Statistică - Cursul 10

Olariu E. Florentin

Aprilie, 2016

Table of contents

Intervale de încredere pentru media unei populaţii cu dispersia necunoscută

- Să ne amintim că atunci când dispersia populaţiei, σ^2, este cunoscută, un interval de încredere de nivel $(1-\alpha)$ pentru media populaţiei μ, este

$$\left(\overline{x}_n - z^* \frac{\sigma}{\sqrt{n}}, \overline{x}_n + z^* \frac{\sigma}{\sqrt{n}}\right),$$

- Dacă dispersia este necunoscută, atunci putem folosi ca estimator al erorii standard a mediei valoarea $s/\sqrt{n}$.
- Obţinem o nouă statistică

$$T = \frac{\overline{x}_n - \mu}{s/\sqrt{n}}$$

cunoscută drept **statistica t a lui Student**.

- Aceasata deoarece noua statistică urmează o distribuţie Student cu $(n-1)$ *grade de libertate*, $t(n-1)$ - dacă apopulaţia urmează o lege normală.

Intervale de încredere pentru media unei populaţii cu dispersia necunoscută

- Fie $T : t(n-1)$, valoarea critică, t^*, pentru niveleul de încredere $(1-\alpha)$ se alege astfel încât

$$P(-t^* \leqslant T \leqslant t^*) = 1-\alpha, P(T \leqslant -t^*) = \alpha/2 \text{ sau } P(T \geqslant t^*) = \alpha/2.$$

Proposition 0.1

Un interval de încredere cu nivelul de încredere $(1-\alpha)$ pentru pentru media unei populaţii cu dispersia necunoscută este

$$\left(\overline{x}_n - t^* \frac{s}{\sqrt{n}}, \overline{x}_n + t^* \frac{s}{\sqrt{n}}\right).$$

unde t^* este valoarea critică asociată cu $\alpha/2$.

Intervale de încredere pentru media unei populaţii cu dispersia necunoscută - Exemplu

Exemplu.

- Într-un oraş există 10, 000 locuinţe închiriate. O companie locală de imobiliare întreprinde un studiu asupra acestor locuinţe: sunt intervievaţi 250 chiriaşi aleşi la întâmplare. Chiria medie este găsită a fi 568$, iar deviaţia standard a eşantionului este 385$.
- Determinaţi a un interval de încredere de 99% pentru chiria medie a tuturor celor 10, 000 de locuinţe închiriate

Soluţie.

- Datele colectate

$$\overline{x}_{250} = 568, s = 385, \alpha = 0.05, \alpha/2 = 0.025,$$

$$t^* = 1.9695, t^* \frac{s}{\sqrt{n}} = 48.0535$$

- Intervalul de încredere este (616.0535, 519.9465).

Intervale de încredere pentru medie - Exerciţii

I. Pentru un sondaj de opinie se alege un eşantion aleator simplu de 400 de persoane de cel puţin 25 de ani dintr-un anumit oraş din Appalachia. Numărul total de ani de şcolarizare al membrilor eşantionului este 4635, iar deviaţia standard a eşantionului este 4.1 ani. Determinaţi a un interval de încredere de 95% pentru numărul mediu de ani de şcolarizare al tuturor persoanelor de cel puţin 25 de ani din oraş. (Presupunem că perioada de şcolarizare urmează o lege normală.)

II. Primăria unui oraş vrea să cunoască venitul mediu al celor 25000 de familii din oraş. Pentru aceasta angajează un institut de sondare a opiniei care interogheaza 1000 de familii alese aleator. Venitul total al acestor familii este de 62396714$, deviaţia standard a eşantionului fiind 53000$. Determinaţi un interval de încredere de 99% pentru venitul mediu al unei familii din acest oraş. (Presupunem că venitul unei familii urmează o lege normală.)

Intervale de încredere pentru medie - Exerciţii

III. O universitate are 30000 studenţi; în vederea unui sondaj, 900 dintre aceşti studenţi sunt aleşi aleator. Vârsta medie a eşantionului este de 22.3 ani cu o deviaţie standard a eşantionului de 4.5 ani. Determinaţi intervale de încredere de 90% şi 95% pentru vârsta medie a studenţilor acestei universităţi. (Presupunem că vârsta unui student urmează o lege normală.)

