# Statistică - Cursul 11

Olariu E. Florentin

May, 2016

## Table of contents

## Teste de semnificaţie

Paşii care trebuie urmaţi când întreprindem un test de semnificaţie:

1-2. Se formulează cele două ipoteze: $H_0$ şi $H_a$: $H_a$ va fi acceptată dacă $H_0$ este respinsă.

3. Se alege un nivel de semnificaţie $\alpha$ - cât de semnificative trebuie să fie evidenţele pentru a respinge $H_0$.
4. Se calculează statistica sau scorul testului.
5. Se determină valoarea critică.
6. Se compară scorul cu valoarea critică şi, dacă este cazul, se respinge $H_0$ şi se acceptă $H_a$, altfel nu se acceptă $H_a$.

# Testul $Z$



- Testul $Z$ este un test statistic asupra ipotezelor utilizat pentru statistici care urmează o distribuţie normală atunci când ipoteza nulă este normală.
- Datorită Teoremei Limite Centrale putem folosi testul $Z$ (deja am făcut acest lucru în cazul testului proporţiilor) când populaţia este aproximativ normală, dar doar pentru eşantioane mari ($n \geqslant 30$).
- Testul $Z$ se bazează pe distribuţia normală; pentru eşantioane mici, acest test de semnificaţie atunci când eşantionul provine dintr-o populaţie normală sau foarte aproape de una normală.
- Pentru un nivel de semnificaţie dat testul $Z$ oferă doar o valoare critică.

# Testul $Z$ - Inferenţă asupra mediei unei populaţii ($\sigma$ cunoscută)



- Vom considera o populaţie statistică a cărei dispersie ($\sigma^2$) este cunoscută.
- Populaţia este normal distribuită şi dorim să testăm o ipoteză asupra mediei populaţiei.
- Testul poate fi întreprins chiar dacă populaţia nu este distribuită normal, dacă utilizăm eşantioane destul de mari.
- Dacă $\mu_0$ este media populaţiei (presupusă în ipoteza nulă), atunci următoarea statistică $\dfrac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}$ este distribuită normal standard: $N(0,1)$.
- Testul $Z$ are loc astfel:

# Testul $Z$ - Inferenţă asupra mediei unei populaţii ($\sigma$ cunoscută)

1. Formulăm *ipoteza nulă*, care susţine că media populaţiei ia o anumită valoare:

$$\boxed{H_0 : \mu = \mu_0}$$

2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantion. Putem avea trei tipuri de ipoteză alternativă

$$\boxed{H_a : \mu < \mu_0} \quad (\textit{asimetrică la stânga}) \text{ sau}$$

$$\boxed{H_a : \mu > \mu_0} \quad (\textit{asimetrică la dreapta}) \text{ sau}$$

$$\boxed{H_a : \mu \neq \mu_0} \quad (\textit{simetrică}).$$

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

# Testul $Z$ - Inferenţă asupra mediei unei populaţii ($\sigma$ cunoscută)



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

$$\boxed{z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}}$$

5. Determinăm valoarea critică corespunzătoare lui $\alpha$

$\boxed{z^* = qnorm(\alpha)}$ 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 $Z$ - Inferenţă asupra mediei unei populaţii ($\sigma$ cunoscută)

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 $\alpha$* pentru a respinge ipoteza nulă (*încercarea de a respinge $H_0$ eşuează*).

## Testul $Z$ - Inferenţă asupra mediei unei populaţii - Exemplu

*Exemplu.*

- O colonie de şoareci de laborator constă din câteva mii de animale. Greutatea lor urmează o lege normală cu deviaţia standard $\sigma = 5g$ şi o medie de $30g$.
- Pentru un eşantion de 25 de şoareci se găseşte o medie de $32g$; Este aceasta valoare semnificativă statistic (cu 5% nivel de semnificaţie)? dar cu 1% nivel de semnificaţie?

*Soluţie.*

- Se pare că adevărata medie de greutate a întregii populaţii este diferită de cea afirmată ($\mu_0 = 30g$).
- Ştiind că populaţia urmează o distribuţie normală, şi că deviaţia standard a populaţiei este cunoscută putem întreprinde un test $Z$ asupra mediei.

# Testul $Z$ - Inferenţă asupra mediei unei populaţii - Exemplu

- Colectăm informaţiile legate de populaţie şi de eşantion: $\mu_0 = 30$, $\sigma = 5$, $n = 25$, $\overline{x}_n = 32$.

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

$$H_0 : \mu = 30 \quad H_a : \mu \neq 30.$$

3. $\alpha = 0.05$.

