Statistică - Cursul 12

Olariu E. Florentin

Mai, 2016

Cuprins

Inferenţe asupra a două populaţii

- Bătălia sexelor poate lua diverse forme. Atunci când subiectul este cine e mai bun, mai rapid, mai sigur, lupta poate deveni acerbă!
- După ce lucrurile s-au liniştit cineva întreabă "Cine parcurge drumul cel mai lung până la universitate?". Lungimea drumului poate fi măsurată ca timp sau ca distanţă; dar există mulţi alţi factori care influenţează drumul către universitate. Trăiesc acasă (sau în chirie)? Au o slujbă part-time sau una întreagă? Au obligaţii de familie?
- În acest curs vom lua în discuţie procedurile prin care se fac inferenţe asupra a două populaţii.
- Atunci când se compară două populaţii avem nevoie de două eşantioane, câte unul din fiecare populaţie.

Inferenţe asupra a două populaţii

- Pot fi utilizate două tipuri de eşantioane: independente sau dependente. Dependenţa sau independenţa a două eşantioane is determinată de sursele datelor.
- O sursă poate fi o persoană, un obiect sau orice altceva care poate dă naştere la date. Dacă acelaşi set de surse sau surse înrudite sunt folosite pentru a obţine datele pentru amândouă populaţiile, atunci avem de-a face cu eşantioane dependente.
- Dacă sunt folosite surse care nu au legătură între ele vom obţine eşantioane independente. Un exemplu poate clarifica aceste idei.
- În cursul nostru vom utiliza populaţii/eşantioane independente.

Inferenţe asupra a două populaţii

- Se concepe un test pentru compararea a două noi tipuri de anvelope auto.
- Automobilele selectate vor fi echipate cu cele două tipuri de anvelope şi apoi conduse în condiţii "normale" timp de o lună. Apoi se vor face măsurători pentru a determina comportamentul lor. Au fost propuse două planuri:
  - Plan I: Un eşantion format din maşini alease aleator vor fi echipate cu tipul A de anvelope, apoi vor fi conduse timp de o lună. Un alt eşantion aleator de maşini va fi selectat şi echipat cu anvelopele B, apoi vor fi conduse pentru o lună întreagă.
  - Plan II: Fiecare maşină dintr-un eşantion aleator va avea o anvelopă de tip A şi o anvelopă de tip B (celelalte două anvelope nu participă la test) şi apoi vor fi conduse pentru o lună.
- Planul I oferă date independente (sursele sunt neînrudite), iar planul II dă eşantioane dependente (sursele sunt comune).

Testul Z - Inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 cunoscute)

- Considerăm două populaţii statistice ale căror dispersii (σ_1^2 şi σ_2^2) sunt cunoscute.
- Populaţiile sunt normal distribuite şi dorim să testăm o ipoteză asupra diferenţei mediilor.
- Dorim să ştim, de exemplu, dacă una dintre medii este mai mică decât cealaltă sau dacă sunt diferite.
- Alegem două eşantioane aleatoare simple independente cu mediile de selecţie $\overline{x}_{n_1}$ şi $\overline{x}_{n_2}$. Considerăm (în condiţiile ipotezei nule) că mediile celor două populaţii sunt μ_1 şi μ_2, respectiv.
- Următoare statistică este distribuită normal standard

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Testul Z - Inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 cunoscute)

- Putem utiliza testul Z chiar dacă cele două populaţii sunt doar aproximativ normal distribuite, dacă cele două eşantioane sunt suficient de mari. eşantioane are large ($n_1, n_2 \geqslant 30$).
- Testul se desfăşoară astfel:

1. Formulăm *ipoteza nulă*, care susţine că diferenţa mediilor celor două populaţii ia o anumită valoare:

$$\boxed{H_0 : \ \mu_1 - \mu_2 = m_0}$$

2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantioane.

