4. ESTIMAREA ȘI INTERVALELE DE ÎNCREDERE

- Estimarea punctuală și prin interval de încredere a parametrilor unei populații
- Intervalul de încredere pentru estimarea mediei unei populații
- Intervalul de încredere pentru estimarea diferenței dintre mediile a două populații
- Intervalul de încredere pentru estimarea proporției unei populații
- Intervalul de încredere pentru estimarea diferenței dintre proporțiile a două populații

OBIECTIVELE CURSULUI

La finalizarea acestui capitol, studentul va fi capabil să:

O7-1: să calculeze și să interpreteze o estimare punctuală și prin interval de încredere a mediei unei anumite populații

O7-2: să calculeze și să interpreteze o estimare punctuală și prin interval de încredere a proporției unei anumite populații

O7-3: să calculeze și să interpreteze o estimare prin interval de încredere a diferenței dintre mediile a două populații

O7-4: să calculeze și să interpreteze o estimare prin interval de încredere a diferenței dintre proporțiile a două populații

INTRODUCERE

În cursul precedent am trecut în revistă principalele metode de eșantionare (simplă aleatoare – în care fiecare membru al populației generale are aceeași șansă de a fi selectat în eșantion, sistematică, stratificată, tip cluster și de convenienţă) și motivele care stau la baza acestui proces de selectare a unităților statistice dintr-o populație (timp îndelungat, costuri ridicate, rezultatele la nivel de eșantion sunt adecvate, natura distructivă a unor teste, imposibilitatea verificării tuturor observațiilor).

În cadrul exemplelor privind distribuțiile de sondaj ale mediilor eșantioanelor am considerat faptul că media, deviația standard și forma distribuției populației de referință sunt cunoscute. În cele mai multe situații reale care pot apare în practică însă asemenea informații nu sunt disponibile, principalul scop al cercetării statistice prin sondaj fiind estimarea acestor parametri ai populației studiate în baza statisticilor calculate la nivel de eșantion (de exemplu, prin selectarea în mod aleator a unui eșantion dintr-o anumită colectivitate de interes, media calculată la nivelul acestuia este utilizată în vederea obținerii unei estimații a mediei populației generale).

În acest curs, vom prezenta câteva aspecte importante ale procesului de cercetare statistică prin sondaj pornind de la estimarea punctuală a parametrilor unei anumite populații, respectiv utilizarea unei valori unice, calculată pe baza datelor din eșantion, în vederea obținerii unei estimații a valorii necunoscute a parametrului corespunzător din cadrul populației. O abordare mult mai consistentă din punct de vedere informațional constă însă în prezentarea unui interval de valori în care ne așteptăm, cu o anumită probabilitate sau nivel de încredere, să se afle valoarea parametrului populației studiate.

ESTIMAREA PUNCTUALĂ ȘI PRIN INTERVAL DE ÎNCREDERE A PARAMETRILOR UNEI POPULAȚII

O ESTIMARE PUNCTUALĂ a unui parametru unei populații reprezintă o valoare unică a statisticii corespunzătoare, determinată la nivelul unui eșantion reprezentativ extras aleator din aceasta. De exemplu, turismul este o sursă foarte importantă de venituri pentru majoritatea țărilor din Marea Caraibilor. Presupunem că Biroul pentru turism din Aruba dorește să obțină o valoare estimativă a volumului mediu de cheltuieli zilnice realizat de turiștii care vizitează insula. Deoarece nu ar fi rezonabilă contactarea fiecărui persoane venită să viziteze această regiune, un eșantion de 500 de turiști este selectat în mod aleator în momentul părăsirii țării și fiecare persoană este chestionată în legătură cu volumul de cheltuieli pe care l-a presupus vizita în Aruba. Valoarea medie zilnică cheltuită la nivelul eșantionului celor 500 de turiști (de ex. 400 USD) reprezintă o estimare punctuală a mediei populației acestora.

Un estimator punctual pentru valoarea unui parametru al unei populații nu ne oferă însă posibilitatea conturării unei imagini complete asupra fenomenului. Deși ne așteptăm ca valoarea estimatorului calculată la nivel de eșantion să fie apropiată de valoarea parametrului corespunzător din cadrul populației, în analiză dorim să cunoaștem în același timp și cât de aproape este acesta. Cu alte cuvinte, încercăm să cuantificăm incertitudinea care planează asupra estimării noastre punctuale sau nivelul de încredere atașat acestei presupuneri.

Astfel, ar fi mult mai interesant să putem afirma faptul că există 95% șanse ca valoarea medie „adevărată” a cheltuielilor zilnice ale unui turist în Aruba să fie cuprinsă între 380 USD și 420 USD. Sau, într-un mod mai general, 95% din intervalele de încredere construite prin utilizarea tuturor eșantioanelor de volum n = 500 extrase în mod aleator din populația turiștilor din Aruba vor conține valoarea adevărată a parametrului – „cheltuielile medii zilnice ale unui turist”.

În general, intervalul de încredere pentru media unei populații μ, calculat pe baza mediei eșantionului $\bar{x}$ și a erorii limită admisibile d este:

$$\bar{x} - d < \mu < \bar{x} + d$$

INTERVALUL DE ÎNCREDERE reprezintă o plajă de valori construită pe baza datelor din eșantion în care este probabil să fie cuprinsă valoarea „necunoscută” și „adevărată” a parametrului corespunzător al populației generale, din care eșantionul fost extras.