4. Scorul $z$

$$z = \frac{\overline{x}_n - \mu_0}{\sigma / \sqrt{n}} = \frac{32 - 30}{5/\sqrt{25}} = 2.$$

5. Valoarea critică este $z^* = -qnorm(\alpha/2) = 1.9599$, pentru $\alpha = 5\%$.

6. Cum $z > |z^*|$, putem respinge ipoteza nulă, şi să acceptăm că adevărata medie a populaţiei nu este $\mu_0 = 30g$.

## Testul $Z$ - Inferenţă asupra mediei unei populaţii - Exemplu



- Reluăm ultimii doi paşi pentru celălalt nivel de semnificaţie:

5'. Pentru $\alpha = 1\%$ 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 adevărata medie a populaţiei este diferită de $30g$).

## Testul $Z$ - Inferenţă asupra mediei unei populaţii - Exemplu revăzut

- Privind din nou informaţiile din eşantion putem observa că media de selecţie, $\overline{x}_n$, este mai mare decât media presupusă a populaţiei.
- Într-un astfel de caz am putea formula o ipoteză alternativă asimetrică la dreapta.

1-2. Noile ipoteze sunt

$$H_0 : \mu = 30 \quad H_a : \mu > 30.$$

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

$$z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{32 - 30}{5/\sqrt{25}} = 2.$$

## Testul $Z$ - Inferenţă asupra mediei unei 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 adevărată a populaţiei este mai mare decât $\mu_0 = 30g$.

- Pentru celălalt nivel de semnificaţie $(1\%)$:

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

6'. Deoarece $z < z^*$, nu putem respinge ipotez nulă cu $1\%$ nivel de semnificaţie (nu există dovezi suficiente cu $1\%$ nivel de semnificaţie pentru a susţine că adevărata media a populaţiei este mai mare decât $30g$).

# Testul $Z$ - Inferenţă asupra mediei unei populaţii - Observaţii

- Merită observat că pentru diverse niveluri de semnificaţie putem avea concluzii diferite: ipoteza nulă poate fi respinsă pentru un anumit nivel de semnificaţie iar cu un alt nivel încercarea poate eşua.
- Dacă ipoteza nulă este respinsă cu 1% nivel de semnificaţie, atunci va fi respinsă şi cu 5%; altfel spus, dacă ipoteza nulă nu poate fi respinsă pentru 5%, nu va putea fi respinsă nici pentru 1%.
- Ipoteza alternativă trebuie formulată conform datelor din eşantion: când $\mu_0 << \overline{x}_n$ putem formula o ipoteză alternativă asimetrică la dreapta, $H_a : \mu > \mu_0$, când $\mu_0 >> \overline{x}_n$ putem formula o ipoteză alternativă asimetrică la stânga, $H_a : \mu < \mu_0$.
- Dacă media de selecţie nu este nici atât de mare şi nici atât de mică prin comparaţie cu media presupusă a populaţiei, putem presupune că media este doar diferită de valoarea din ipoteza nulă, $H_a : \mu \neq \mu_0$.

# Testul $Z$ - Exerciţii



I. Reluaţi testul $Z$ pentru exerciţiul anterior cu $\overline{x}_n = 27g$ şi $\sigma = 6$. (Folosiţi ambele nivele de semnificaţie.)

II. Se afirmă că studenţii unei anumite universităţi vor obţine la un anumit test o medie de 35 de puncte cu $\sigma = 4$. Are această afirmaţie susţinere ştiind că pentru un eşantion aleator se obţin următoarele rezultate la test: $33, 42, 38, 37, 30, 42$? Întreprindeţi un test corespunzător pentru $\alpha = 5\%$. Se presupune că rezultatele la acest test sunt normal distribuite.

III. Conform National Center for Health Statistics, înălţimea medie a femeilor din SUA (care este normal distribuită) este de 63.7 in cu o deviaţie standard cunoscută de $\sigma = 2.75$ in. Pentru un eşantion aleator simplu de 50 femei care lucrează în domeniul serviciilor medicale se găseşte o înăţime de 65.2. Testaţi ipoteza că media de înăţime a femeilor care lucrează în sănătate este diferită de 63.7 in. 5% nivel de semnificaţie.