Testul Z - Inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 cunoscute)

Putem avea trei tipuri de ipoteză alternativă

$\boxed{H_a : \mu_1 - \mu_2 < m_0}$ (*asimetrică la stânga*) sau

$\boxed{H_a : \mu_1 - \mu_2 > m_0}$ (*asimetrică la dreapta*) sau

$\boxed{H_a : \mu_1 - \mu_2 \neq m_0}$ (*simetrică*).

Ipotezele asimetrice se mai numesc *one-tailed*, iar cea simetrică *two-tailed*.

3. Alegem un nivel de semnificaţie $\alpha \in \{1\%, 5\%\}$.

Testul Z - Inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 cunoscute)

4. Calculăm *scorul z* (*statistica* testului)

$$\boxed{z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}}$$

5. Determinăm valoarea critică corespunzătoare lui α

$\boxed{z^* = qnorm(\alpha)}$ pentru H_a asimetrică la stânga ($z^* < 0$),

$\boxed{z^* = qnorm(1-\alpha)}$ pentru H_a asimetrică la dreapta ($z^* > 0$),

$\boxed{z^* = -qnorm(\alpha/2) = qnorm(1-\alpha/2)}$ pentru H_a simetrică ($z^* > 0$

Testul Z - Inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 cunoscute)

6. Comparăm valoarea critică cu scorul z; dacă scorul z aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:

$$(-\infty, z^*] \text{ pentru } H_a \text{ asimetrică la stânga,}$$

$$[z^*, +\infty) \text{ pentru } H_a \text{ asimetrică la dreapta,}$$

$$(-\infty, -|z^*|] \cup [|z^*|, +\infty) \text{ pentru } H_a \text{ simetrică.}$$

Dacă scorul z nu aparţine zonei de respingere spunem că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).

Testul Z - Inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 cunoscute)

Observaţii

- Atunci când întreprindem un test asupra mediilor a două populaţii, cel mai adesea vom presupune, în ipoteza nulă, că cele două medii sunt egale (i. e., $\mu_1 - \mu_2 = 0$, sau $m_0 = 0$).
- Sintagma "să se testeze mediile" are următorul înţeles: testăm $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 \neq \mu_2$.
- O altă observaţie este aceea că paşii 3, 5 şi 6 ai testului curent sunt identici cu cei ai testului Z asupra mediei unei populaţii (vezi cursul 11).

Testul Z - Inferenţe asupra mediilor a două populaţii - Exemplu

Examplu.

- Un studiu arată că studenţii din primul an lucrează în fiecare săptămână având o deviaţie standard de 8.7 ore (cei din universităţile publice) şi cu o deviaţie standard de 8.9 (cei din universităţile private). Folosind eşantioane aleatoare independente (unul din universităţile publice de dimensiune 900 şi unul din universităţile private de dimensiune 1000) se determină mediile de selecţie 16.1 şi 15.2 de ore pe săptămână, respectiv.
- Această diferenţă între medii este datorată şansei? Dacă nu, ce altă explicaţie ar putea avea? Folosiţi 5% şi 1% nivel de semnificaţie.

Soluţie.

- Este posibil ca cele două medii ale populaţiilor să fie diferite.
- Nu ştim dacă cele două populaţii urmează distribuţii normale, dar eşantioanele sunt suficient de mari pentru a folosi un test Z.

Testul Z - Inferenţe asupra mediilor a două populaţii - Exemplu

- Datele privindcele două populaţii/eşantioane sunt: $m_0 = 0$, $\sigma_1 = 8.7$, $n_1 = 1000$, $\overline{x}_{n_1} = 16.1$, $\sigma_2 = 8.9$, $n_2 = 900$, and $\overline{x}_{n_2} = 15.2$.