Probabilitatea specifică estimării acestui interval se numește NIVEL DE ÎNCREDERE și are de obicei valoarea de 95%. Aceasta exprimă siguranța cu care se afirmă faptul că intervalul de încredere cuprinde parametrul estimat, în cazul exemplului nostru, media populației generale μ:

$$P(\bar{x} - d < \mu < \bar{x} + d) = 1 - \alpha = 0,95 \text{ sau } 95\%$$

$\alpha = 1 - 0,95 = 0,05$ sau 5% reprezintă pragul sau nivelul de semnificație asociat estimării.

In concluzie, intervalul de încredere evidențiază nivelul de precizie al estimării parametrului din populația de referință.

INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA MEDIEI UNEI POPULAȚII

Conform teoremei limită centrală, în situația când volumul eșantionului este mare $(n \geq 30)$ distribuția de sondaj a mediilor eșantioanelor poate fi aproximată printr-o DISTRIBUȚIE NORMALĂ având media μ și deviația standard $\sigma/\sqrt{n}$. Chiar dacă media populației generale nu este cunoscută, distribuția de sondaj a mediilor eșantioanelor ne arată cum sunt repartizate valorile $\overline{x}$ în jurul lui μ. Astfel, aflăm informații despre toate diferențele $\overline{x} - \mu$ sau erorile de sondaj posibile.

Așa cum am văzut anterior, orice distribuție probabilistică normală *(inclusiv distribuția de sondaj a mediilor eșantioanelor)* poate fi convertită într-o distribuție normală standard prin schimbarea de variabilă:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Deoarece utilizăm în calcule distribuția de sondaj a mediilor eșantioanelor în locul distribuției variantelor individuale ale variabilei x_i, vom folosi în locul deviației standard a populației generale σ, deviația standard a distribuției de sondaj a mediilor tuturor eșantioanelor extrase aleator din aceasta sau eroarea standard a mediei (SEM):

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Valoarea $z_{\overline{x}}$ corespunzătoare mediei eșantionului $\overline{x}$ extras din populația generală cu media μ devine:

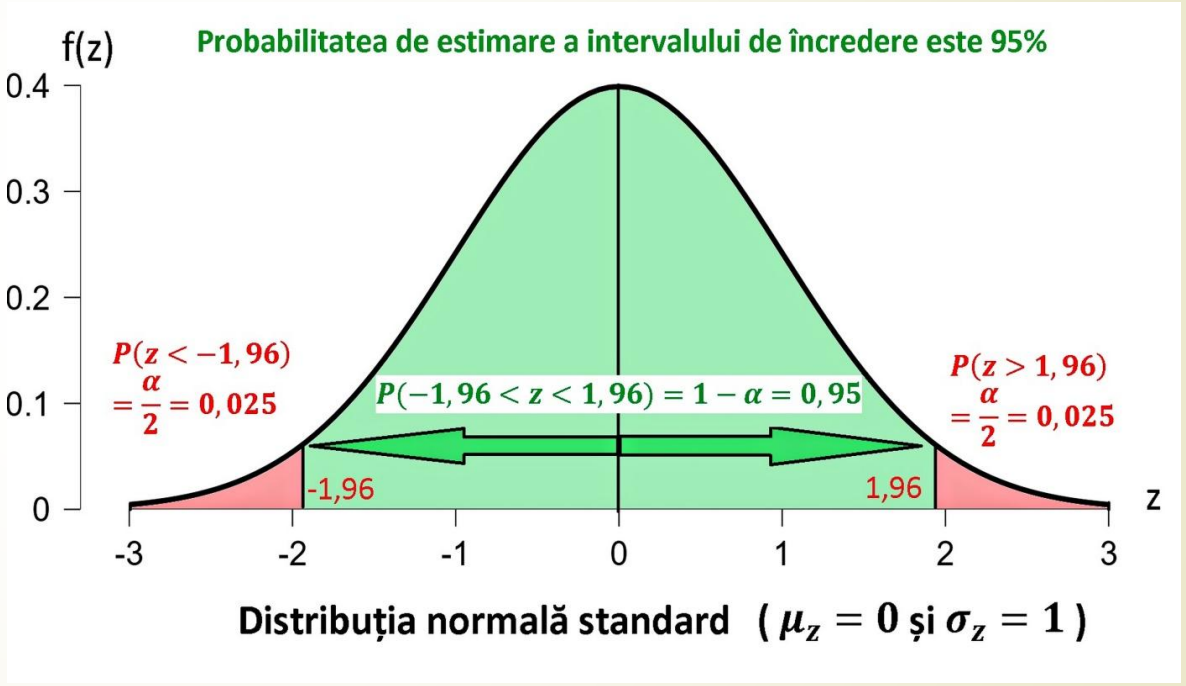
$$z_{\overline{x}} = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Vom utiliza litera α pentru a indica probabilitatea că eroarea de sondaj $|\bar{x} - \mu|$ este mai mare decât cea specifică nivelului de precizie dorit, aceasta fiind repartizată simetric, jumătate $(\alpha/2)$ în partea stângă și cealaltă jumătate $(\alpha/2)$ în partea dreaptă a distribuției de sondaj a mediilor eșantioanelor. $1 - \alpha$ va reprezenta probabilitatea ca media eșantionului nostru $\bar{x}$ să genereze o eroare de sondaj $|\bar{x} - \mu|$ mai mică sau egală cu cea specifică nivelului de precizie cerut de un anumit studiu.

$$-z_{1-\frac{\alpha}{2}} < z_{\bar{x}} < z_{1-\frac{\alpha}{2}}$$

$$-z_{1-\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{1-\frac{\alpha}{2}}$$

$$\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$



Pentru un nivel de încredere:

$1 - \alpha = 90\%$ avem $z_{1-\frac{\alpha}{2}} = 1,64$

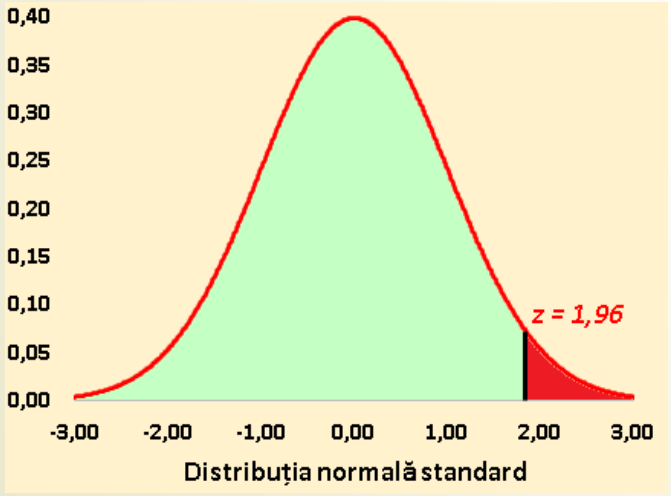
$1 - \alpha = 95\%$ avem $z_{1-\frac{\alpha}{2}} = 1,96$

$1 - \alpha = 99\%$ avem $z_{1-\frac{\alpha}{2}} = 2,58$

$1 - \alpha$ reprezintă coeficientul sau nivelul de încredere și $z_{1-\frac{\alpha}{2}}$ reprezintă acea valoare a lui z care delimitează o suprafață de $\alpha/2$ în extremitatea superioară a distribuției normale standard.

Sintetizând, putem afirma faptul că, există o probabilitate $1-\alpha$ ca intervalul $\overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ să conțină media adevărată a populației generale μ din care eșantionul de medie $\overline{x}$ a fost extras.

Pentru determinarea lui $z_{1-\alpha/2}$ putem folosi tabelele distribuției normale standard sau putem genera direct cu ajutorul programului Excel această valoare, utilizând, de exemplu, pentru un nivel de încredere de 95% ($1-\alpha = 0,95$ și $\alpha = 0,05$) funcția: =NORM.S.INV(1-0,05/2)= 1,96



DISTRIBUTIA NORMALA STANDARD

z	Probabilitati cumulate									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
...	...	...	...	...	...	...	...	...	...	...
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857

$z_{1-\frac{\alpha}{2}} = 1,96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \frac{\alpha}{2} = 1 - \frac{0,05}{2} = 1 - 0,025 = 0,975$

Intervalele de încredere pentru media unei populații μ au fost determinate anterior după formula

$$\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

presupunând cunoscută deviația standard a populației σ.

În cele mai multe situații practice însă, σ este necunoscută și vom utiliza ca estimator punctual al acesteia deviația standard ajustată, calculată la nivel de eșantion, prin raportarea sumei pătratelor abaterilor nivelurilor individuale de la media eșantionului la $n-1$ în loc de n, astfel:

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

Astfel, estimatorul deviației standard a distribuției de sondaj a mediilor eșantioanelor devine:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

În situația când deviația standard a populației generale σ nu este cunoscută distribuția de sondaj a mediilor eșantioanelor urmează o distribuție probabilistică continuă STUDENT (t), care este foarte apropiată de distribuția normală standard.

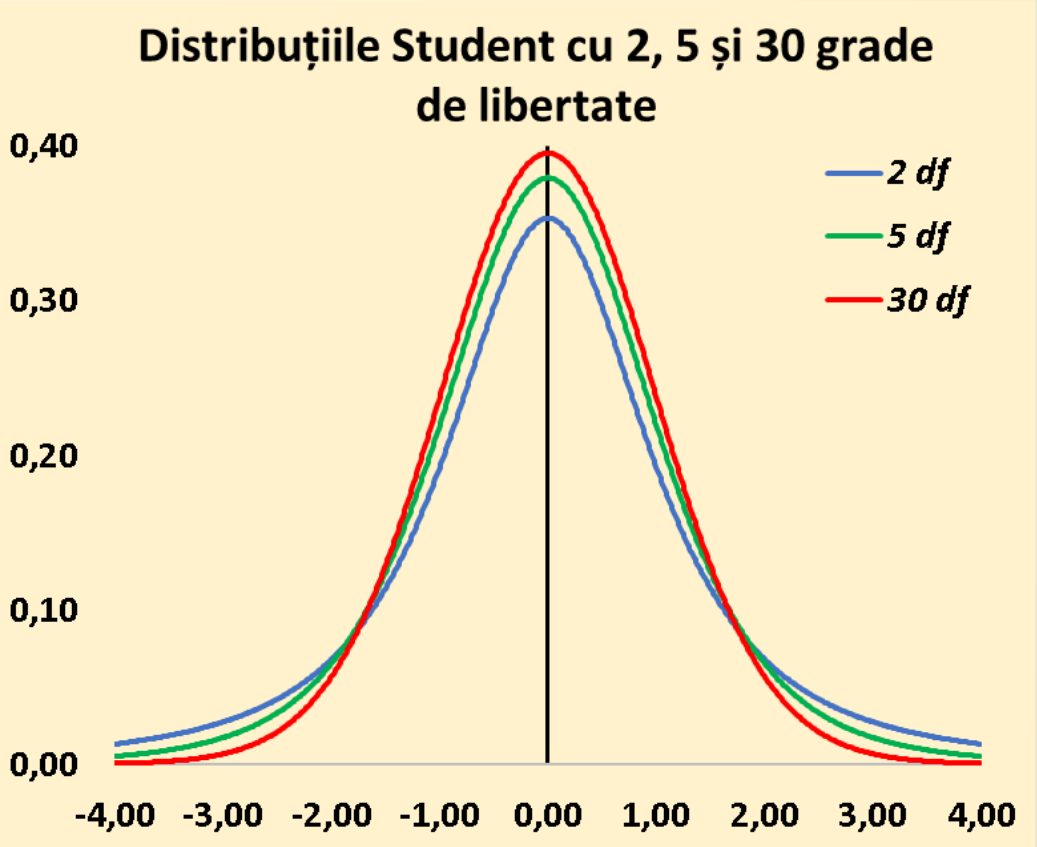
Distribuția Student prezintă următoarele caracteristici:

- are o formă de clopot, fiind distribuită simetric in jurul mediei de valoare zero
- are o deviație standard mai mare decât 1
- există o familie de distribuții t, câte una pentru fiecare grad de libertate
- pe măsură ce volumul eșantionului crește, distribuția t tinde către distribuția normală

Gradele de libertate

sunt specifice variantelor x_i ale unei variabile statistice X care au posibilitatea să varieze liber în momentul în care valoarea indicatorului statistic calculat pe baza lor a fost fixată.

În cazul mediei aritmetice avem $n-1$ grade de libertate. Astfel, pentru media unei grupe de 12 de studenți la un examen, care are valoarea 7,50 notele unui număr de n – 1 = 11 studenți pot varia în mod liber, dar nota ultimului student luată în calculul acestui indicator statistic va fi „fixată” la valoarea necesară obținerii nivelului mediu de 7,50.



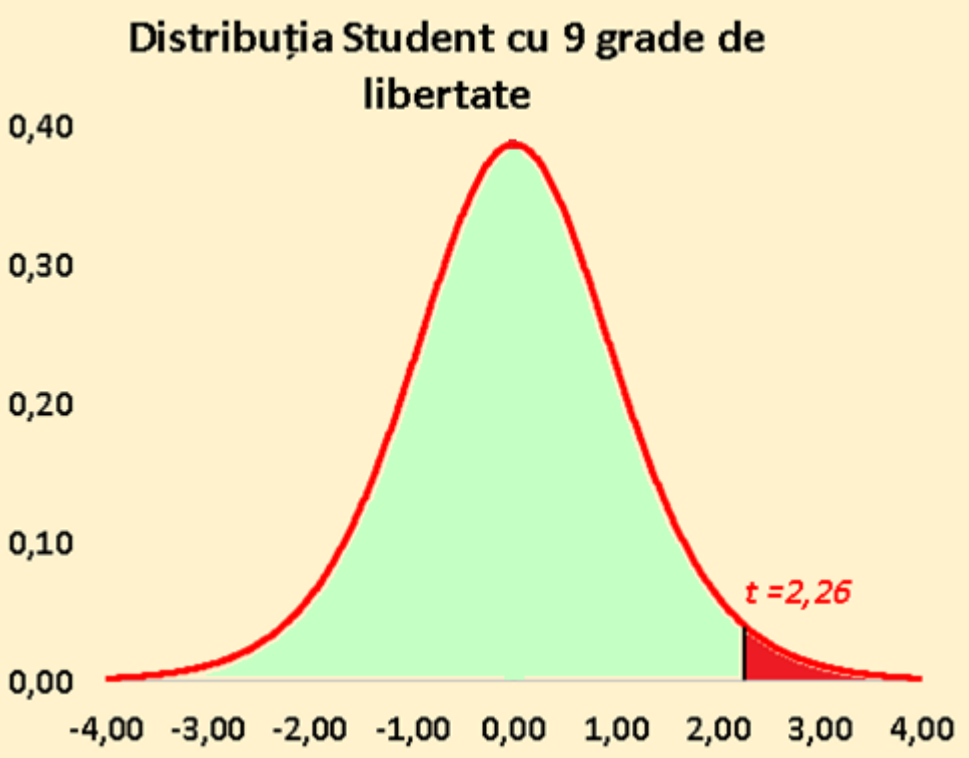
- deoarece distribuția t prezintă o împrăștiere mai mare decât distribuția z, valoarea absolută a lui t pentru un anumit nivel de încredere specificat este superioară valorii corespunzătoare a lui z.
- pe măsura utilizării unor eșantioane de volum superior ($n \geq 30$) diferența dintre intervalele de încredere generate prin utilizarea valorilor t și z devine neglijabilă; pentru $n \to \infty$, în cazul unui nivel de încredere de 95% ($1 - \alpha = 0,95$): $t_{n-1,1-\frac{0,05}{2}} \approx z_{1-\frac{0,05}{2}} = 1,96$.

În concluzie, intervalul de încredere pentru media unei populații μ având σ necunoscută se obține astfel:

$$\overline{x} - t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

unde $t_{n-1,\,1-\frac{\alpha}{2}}$ reprezintă valoarea lui t care delimitează o suprafață de $\alpha/2$ în extremitatea dreaptă a distribuției Student cu $n - 1$ grade de libertate

Pentru determinarea lui $t_{n-1,1-\frac{\alpha}{2}}$ putem folosi tabelele distribuției Student sau putem genera direct cu ajutorul programului Excel această valoare, utilizând, de exemplu, pentru un nivel de încredere de 95% ($1-\alpha=0,95$ și $\alpha=0,05$) funcția: = T.INV(1-0,05/2;9) = T.INV(0,975;9) = 2,26. Deoarece intervalul de încredere 95%, lasă 5% din suprafața de sub curba distribuției Student egal împărțită între cele 2 cozi laterale (2,5% în stânga și 2,5% în dreapta), avem nevoie de acea valoare a lui t în dreapta căreia să rămână 2,5% iar în partea stângă restul de 97,5%.