# Testul $T$



- Testul $T$ testează ipotezele statistice pentru statistici care urmează o distribuţie Student.
- Vom descrie mai întâi un test $T$ pentru a compara media unei populaţii cu o valoare cunoscută din $H_0$.
- Testul $T$ se foloseşte atunci când deviaţia standard a populaţiei (distribuită normal) este necunoscută.
- Un test $T$ este de asemeni potrivit când avem de-a face cu eşantioane mici ($n < 30$) pentru populaţii care sunt aproximativ normal (din Teorema Limită Centrală).
- Urmând aceste observaţii putem spune că un test $T$ pentru media unei populaţii este complementar unui test $Z$.
- În secţiunea următoare descriem testul $T$ pentru media unei populaţii cu dispersia necunoscută.

## Testul $T$ - Inferenţă asupra mediei unei populaţii ($\sigma$ necunoscută)

- Considerăm o populaţie statistică a cărei dispersie ($\sigma^2$) este necunoscută.
- Populaţia este normal distribuită şi dorim să testăm o ipoteză asupra adevăratei medii a populaţiei.
- Testul poate fi întreprins dacă populaţia are o distribuţie foarte apropiată de cea normală, când eşantionul la dispoziţie este mic (altfel se utilizează un test $Z$).
- Dacă $\mu_0$ este media populaţiei (presupusă în ipoteza nulă), atunci următoarea statistică $\dfrac{\overline{x}_n - \mu_0}{s/\sqrt{n}}$ este distribuită Student cu $(n-1)$ grade de libertate: $T(n-1)$.
- O diferenţă faţă de un test $Z$ constă în înlocuirea deviaţiei standard a populaţiei, $\sigma$, cu deviaţia standard a eşantionului $s$.
- Testul $T$ decurge astfel:

## Testul $T$ - Inferenţă asupra mediei unei populaţii ($\sigma$ necunoscută)

1. Formulăm mai întâi *ipoteza nulă*, care susţine că media populaţiei ia o anumită valoare:

$$\boxed{H_0: \ \mu = \mu_0}$$

2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantion. Putem avea trei tipuri de ipoteză alternativă

$$\boxed{H_a: \ \mu < \mu_0} \quad (\textit{asimetrică la stânga}) \text{ sau}$$

$$\boxed{H_a: \ \mu > \mu_0} \quad (\textit{asimetrică la dreapta}) \text{ sau}$$

$$\boxed{H_a: \ \mu \neq \mu_0} \quad (\textit{simetrică}).$$

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

## Testul $T$ - Inferenţă asupra mediei unei populaţii ($\sigma$ necunoscută)

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

$$\boxed{t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}}}$$

5. Determinăm valoarea critică corespunzătoare lui $\alpha$

$\boxed{t^* = qt(\alpha, n-1)}$ 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 ($t^*$

## Testul $T$ - Inferenţă asupra mediei unei populaţii ($\sigma$ necunoscută)



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

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

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

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

Dacă scorul $t$ nu aparţine zonei de respingere vom spune că *nu există suficiente dovezi cu nivelul de semnificaţie* $\alpha$ pentru a respinge ipoteza nulă (*încercarea de a respinge* $H_0$ *eşuează*).

## Testul $T$ - Inferenţă asupra mediei unei populaţii - Exemplu

*Example*

- Concentraţia de CO (monixid de carbon) se măsoară cu un aparat numit spectrofotometru care are o precizie de aporape 100 ppm. Aceste aparate trebuie calibrate zilnic măsurând concentraţia de CO din eşantioane de gaz industrial care au o concentraţie controlată de 70 ppm. Dacă aparatul dă valori "apropiate" de 70 ppm poate fi folosit, dacă nu, trebuie ajustat.
- Presupunem că această concentraţie urmează o distribuţie normală dar deviaţia standard este necunoscută. Într-o anumită zi se obţin următoarele valori
$$58 \;\; 71 \;\; 67 \;\; 64 \;\; 62.$$
- Patru dintre aceste valori sunt mai mici decât 70; putem explica aceasta doar pe seama întâmplării? Sau aceasta arată că aparatul trebuie ajustat?

# Testul $T$ - Inferenţă asupra mediei unei populaţii - Exemplu

*Solution*

- Este posibil ca aparatul să necesite o ajustare, deci vom testa ipoteza că media este diferită de $\mu_0 = 70$.
- Cum populaţia urmează o distribuţie normală, iar că deviaţia standard a populaţiei este necunoscută putem întreprinde un test $T$ asupra mediei.
- Informaţiile privind populaţia şi eşantionul: $\mu_0 = 70$, $s = 4.9295$, $n = 5$, $\overline{x}_n = 64.4$.