1-2. Formulăm ipoteza nulă şi o ipoteză alternativă simetrică

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0.$$

3. $\alpha = 0.05$.
4. Scorul z

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 2.2244.$$

5. Valoarea critică este $z^* = -qnorm(\alpha/2) = 1.9599$, pentru $\alpha = 5\%$.
6. Deoarece $|z| > |z^*|$, putem respinge ipoteza nulă, şi să acceptăm că mediile celor două populaţii diferă: the diferenţa dintre mediile eşantioanelor nu se datorează hazardului.

Testul Z - Inferenţe asupra mediilor a două populaţii - Exemplu

- Pentru celălalt nivel de semnificaţie:

5'. Cu $\alpha = 1\%$ valoarea critică este $z^* = -qnorm(\alpha/2) = 2.5758$.

6'. Cum $|z| < |z^*|$, încercarea de a respinge ipoteza nulă eşuează cu 1% nivel de semnificaţie (nu există suficiente dovezi pentru nivelul de 1% pentru a susţine că mediile celor două populaţii diferă).

Testul Z - Inferenţe asupra mediilor a două populaţii - Exemplu revăzut

- Privind din nou informaţiile din eşantion putem observa că media de selecţie a studenţilor din universităţile publice este mai mare decât aceea a celor din universităţile private.
- Am putea formula o ipoteză alternativă asimetrică la dreapta

1-2. Noile ipoteze sunt

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0.$$

3. $\alpha = 0.05.$
4. Scorul z

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 2.2244.$$

Testul Z - Inferenţe asupra mediilor a două populaţii - Exemplu revăzut

5. Valoarea critică: $z^* = qnorm(1-\alpha) = 1.6448$, pentru $\alpha = 5\%$.
6. Deoarece $z > z^*$, putem respinge ipoteza nulă, şi să acceptăm că media primei populaţii este mai mare decât a celei de-a doua: în medie un student la o instituţie publică munceşte mai mult decât unul de la o instituţie privată.

- Pentru 1% nivel de semnificaţie:

5'. Valoarea critică este $z^* = qnorm(1-\alpha) = 2.3263$.

6'. Cum $z < z^*$, nu putem respinge ipotez nulă cu 1% nivel de semnificaţie (nu există dovezi suficiente, cu 1% nivel de semnificaţie, pentru a susţine că prima medie este mai mare decât cea de-a doua).

Testul Z - Exerciţii

I. Întreprindeţi din nou testul Z anterior modificând $n_1 = 1000$.

II. Se ştie că un colegiu privat are costuri mai mari decât unul public. Este această diferenţă valabilăşi pentru costurile manualelor universitare? Aceste costuri per curs urmează legi normale cu deviaţia standard de 18.55$ (pentru colegiile publice) and 13.12$ (pentru colegiile private). Se aleg două eşantioane aleatoare de dimensiune. Folosind $\alpha = 5\%$ să se determine dacă costurile medii per curs ale manualelor şcolare diferă în cele două tipuri de colegii

$$Public:$$

64.69 89.6 101.49 101.75 103.59 106.38 106.77 110.69 118.94 135.94

$$Privat:$$

71.00 96.19 97.14 96.47 98.56 98.94 107.79 112.58 114.00 116.55

Inferenţe asupra raportului dintre dispersii

- Când se compară două populaţii, în mod natural vom pune fa ctă în faţă parametrii fundamentali ai distribuţiei, "centrele" şi "împrăştierea", adică mediile şi deviaţiile standard.
- Am descris deja o procedură pentru compararea mediilor a două populaţii folosond eşantioane independente, când se ştiu dispersiile.
- Există şi o altă procedură pentru cazul când dispersiile nu se cunosc. Dar pentru acest caz trebuie mai întâi ştiut dacă dispersiile sunt egale.
- Următorul pas (logic) va fi compararea deviaţiilor standard.