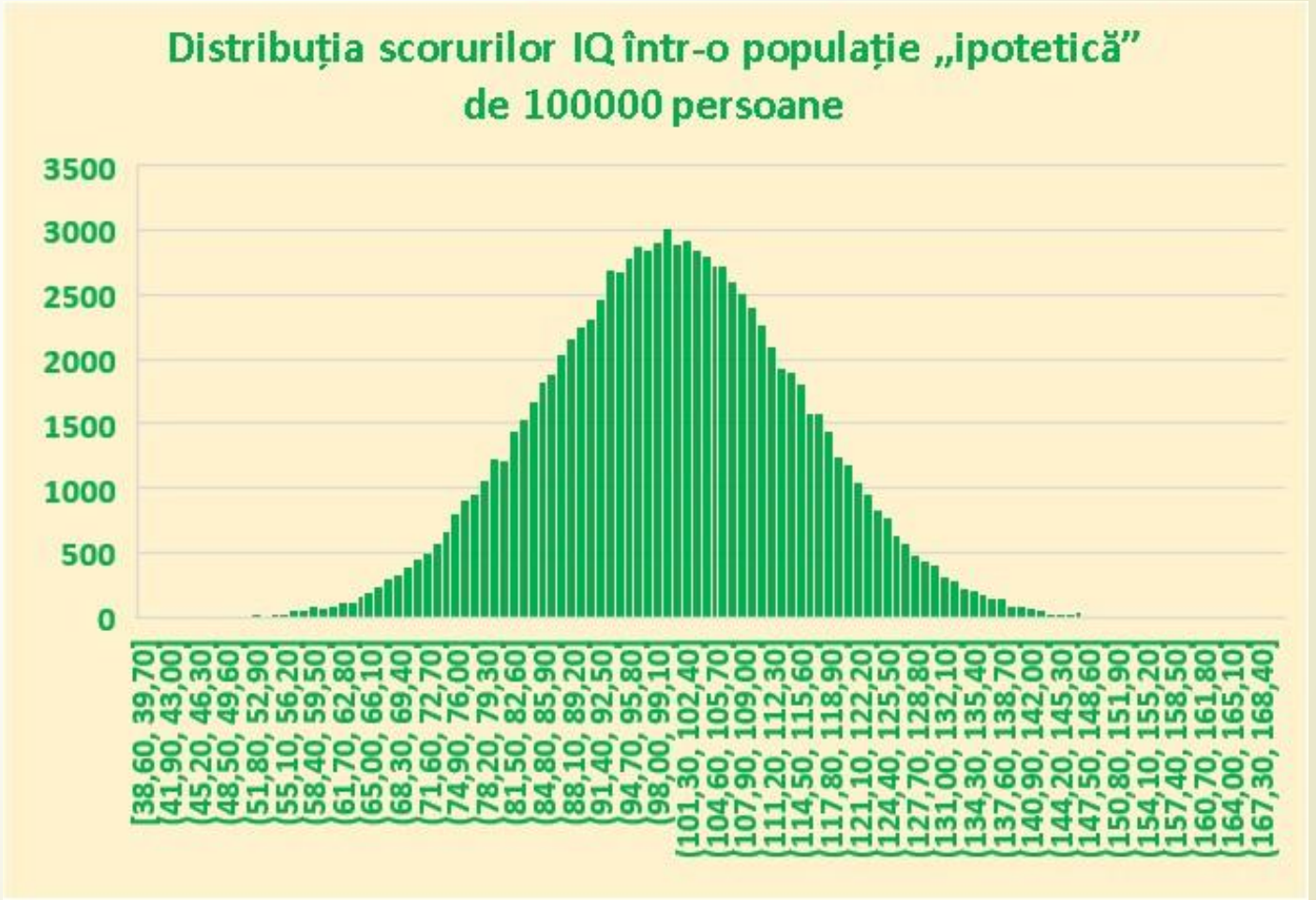


DISTRIBUȚIA STUDENT "t"

Grade de libertate	Suprafața delimitată în extremitatea dreaptă				
	0,10	0,05	0,025	0,01	0,005
1,00	3,0777	6,3138	12,7062	31,8205	63,6567
2,00	1,8856	2,9200	4,3027	6,9646	9,9248
3,00	1,6377	2,3534	3,1824	4,5407	5,8409
4,00	1,5332	2,1318	2,7764	3,7469	4,6041
5,00	1,4759	2,0150	2,5706	3,3649	4,0321
6,00	1,4398	1,9432	2,4469	3,1427	3,7074
7,00	1,4149	1,8946	2,3646	2,9980	3,4995
8,00	1,3968	1,8595	2,3060	2,8965	3,3554
9,00	1,3830	1,8331	2,2622	2,8214	3,2498
10,00	1,3722	1,8125	2,2281	2,7638	3,1693
...	...	...	...	...	...

Pentru exemplificare, considerăm distribuția scorurilor IQ într-o populație „ipotetică” de 100000 persoane, având media de 100,13 și deviația standard 15,04. Scorurile IQ variază pe o plajă de valori cuprinsă între 38,60 și 170,18, fiind distribuite aproximativ normal.

