EŞANTIOANE MARI

- Când n_1 , $n_2 > 30$, chiar estimând σ_i prin s_i se poate aplica acelaşi test (cu aproximație).
- Exemplu. Pe un eşantion de 50 indivizi dintr-o populație se obține o medie de 57,5 şi o deviație standard de 6,2, iar pe un eşantion de 60 de indivizi dintr-o altă populație, aceeaşi caracteristică măsurată dă media 54,4 şi deviația standard 10,6. Să se dea un interval de încredere (0,05) pentru diferența mediilor celor două populații.

$$(\overline{x}_1 - \overline{x}_2) \pm z_{critic}(0{,}025) \cdot \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2} = 3{,}1 \pm 4{,}19$$

- Intervalul este (-1,09; +7,29). La nivel 0,05, se respinge, de exemplu, ipoteza "$\mu_1 - \mu_2 = 10$".

INFERENȚE ASUPRA A DOUĂ DISPERSII

1. Egalitatea a două dispersii
2. Estimarea raportului σ_1^2 / σ_2^2 a două dispersii.

- Două eşantioane independente, de n_1, respectiv n_2 indivizi (din cele două populații normale).
- Statistica este $F = s_1^2 / s_2^2$.
- În condițiile de mai sus, statistica are distribuție **F**.
  - Nenegativă; asimetrică;
  - Câte o distribuție F pentru fiecare pereche de grade de libertate;
  - Valori critice $F(df_n, df_d, \alpha)$;
  - $F(df_1, df_2, 1-\alpha) = 1 / F(df_2, df_1, \alpha)$.

EXEMPLUL I

- Maşina existentă *e*: 22 teste, $s_e^2 = 0{,}0008$;
- Maşina rapidă *r*: 25 teste, $s_r^2 = 0{,}0018$.
- Se poate respinge ($\alpha = 0{,}01$) ipoteza companiei că maşina mai rapidă nu are dispersie mai mare?
- **$H_0 : \sigma_1^2 = \sigma_2^2$ (sau $\sigma_1^2 / \sigma_2^2 = 1$);**
- **$H_a : \sigma_1^2 > \sigma_2^2$ (sau $\sigma_1^2 / \sigma_2^2 > 1$).**
- **$F_{critic} = F(24; 21; 0{,}01) = 2{,}80$.**
- **$F_{eşantion} = s_1^2 / s_2^2 = 0{,}0018 / 0{,}0008 = 2{,}25$.**
- **Nu se poate respinge H_0.**
- **Interval de încredere pentru σ_1^2 / σ_2^2 :**

$\left((s_A^2 / s_B^2) / F(df_A; df_B; \alpha/2) ;\right.$

$\left. (s_A^2 / s_B^2) / F(df_A; df_B; 1-\alpha/2) \right)$

INFERENȚE ASUPRA DIFERENȚEI DINTRE DOUĂ MEDII INDEPENDENTE

(în cazul dispersiilor necunoscute şi al eşantioanelor mici)

- Populații aproximativ normale.
- Statistică de distribuție t.
- Cazuri: 1.- $\sigma_1^2 = \sigma_2^2$; 2.- $\sigma_1^2 \neq \sigma_2^2$.

1.- Statistica:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}}, \quad \text{unde}$$

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

- s_p este estimarea deviației standard din eşantioanele "reunite".

EXEMPLUL II (1)

- Studiindu-se necesitățile financiare ale studenților, s-a ridicat întrebarea dacă fetele şi băieții cheltuiesc la fel de mult pentru rechizite / cărți.
- Pentru a se afla răspunsul, s-au luat două eşantioane de câte 25 de persoane. Pe baza datelor, se poate respinge ($\alpha = 0,10$) ipoteza nulă că fetele şi băieții cheltuiesc la fel de mult la acest capitol?
- Fete: medie 10,55 (sute mii lei); $s^2 = 24,47$;
- Băieți: medie 10,22 (sute mii lei); $s^2 = 33,95$.
- <u>Soluție</u>. Cum dispersiile sunt necunoscute, trebuie mai întâi testat dacă ele sunt egale sau nu, apoi aplicat cazul corespunzător pentru medii.

EXEMPLUL II (2)

- **Prima ipoteză**:
- $H_0 : \sigma_b^2 = \sigma_f^2$; $H_a : \sigma_1^2 \neq \sigma_2^2$.
- $F_{critic\_dr} = F(24; 24; 0{,}05) = 1{,}98$;
- $F_{critic\_st} = 1 / F(24; 24; 0{,}95) = 1 / 1{,}98 = 0{,}505$.
- $F_{eşantion} = s_b^2 / s_f^2 = 33{,}95 / 24{,}47 = 1{,}387$.
- Nu se poate respinge H_0. Deci suntem în cazul 1.
- **A doua ipoteză**:
- $H_0 : \mu_b = \mu_f$; $H_a : \mu_b \neq \mu_f$.
- $t_{critic} = t(48; 0{,}05) = 1{,}65$.
- $t_{eşantion} = -0{,}2158$. Nu se poate respinge H_0.