Inferenţe asupra raportului dintre dispersii

- Procedura de inferenţă prezentată aici va fi un test de semnificaţie pentru deviaţiile standard (sau dispersiile) a două populaţii normale.
- Pentru acest test prezumţia de normalitate este foarte importantă.
- Alegem două eşantioane aleatoare independente (de dimensiuni n_1 şi n_2, respectiv) cu deviaţiile standard s_1 şi s_2.
- Considerăm că, în condiţile ipotezei nule, adevăratele deviaţii standard σ_1 şi σ_2 sunt egale.
- Următoarea statistică este distribuită Fisher

$$F = \frac{s_1^2}{s_2^2}.$$

Distribuţia Fisher, $F(r_1, r_2)$

- Distribuţia Fisher, face parte dintr-o familie de distribuţii. Fiecare distribuţie F se identifică prin două numere de grade de libertate (cîte unul pentru fiecare eşantion).
- Proprietăţi ale distribuţiei F:
  - F are valori nenegativă;
  - F este asimetrică;
- Pentru inferenţele din această secţiune numerele de grade de libertate sunt $r_1 = n_1 - 1$, $r_2 = n_2 - 1$.

Testul F - Inferenţe asupra raportului dintre dispersii

- Testul se desfăşoară astfel:

1. Formulăm *ipoteza nulă*:

$$\boxed{H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1}$$

2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantion. Putem avea două[1] tipuri de ipoteză alternativă

$$\boxed{H_a : \frac{\sigma_1}{\sigma_2} > 1}$$ (*asimetrică la dreapta*) pentru un *un test one-taile*

$$\boxed{H_a : \frac{\sigma_1}{\sigma_2} \neq 1}$$ (*simetrică*) pentru un *un test two-tailed*.

[1] Este recomandat ca dispersia mai mare sau care e de aşteptat sa fie mai mare să fie pusă la numărător.

Testul F - Inferenţe asupra raportului dintre dispersii

3. Alegem nivelul de semnificaţie $\alpha \in \{1\%, 5\%\}$.
4. Calculăm *scorul F* (*statistica* testului)

$$\boxed{F = \frac{s_1^2}{s_2^2}}$$

5. Determinăm valoarea critică corespunzătoare pentru α

$$\boxed{F^* = qf(\alpha, n_1 - 1, n_2 - 1)}$$ pentru H_a asimetrică la dreapta ,

$$\boxed{F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1), F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1)}$$

pentru H_a simetrică.

Testul F - Inferenţe asupra raportului dintre dispersii

6. Comparăm valoarea critică cu scorul F; dacă scorul F aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:

$$[F^*, +\infty) \text{ pentru } H_a \text{ asimetrică la stânga,}$$

$$(0, F_s^*] \cup [F_d^*, +\infty) \text{ pentru } H_a \text{ simetrică.}$$

Dacă scorul F nu aparţine zonei de respingere spunem că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).

Testul F - Inferenţă asupra raportului dintre dispersii - Exemplu

Exemplu.

- O companie care îmbuteliază băuturi răcoritoare deţine o maşină care umple sticle de 16 uncii. Compania trebuie să controleze deviaţia standard σ (sau dispersia σ^2) cantităţii de băutură din sticle. O medie corectă a cantităţii de bătură nu garantează faptul ca maşina este bine calibrată. Dacă dispersia este prea mare multe sticle vor avea prea multă sau prea puţină băutură.
- Astfel, maşina trebuie să aibă o deviaţie standard (sau dispersie) cât mai mică posibil.
- Compania doreşte să decidă dacă să instaleze o maşină de îmbuteliat mai modernă şi mai rapidă.
- Una dintre îngrijorări este aceea că o viteză mai mare de îmbuteliere corespunde unei dispersii mai mari, iar o asemenea creştere nu este acceptabilă.

Testul F - Inferenţe asupra raportului dintre dispersii - Exemplu

- Producătorul noii maşini de îmbuteliat susţine că dispersia nu este mai mare decât cea a vechii maşini.
- Un eşantion de 25 de sticle îmbuteliate de maşina nouă are o dispersie de 0.0018, în timp ce un eşantion de 22 sticle îmbuteliate maşină curentă are o dispersie de 0.0008.