Persoana	IQ
1	76,11
2	86,35
3	94,66
4	113,33
5	95,97
6	106,01
7	81,19
8	120,67
9	81,62
10	86,30
...	...
99998	136,34
99999	120,74
100000	69,65



Vom calcula intervalele de încredere pentru media populației generale, prin utilizarea unor eșantioane de volume 10, 20 și 30 extrase aleator din colectivitatea generală a celor 100000 de persoane analizate prin prisma scorurilor IQ. În toate cazurile, vom determina intervalele de confidență pentru un nivel de încredere de 95% atât în ipoteza cunoașterii deviației standard σ a populației studiate cât și în situația utilizării, în calitate de estimator al acesteia, a deviației standard ajustate s, calculate la nivelul fiecărui eșantion în parte.

a) Eșantion de volum n = 10 ($\overline{x} = 101,78$, $s = 8,22$) extras în mod aleator din populația generală de volum N = 100000 persoane ($\mu = 100,13$, $\sigma = 15,04$)

IQ
100,20
107,37
94,93
102,90
100,97
97,36
102,80
87,29
105,36
118,58

❑ Dacă cunoaștem deviația standard a populației $\sigma = 15,04$, intervalul de încredere pentru media populației generale μ cu o probabilitate de 95% ($\alpha = 0,05$, $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0,05}{2}} = z_{0,975} = 1,96$):

$$\overline{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$101,78 - 1,96 \cdot \frac{15,04}{\sqrt{10}} \leq \mu \leq 101,78 + 1,96 \cdot \frac{15,04}{\sqrt{10}} \quad 92,46 \leq \mu \leq 111,10$$

❑ Dacă nu cunoaștem deviația standard a populației, intervalul de încredere pentru media populației generale μ cu o probabilitate de 95% ($\alpha = 0,05$, $t_{n-1,1-\frac{\alpha}{2}} = t_{10-1,1-\frac{0,05}{2}} = 2,26$):

$$\overline{x} - t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

$$101,78 - 2,26 \cdot \frac{8,22}{\sqrt{10}} \leq \mu \leq 101,78 + 2,26 \cdot \frac{8,22}{\sqrt{10}} \quad 95,90 \leq \mu \leq 107,65$$

b) Eșantion de volum n = 20 ($\bar{x} = 99,56$, $s = 10,75$) extras în mod aleator din populația generală de volum N = 100000 persoane ($\mu = 100,13$, $\sigma = 15,04$)

❑ **Dacă cunoaștem deviația standard a populației, IC 95%:**

$$99,56 - 1,96 \cdot \frac{15,04}{\sqrt{20}} \leq \mu \leq 99,56 + 1,96 \cdot \frac{15,04}{\sqrt{20}} \qquad 92,96 \leq \mu \leq 106,15$$

❑ **Dacă nu cunoaștem deviația standard a populației, IC 95%:**

$$99,56 - 2,26 \cdot \frac{10,75}{\sqrt{20}} \leq \mu \leq 99,56 + 2,26 \cdot \frac{10,75}{\sqrt{20}} \qquad 94,52 \leq \mu \leq 104,59$$

IQ	
89,56	93,51
106,13	112,25
100,25	103,01
87,10	104,90
90,15	106,89
81,37	88,17
96,81	91,77
94,34	96,69
112,27	97,23
120,40	118,31

c) Eșantion de volum n = 30 ($\bar{x} = 102,08$, $s = 12,38$) extras în mod aleator din populația generală de volum N = 100000 persoane ($\mu = 100,13$, $\sigma = 15,04$)

❑ **Dacă cunoaștem deviația standard a populației, IC 95%:**

$$102,08 - 1,96 \cdot \frac{15,04}{\sqrt{30}} \leq \mu \leq 102,08 + 1,96 \cdot \frac{15,04}{\sqrt{30}} \qquad 96,70 \leq \mu \leq 107,46$$

❑ **Dacă nu cunoaștem deviația standard a populației, IC 95%:**

$$102,08 - 2,26 \cdot \frac{12,38}{\sqrt{30}} \leq \mu \leq 102,08 + 2,26 \cdot \frac{12,38}{\sqrt{30}} \qquad 97,46 \leq \mu \leq 106,70$$

IQ	
97,46	111,04
97,50	97,17
108,34	104,38
87,70	98,97
98,31	81,49
82,91	.....
97,90	109,87

- **INTERPRETAREA PROBABILISTICĂ A INTERVALELOR DE ÎNCREDERE**: dacă replicăm de un număr foarte mare de ori experimentul de extragere a eșantioanelor de volum *n* dintr-o populație normal distribuită de volum *N*, având o deviație standard cunoscută în prealabil (σ) sau necunoscută (s), $100(1-\alpha)$ procente din toate intervalele de forma $\overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ sau $\overline{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, după caz, vor include media aritmetică adevărată a populației generale (μ). De obicei, $100(1-\alpha) = 95\%$ dintre toate intervalele de încredere construite în această manieră vor conține în interiorul lor media populației generale (μ).
- **INTERPRETAREA PRACTICĂ A INTERVALELOR DE ÎNCREDERE**: atunci când eșantionul a fost extras în mod aleator dintr-o populație normal distribuită, având o deviație standard cunoscută în prealabil (σ) sau necunoscută (s), apreciem cu un nivel de confidență de $100(1-\alpha)$ procente că intervalul de forma $\overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ sau $\overline{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, după caz, va include media aritmetică adevărată a populației generale (μ).