1-2. Putem formula ipoteza nulă şi o ipoteză alternativă simetrică

$$H_0 : \mu = 70 \quad H_a : \mu \neq 70.$$

# Testul $T$ - Inferenţă asupra mediei unei populaţii - Exemplu



3. $\alpha = 0.05$.
4. Scorul $t$
$$t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} = \frac{64.4 - 70}{4.9295/\sqrt{5}} = -2.5402.$$
5. Valoarea critică este $t^* = -qt(\alpha/2, 4) = 2.7764$, for $\alpha = 5\%$.
6. Cum $|t| < |t^*|$, încercarea de a respinge ipoteza nulă eşuează (nu există suficiente dovezi pentru nivelul de 5% pentru a susţine că adevărata medie a populaţiei este diferită de 70 ppm).

## Testul $T$ - Inferenţă asupra mediei unei populaţii - Exemplu



- Reluăm ultimii doi paşi pentru celălalt nivel de semnificaţie:

5'. Pentru $\alpha = 1\%$ valoarea critică este $t^* = -qt(\alpha/2, 4) = 4.6040$.

6'. Since $|t| < |t^*|$, încercarea de a respinge ipoteza nulă eşuează (nu există suficiente dovezi pentru nivelul de 1% pentru a susţine că adevărata medie a populaţiei este diferită de 70 ppm).

- Merită observat că reluarea ultimilor doi paşi cu un $\alpha$ mai mic nu era necesară, deoarece, ştim deja, ipoteza nulă nu a putut fi respinsă cu 5% nivel de semnificaţie.

## Testul $T$ - Inferenţă asupra mediei unei populaţii - Exemplu revăzut



- Privind din nou la datele din eşantion putem observa că media de selecţie, $\overline{x}_n$, este mai mică decât media presupusă a populaţiei.
- Într-o astfel de situaţie putem formula o ipoteză alternativă asimetrică la stânga.

1-2. Noile ipoteze sunt

$$H_0 : \mu = 70 \quad H_a : \mu < 70.$$

3. $\alpha = 0.05$.

4. Scorul $t$

$$t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} = \frac{64.4 - 70}{4.9295/\sqrt{5}} = -2.5402.$$

## Testul $T$ - Inferenţă asupra mediei unei populaţii - Exemplu revăzut



5. Valoarea critică este $t^* = qt(\alpha, 4) = -2.1318$, pentru $\alpha = 5\%$.
6. Deoarece $t < t^*$, putem respinge ipoteza nulă şi să acceptăm că media adevărată a populaţiei este mai mică decât 70 ppm.

- Cu celălalt nivel de semnificaţie (1%):

5'. Valoarea critică este $t^* = qt(\alpha, 4) = -3.7469$.

6'. Deoarece $t > t^*$, nu putem respinge ipotez nulă cu 1% nivel de semnificaţie (nu există dovezi suficiente cu 1% nivel de semnificaţie pentru a susţine că adevărata media a populaţiei este mai mică decât 70 ppm.

## Testul $T$ - Exerciţii



I. O asociaţie studenţească susţine că un student călătoresc, în fiecare zi, în medie, 25 de minute pe drumul către universitate. Universitatea determină un eşantion aleator obţinut de la 32 de studenţi. Eşantionul are o medie de selecţie de 19.4 minute şi o deviaţie standard de 9.6 minute. Datele acestea sunt suficiente statistic pentru a respinge afirmaţiile asociaţiei? Folosiţi $\alpha = 0.01$. (Presupunem că timpul unei călătorii urmează o distribuţie normală.)

II. Se ştie că tinerii adulţi cheltuie săptămânal 40$ pentru mâncare de tip fast food. Un studiu bazat pe un eşantion aleator fromat din 1000 de tineri adulţi (Greenfield Online - USA Today Snapshot) determină o medie săptămânală de 35$ cheltuiţi pe fast food cu o deviaţie standard de 14.50$. Presupunând ca aceste cheltuieli urmează o distribuţie normală întreprindeţi un test statistic potrivit asupra mediei acestor cheltuieli pentru întreaga populaţie.

## Testul $T$ - Exerciţii

III. Locuinţele dintr-un oraş au o valoare medie de 88950\$. Se presupune că locuinţele din apropierea universităţii au un preţ mediu mai mare. Pentru a testa aceasta, se foloseşte un eşantion aleator simplu format din 12 case alese în apropierea universităţii. Media lor de selecţie este 92460\$, cu o deviaţie standard de 5200\$. Întreprindeţi un test statistic pe baza acestor date cu $\alpha = 5\%$. (Se presupune că preţurile caselor sunt normal distribuite.)

**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.