Soluţie.

- Vom testa amândouă tipurile de ipoteză alternativă: mai întâi faptul că dispersiile celor două maşini sunt diferite, iar apoi că noua maşină are o dispersie mai mare.
- Datele relativ la cele două populaţii/eşantioane: $n_1 = 25$, $s_1^2 = 0.0018$, $n_2 = 22$ şi $s_2^2 = 0.0008$.

Testul F - Inferenţe asupra raportului dintre dispersii - Exemplu

1-2. Formulăm ipoteza nulă şi o ipoteză alternativă simetrică

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$

3. Alegem $\alpha = 5\%$.
4. Calculăm scorul F al testului

$$F = \frac{s_1^2}{s_2^2} = 2.2500$$

5. Valorile critice sunt $F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1) = 0.4327, F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1) = 2.3675$, pentru $\alpha = 5\%$.

Testul F - Inferenţe asupra raportului dintre dispersii - Exemplu

6. Deoarece $F \in [F_s^*, F_d^*]$, încercarea de a respinge ipoteza nulă eşuează cu 5% nivel de semnificaţie (nu există suficiente dovezi pentru nivelul de 5% pentru a susţine că dispersiile celor două maşini sunt diferite): diferenţele sunt datorate hazardului.

- Reluăm ultimii doi paşi pentru celălalt nivel de semnificaţie:

5'. Pentru $\alpha = 1\%$ valorile critice vor da un interval mai mare: $F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1) = 0.3294$, $F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1) = 3.1473$.

6'. Concluzia este aceeaşi: nu putem respinge ipoteza nulă.

- La fel ca şi în cazul testelor Z şi T, dacă H_0 nu se respinge pentru $\alpha = 5\%$ nu se va respinge nici pentru $\alpha = 1\%$ (pentru că zona de respingere în al doilea caz o înglobează pe prima).

Testul F - Inferenţe asupra raportului dintre dispersii - Exemplu revizuit

- Deoarece dispersia primului eşantion este mai mare decât cea de-a doua vom întreprinde un al doilea test cu o ipoteză alternativă asimetrică la dreapta.

1-2. Noile ipoteze sunt

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a : \frac{\sigma_1^2}{\sigma_2^2} > 1.$$

3. Alegem $\alpha = 5\%$.
4. Scorul testului va fi acelaşi

$$F = \frac{s_1^2}{s_2^2} = 2.2500$$

5. Valoarea critică este $F^* = qf(1-\alpha, n_1 - 1, n_2 - 1) = 2.0540$, pentru $\alpha = 5\%$.

Testul F - Inferenţe asupra raportului dintre dispersii - Exemplu revizuit

6. Deoarece $F \in [F_d^*, +\infty)$, putem să respingem ipoteza nulă şi să acceptăm ipoteza alternativă: dispersia noii maşini este mai mare decât a celei vechi.

- Pentru $\alpha = 1\%$

5'. Valoarea critică este $F^* = qf(1-\alpha, n_1 - 1, n_2 - 1) = 2.8010$.

6'. Cum $F < F^*$, încercarea de a respinge ipoteza nulă eşuează cu 1% nivel de semnificaţie (nu există suficiente dovezi pentru nivelul de 1% pentru a susţine că maşina nouă are o dispersie mai mare).

Testul T - inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 necunoscute)

- Considerăm două populaţii statistice ale căror dispersii (σ_1^2 şi σ_2^2) sunt necunoscute.
- Populaţiile sunt normal distribuite şi dorim să testăm o ipoteză asupra diferenţei mediilor.
- Dorim să ştim, de exemplu, dacă una dintre medii este mai mare decât cealaltă sau dacă sunt diferite.
- Alegem două eşantioane aleatoare simple independente cu mediile de selecţie $\overline{x}_{n_1}$ şi $\overline{x}_{n_2}$ şi deviaţiile standard s_1 şi s_2, respectiv.
- Considerăm (în condiţiile ipotezei nule) că mediile celor două populaţii sunt μ_1 şi μ_2, respectiv.
- Expresia scorului testului depinde de faptul dacă cele două dispersii sunt egale sau diferite. Astfel, **înainte de întreprinde testul T vom desfăşura un test F relativ la raportul celor două dispersii**.