În toate cazurile prezentate anterior, riscul α (de obicei 5%) a fost distribuit bilateral și simetric ($\frac{\alpha}{2} = 2,5\%$ în stânga și $\frac{\alpha}{2} = 2,5\%$ în dreapta). În funcție de a abordarea cercetării putem avea și intervale de încredere unilaterale (la stânga sau la dreapta), situații în care riscul este distribuit integral doar într-o singură parte.

Eroarea limită admisibilă $\Delta_{\overline{x}} = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ sau $\Delta_{\overline{x}} = t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ poate fi calculată cu ajutorul programului Excel utilizând funcțiile: =CONFIDENCE.NORM(α, σ, *n*) sau =CONFIDENCE.T(α, *s*, *n*).

INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA DIFERENȚEI DINTRE MEDIILE A DOUĂ POPULAȚII

Uneori apare în practică necesitatea estimării diferenței între mediile a două populații distribuite normal μ_1 și μ_2, din cadrul cărora a fost extras aleator câte un eșantion de volum n_1 și n_2, având mediile $\bar{x}_1$ și $\bar{x}_2$.

Ca și în cazul mediei aritmetice, vom construi distribuția de sondaj a diferenței mediilor tuturor eșantioanelor de volum *n*, luând în calcul toate perechile posibile de 2 medii determinate la nivel de eșantion. Aceste diferențe ale mediilor eșantioanelor $\bar{x}_1 - \bar{x}_2$ vor fi grupate în funcție de frecvența lor de apariție, fiind obținută în acest mod o distribuție normală, având media $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$ și varianța $\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Deviația standard a acestei distribuții sau eroarea standard a diferenței mediilor eșantioanelor va fi $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Similar cu situația intervalului de încredere pentru estimarea mediei unei populații, $(\bar{x}_1 - \bar{x}_2)$ reprezintă un estimator nedeplasat al diferenței mediilor celor 2 populații $(\mu_1 - \mu_2)$, având varianța $\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Atunci când dispersiile populațiilor (σ_1^2 și σ_2^2) sunt cunoscute, intervalul de încredere pentru $\mu_1 - \mu_2$ garantat cu probabilitatea de $100(1-\alpha)$ este:

$$(\overline{x}_1 - \overline{x}_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Simpla examinare a intervalului de încredere pentru diferența dintre mediile celor 2 populații studiate ne oferă suficientă informație pentru a putea decide dacă acestea pot fi sau nu considerate egale. Atunci când intervalul de încredere NU INCLUDE VALOAREA ZERO, putem afirma faptul că acesta oferă dovezi în sprijinul ideii că mediile celor 2 populații NU sunt egale. În situația contrară, când intervalul conține valoarea zero, concluzionăm că mediile populațiilor analizate pot fi egale.

Determinarea intervalelor de încredere pentru diferența mediilor a 2 populații care NU urmează o distribuție normală se realizează într-o manieră similară în situația când volumele ambelor eșantioane n_1 și n_2 sunt mai mari de 30 de unități (Teorema limită centrală). Dacă variațele celor 2 populații (σ_1^2 și σ_2^2) nu sunt cunoscute, în calcule vom folosi valorile calculate la nivelul fiecărui eșantion (s_1^2 și s_2^2) pentru a le estima.

Similar cu estimarea intervalului de încredere pentru media unei populații, în situația când deviațiile standard ale celor 2 populații (σ_1^2 și σ_2^2) nu sunt cunoscute distribuția de sondaj a diferenței mediilor eșantioanelor urmează o repartiție probabilistică continuă STUDENT (t), care este foarte apropiată de distribuția normală standard. Utilizarea distribuției „t" se bazează pe faptul că distribuțiile celor 2 populații studiate sunt normale. Deosebim 2 situații distincte:

a) VARIANȚELE CELOR DOUĂ POPULAȚII SUNT EGALE

În situația când putem considera că variațele celor 2 populații studiate sunt egale, variațele ajustate, calculate la nivelul fiecărui eșantion, pot fi considerate drept estimări ale variației comune. Pentru a obține un estimator „combinat" al variației comune vom calcula media aritmetică a variațelor celor două eșantioane ponderate cu gradele de libertate ale fiecăruia $(n_1 - 1)$, respectiv $(n_2 - 1)$:

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

Eroarea standard a diferențelor dintre mediile eșantioanelor:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Intervalul de încredere pentru $\mu_1 - \mu_2$ garantat cu probabilitatea de $100(1-\alpha)$ este:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

b) VARIANȚELE CELOR DOUĂ POPULAȚII NU SUNT EGALE

Dacă nu putem considera faptul că cele 2 variante ale populațiilor analizate sunt egale, în calcule vom utiliza factorul de încredere : $t'_{n_1+n_2-2,1-\frac{\alpha}{2}} = \dfrac{\frac{s_1^2}{n_1} \cdot t_{n_1-1,1-\frac{\alpha}{2}} + \frac{s_2^2}{n_1} \cdot t_{n_2-1,1-\frac{\alpha}{2}}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$

Intervalul de încredere pentru $\mu_1 - \mu_2$ garantat cu probabilitatea de $100(1-\alpha)$ este:

$$(\bar{x}_1 - \bar{x}_2) \pm t'_{n_1+n_2-2,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA PROPORȚIEI UNEI POPULAȚII

Alături de estimarea mediei, un loc important în analiza statistică este deținut de studiul proporțiilor (ponderilor) deținute de diferite categorii/grupe în ansamblul populației de referință (proporția persoanelor de sex masculin dintr-o localitate, proporția pacienților care s-au vindecat în urma aplicării unui tratament, ponderea persoanelor imune la o anumită boală dintr-o zonă, etc.). Estimarea proporției unei populații se realizează într-o manieră similară cu cea a mediei, proporția putând fi asimilată unui caz particular al mediei aritmetice, întâlnit în situația unei variabile alternative (binare), care poate lua doar 2 valori (1 – în cazul variantelor variabilei care posedă caracteristica și 0 în rest).