CAZUL II: DISPERSII INEGALE

- Dacă dispersiile sunt inegale, atunci nu se mai pot unifica eşantioanele, astfel că deviația standard a diferenței mediilor de selecție se modifică, distribuția statisticii rămânând aceeaşi - Student:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

- Numărul de grade de libertate este $\min(n_1-1, n_2-1)$.
- Temă. Studenții se plâng că automatul de cafea din corpul A toarnă mai puțin lichid decât cel din corpul B. La 10 cafele A rezultă o medie de 5,38 cu deviația standard observată 1,59; la 12 cafele B, media este 5,92 şi deviația standard 0,83. Se susține (α = 0,05) plângerea?

INFERENȚE ASUPRA DIFERENȚEI DINTRE DOUĂ MEDII DEPENDENTE

- Observațiile se grupează în perechi, pentru care se calculează diferențele. Populația de diferențe se presupune aproximativ normală cu media presupusă μ_d şi dispersia necunoscută σ^2 (estimată prin s_d).
- Din eşantion se calculează $\overline{d}$, media diferențelor din eşantion, care are deviația standard s_d.
- Statistica este t cu n-1 grade de libertate:

$$t = \frac{\overline{d} - \mu_d}{s_d / \sqrt{n}}$$

- Se urmează paşii de la ipoteze t.

ANALIZA DISPERSIONALĂ - ANOVA

- Testare de ipoteze asupra mai multor medii.
- Exemplu. $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.
- Cu tehnicile deja cunoscute, ar însemna testarea a 10 ipoteze asupra a două medii fiecare, cu α mult mai mare pentru întreg testul decât pentru fiecare sub-test în parte.
- ANOVA: un singur test, cu α prescris.
- Cea mai simplă variantă ANOVA: cea cu un singur factor.
- Exemplu. Într-o fabrică, temperatura pare a influența producția. Se numără piesele realizate într-o oră la trei temperaturi t_1, t_2, t_3: de 4 ori la t_1, de 5 ori la t_2, de 4 ori la t_3.

EXEMPLUL III (1)

- t_1 : 10, 12, 10, 9(total $C_1 = 41$; medie 10,25; $k_1=4$);
- t_2 : 7,6,7,8,7 (total $C_2 = 35$; medie 7,0; $k_2=5$);
- t_3 : 3,3,5,4 (total $C_3 = 15$; medie 3,75; $k_3=4$).
- $n = k_1 + k_2 + k_3 = 13$.
- $H_0 : \mu_1 = \mu_2 = \mu_3$.
- H_a : cel puțin o medie diferă de celelalte.
- Statistica şi distribuția F (raport de dispersii).
- Se partiționează suma pătratelor abaterilor în partea de sumă datorată factorului studiat şi partea de sumă datorată erorilor (de eşantionare):
- SPA(total) = SPA(factor) + SPA(eroare)

EXEMPLUL III (2)

- $SPA(factor)=(C_1^2 /k_1+C_2^2 /k_2+C_3^2 /k_3+\ldots)-(\Sigma x)^2 /n.$
- $SPA(temp)=41^2 /4 + 35^2 /5 + 15^2 /4 - 91^2/13 = 84{,}5.$
- $SPA(eroare)=\Sigma(x^2) - (C_1^2 /k_1+C_2^2 /k_2+C_3^2 /k_3+\ldots)$
- SPA(er_exp) = 731-721,5 = 9,5.

Sursa	SPA	df	MS=SPA/df
Factor	84,5	2 = 3-1	42,25
Eroare	9,5	10 = 13-3	0,95
Total	94	12 = 13-1	-

EXEMPLUL III (3)

- Statistica este $F_{esantioane}$ = MS(factor) / MS (eroare)
- În exemplu:
- $F_{esantioane}$ = MS(temperatură) / MS(er_exp) = =42,25/0,95 = 44,47.
- F_{critic} = F(2; 10; 0,05) = 4,10.
- Se respinge ipoteza H_0.
- Intuitiv: Se compară MS(factor) – variația între niveluri – cu MS(eroare) – variația în interiorul nivelurilor. Dacă MS(factor) este în mod semnificativ mai mare decât MS(eroare), atunci se decide că mediile nu sunt egale.