Testul T - inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 necunoscute)

- Atunci când cele două dispersii sunt egale următoarea statistică este repartizată Student cu $(n_1 + n_2 - 2)$ grade de libertate

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}},$$

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

- Dacă cele două dispersii sunt diferite următoarea statistică este repartizată Student cu $\min(n_1 - 1, n_2 - 1)$ grade de libertate

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Testul T - inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 necunoscute)

Testul decurge astfel:

1. Formulăm *ipoteza nulă*, care susţine că diferenţa mediilor celor două populaţii ia o anumită valoare:

$$\boxed{H_0 : \ \mu_1 - \mu_2 = m_0}$$

2. Formulăm *ipoteza alternativă* conform datelor din eşantioane. Putem avea trei tipuri de ipoteză alternativă

$$\boxed{H_a : \ \mu_1 - \mu_2 < m_0} \quad (\textit{asimetrică la stânga}) \text{ sau}$$

$$\boxed{H_a : \ \mu_1 - \mu_2 > m_0} \quad (\textit{asimetrică la dreapta}) \text{ sau}$$

$$\boxed{H_a : \ \mu_1 - \mu_2 \neq m_0} \quad (\textit{simetrică}).$$

Ipotezele asimetrice se mai numesc *one-tailed*, iar cea simetrică *two-tailed*.

3. Alegem nivelul de semnificaţie $\alpha \in \{1\%, 5\%\}$.

Testul T - inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 necunoscute)

4. Calculăm *scorul t* (*statistica* testului)

a) dacă dispersiile sunt egale:

$$\boxed{t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}}$$

unde $s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, iar numărul de grade de libertate este $df = n_1 + n_2 - 2$.

b) dacă dispersiile sunt diferite:

$$\boxed{t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

numărul de grade de libertate fiind $df = \min(n_1 - 1, n_2 - 1)$.

Testul T - inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 necunoscute)

5. Determinăm valoarea critică corespunzătoare

   $\boxed{t^* = qt(\alpha, n-1)}\, mbox pentru H_a$ asimetrică la stânga ($t^* < 0$),

   $\boxed{t^* = qt(1-\alpha, n-1)}$ pentru H_a asimetrică la dreapta ($t^* > 0$),

   $\boxed{t^* = -qt(\alpha/2, n-1) = qt(1-\alpha/2, n-1)}$ pentru H_a simetrică (t^*

6. Comparăm valoarea critică cu scorul z; dacă scorul z aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:

$$(-\infty, t^*] \text{ pentru } H_a \text{ asimetrică la stânga,}$$

$$[t^*, +\infty) \text{ pentru } H_a \text{ asimetrică la dreapta,}$$

$$(-\infty, -|t^*|] \cup [|t^*|, +\infty) \text{ pentru } H_a \text{ simetrică.}$$

Testul T - inferenţe asupra mediilor a două populaţii - Exemplu

Dacă scorul t nu aparţine zonei de respingere spunem că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).

Exemplu.

- National Assesment of Educational Progress (NAEP) monitorizează tendinţele performanţei şcolare. În fiecare an NAEP desfăşoară la nivel naţional teste asupra mai multor subiecte pe eşantioane de elevi care au 17 ani. Se ştie că rezultatele acestor teste sunt normal distribuite.
- Testul de lectură a fost dat în 1990 şi din nou în 2004. Pentru două eşantioane de 1000 de elevi scorul mediu a scăzut de la 290 la 285, deviaţiile standard ale eşantioanelor fiind 40 şi 37, respectiv.
- Diferenţa este de 5 puncte; este aceasta datorată hazardului sau este o diferenţă reală? ($\alpha = 1\%$)

Testul T - inferenţe asupra mediilor a două populaţii - Exemplu

Soluţie.