În situația când volumul eșantioanelor este suficient de mare ($n \cdot p \geq 5$ *și* $n \cdot (1-p) \geq 5$) distribuția de sondaj a tuturor proporțiilor eșantioanelor de volum n extrase dintr-o populație de volum N tinde către repartiția normală. Dacă populația generală de volum *N* are o proporție π și o deviație standard $\sigma = \sqrt{\pi \cdot (1-\pi)}$, atunci distribuția de sondaj a proporțiilor tuturor eșantioanelor de volum *n* extrase aleator din aceasta are de asemenea o medie a proporțiilor π și o eroare standard a proporțiilor SEP (deviație standard a distribuției de sondaj a proporțiilor eșantioanelor σ_p) egală cu:

$$SEP = \sigma_p = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

Deoarece parametrul π pe care dorim să-l estimăm este necunoscut, vom utiliza în locul lui ca estimator punctual proporția *p* a eșantionului extras aleator din cadrul populației.
Distribuția de sondaj a proporțiilor eșantioanelor) poate fi convertită într-o distribuție normală standard prin schimbarea de variabilă:

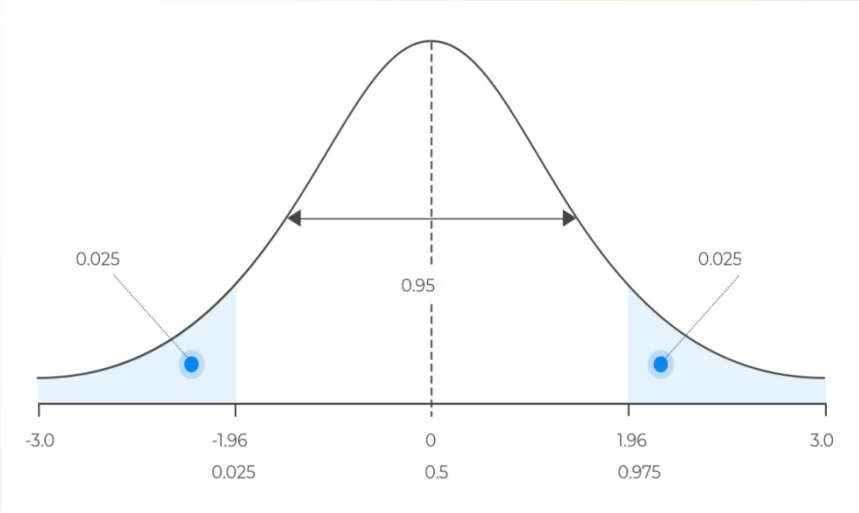
$$z_p = \frac{p-\pi}{\sigma_p} = \frac{p-\pi}{\sigma/\sqrt{n}} = \frac{p-\pi}{\sqrt{\frac{p\cdot(1-p)}{n}}}$$

Intervalul de încredere pentru proporția populației generale (π), garantat cu probabilitatea de $100(1-\alpha)$, se obține aplicând formula generală:

$$P\left(-z_{1-\frac{\alpha}{2}} < z_p < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \qquad -z_{1-\frac{\alpha}{2}} < z_p < z_{1-\frac{\alpha}{2}} \Rightarrow -z_{1-\frac{\alpha}{2}} < \frac{p-\pi}{\sqrt{p\cdot(1-p)/n}} < z_{1-\frac{\alpha}{2}}$$

$$p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}} < \pi < p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}}$$

De obicei, în practică utilizăm un nivel de semnificație $\alpha = 0,05$ (sau 5%) simetric, în care $z_{1-\frac{\alpha}{2}} = 1,96$ reprezintă abscisa corespunzătoare probabilității cumulate: $P = 1-\frac{\alpha}{2} = 1-\frac{0,05}{2} = 1-0,025 = 0,975$



INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA DIFERENȚEI DINTRE PROPORȚIILE A DOUĂ POPULAȚII

Evidențierea magnitudinii diferențelor dintre proporțiile a două populații este în foarte multe cazuri elementul central al unei cercetări. Putem astfel efectua comparații între 2 grupe de vârstă, 2 grupuri socio-economice, 2 grupuri de pacienți diagnosticați în funcție de proporția pe care o posedă o anumită caracteristică de interes în studiu.

Un estimator nedeplasat al diferenței dintre proporțiile a două populații $(\pi_1 - \pi_2)$ este reprezentat diferența dintre proporțiile a două eșantioane extrase în mod aleator din fiecare dintre acestea $(p_1 - p_2)$. După cum a fost prezentat în cursurile anterioare, în situația când volumele eșantioanelor sunt suficient de mari $n_1 \cdot p_1 \geq 5, n_1 \cdot (1 - p_1) \geq 5, n_2 \cdot p_2 \geq 5$ și $n_2 \cdot (1 - p_2) \geq 5$ și proporțiile celor două populații nu sunt foarte aproape de 0 sau 1, se aplică teorema limită centrală, astfel încât putem utiliza și în acest caz distribuția normală pentru a construi intervalele de încredere.

Intervalul de încredere pentru diferența dintre proporțiile a două populații $(\pi_1 - \pi_2)$, garantat cu probabilitatea de $100(1 - \alpha)$:

$$(p_1 - p_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2}}$$