- Deoarece populaţii sunt distribuite normal, iar deviaţiile standard ale populaţiilor nu sunt cunoscute vom întreprinde un test T pentru cele două medii.
- *Mai întâi întreprindem testul F asupra raportului celor două dispersii*.
- Informaţiile legate de eşantioane sunt $s_1 = 40$, $n_1 = 1000$, $\overline{x}_{n_1} = 290$, $s_2 = 37$, $n_2 = 1000$, $\overline{x}_{n_2} = 285$.

1-2. $H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1.$

Testul T - inferenţe asupra mediilor a două populaţii - Exemplu

3. Alegem $\alpha = 5\%$.
4. Calculăm scorul F al testului

$$F = \frac{s_1^2}{s_2^2} = 1.0810$$

5. Valorile critice sunt $F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1) = 0.8494$, $F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1) = 1.1771$, pentru $\alpha = 5\%$.
6. Deoarece $F \in [F_s^*, F_d^*]$, încercarea de a respinge ipoteza nulă eşuează cu 1% nivel de semnificaţie (nu există suficiente dovezi pentru nivelul de 1% pentru a susţine că cele două dispersii sunt diferite); diferenţa dintre dispersiile celor două eşantioane este datorată întâmplării.

- *În cele ce urmează vom presupune căcele două dispersii sunt egale.*

Testul T - inferenţe asupra mediilor a două populaţii - Exemplu

- Întreprindem acum testul T pentru diferenţa mediilor.

1-2. $H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0.$

3. $\alpha = 5\%,$

4. Calculăm scorul t al testului

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2})}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = 2.9003,$$

unde $s = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 38.5292.$

Testul T - inferenţe asupra mediilor a două populaţii - Exemplu

5. Valoarea critică: $t^* = -qt(\alpha/2, df) = 2.5782$, unde $df = n_1 + n_2 - 2 = 1998$ ($\alpha = 1\%$).
6. Deoarece $|t| > |t^*|$, putem respinge ipoteza nulă, şi să acceptăm că cele două medii sunt diferite: între 1990 şi 2004 rezultatul mediu al testului de lectură s-a schimbat.

- Întreprindem acum un test cu ipoteză alternativă asimetrică la dreapta:

1-2'. $H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0$.

5'. Valoarea critică este $t^* = qt(1 - \alpha, df) = 2.5782$, unde $df = n_1 + n_2 - 2 = 1998$.

6'. Cum $t > t^*$, putem respinge ipoteza nulă, şi să acceptăm că prima medie (din 1990) este mai mare decât cea de-a doua (din 2004): rezultatul mediu a scăzut.

- (Exerciţiu) A fost necesar acest al doilea test T?

Testul T - inferenţe asupra mediilor a două populaţii - Exerciţii

I. Femeile au în medie cu 8 perechi de pantofi mai mult decât bărbaţii (conform unui Snapshot USA Today). Un studiu recent efectuat la un colegiu a determinat următoarele valori

	n	media	dev. std.
Bărbaţi	21	8.48	4.43
Femei	30	26.63	21.83

Diferenţa dintre cele două eşantioane este mai mare decât 8, testaţi această ipoteză cu $\alpha = 5\%$. (Populaţiile sunt are normale.)

II. Se consideră de mult timp că chiria medie a unei maşini în New York este egală cu aceea din Boston. Un studiu din 2007 însă a dat următoarele valori:

	n	media	dev. std.
Boston	10	95.94	7.50
New York	16	127.75	15.83

Este această diferenţă semnficativă sau este datorată şansei? $(1\%, 5\%)$

Sfârşit

Bibliography

- Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.
- Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.
- Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
- Spiegel, M. R., L. J. Stephens, *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw Hill, 3rd edition, 1999.