Testarea ipotezelor statistice

- Decizii trebuie luate în fiecare zi; unele dintre ele sunt mai semnificative decât altele, dar mecanismul luării unei decizii urmează acelaşi şablon.
- Avem două sau mai multe alternative şi trebuie să alegem una dintre ele pe baza evidenţei/convingerilor/contextului/informaţiilor etc.
- Un test statistic de semnificaţie urmează un acelaşi patern cu excepţia faptului că decizia este luată folosind informaţia statistică.
- Aceasta înseamnă că în timpul acestui proces se vor calcula anumite statistici şi pe baza acestora se vor lua deciziile.
- Primul pas este acela de a identifica o situaţie care are un anumit grad de incertitudine şi să formulă două *ipoteze* legate de ea.

Testarea ipotezelor statistice

Definition 1.1

Testarea ipotezelor statistice *este un proces prin care se ia o decizie între două ipoteze opuse.*

Ipotezele statistice *sunt formulate în aşa fel încât întotdeauna una din ele este falsă iar cealaltă adevărată.*

Una dintre ipoteze este testată sperând că se poate arăta că este puţin probabil ca ea să fie adevărată, ceea ce implică faptul că cealaltă ipoteză este probabil adevărată.

- Cele două ipoteze sunt numite *ipoteza nulă* şi *ipoteza alternativă*.
- Un test statistic încearcă sa dovedească faptul că ipoteza nulă este falsă.

Testarea ipotezelor statistice

Definition 1.2

Ipoteza nulă, H_0, *este ipoteza status-quo-ului în ceea ce priveşte populaţia; formal este o afirmaţie legată de populaţie: spre exemplu că are o anumită medie sau dispersie, sau o anumită distribuţie etc.*

Ipoteza alternativă, H_a, *este ipoteza de cercetare, şi susţine un lucru diferit despre obiectul ipotezei nule.*

- Ipoteza nulă este punctul de plecare al studiului şi, în mod conservator, susţine că "nu există nici o diferenţă" sau că "nimic nu se întâmplă" (are tendinţa de a se opune oricărei schimbări).
- În tr-un anumit fel ipoteza alternativă are rolul de a ataca ipoteza nulă şi observă o diferenţă acolo unde cea nulă nu vede nimic.

Testarea ipotezelor statistice - Exemplul I

- O companie de produse lactate cumpără lapte de la mai mulţi distribuitori. Există un dubiu asupra calităţii laptelui astfel cumpărat.
- Temperaturade îngheţ al laptelui urmează o lege normală cu media $\mu_0 = -0.545^\circ C$ şi deviaţia standard $\sigma = 0.008^\circ C$. Dacă se adaugă apă în lapte temperatura de îngheţ creşte.
- Se măsoară temperatura de îngheţ a laptelui pentru cinci loturi diferite primite de la unul dintre distribuitorişi se determină o temperatură medie de $\overline{x}_5 = -0.538^\circ C$. Este aceasta o dovadă ca distribuitorul adaugă apaă în lapte? Sau este doar o valoare datorată hazardului?

Testarea ipotezelor statistice - Exemplul I

- Ipoteza de lucru, adică ipoteza nulă, este aceea că media populaţiei este $\mu = \mu_0$.
- Ipoteza de cercetare, adică ipoteza alternativă, este că media populaţiei este mai mare: $\mu > \mu_0$.
- Relativ la ipoteza alternativă o întrebare mai formală care se poate formula este: în condiţii obişnuite care sunt şansele ca $\mu > \mu_0$?
- Fie X variabila aleatoare asociată temperaturii de îngheţ a laptelui; $X : N(-0.545, 0.008)$.

Testarea ipotezelor statistice - Exemplul I

- Arunci

$$P(X > -0.538) = P\left(\frac{X-(-0.545)}{0.008/\sqrt{5}} > \frac{(-0.538)-(-0.545)}{0.008/\sqrt{5}}\right) =$$

$$= P\left(\frac{X-(-0.545)}{0.008/\sqrt{5}} > 1.95655948\right).$$

- Această probabilitate este $P(Z > 1.95655948)$, unde $Z : N(0,1)$.
- Pentru o variabilă aleatoare normală standard $P(Z > 1.95655948) \sim 0.0250$. Astfel, cu probabilitate 0.025, putem spune că distribuitorul adaugă apă în lapte.

Testarea ipotezelor statistice

- Dacă, presupunând că ipoteza nulă este adevărată, găsim că o anumită statistică obţinută dintr-un eşantion diferă mult de rezultatul aşteptat, atunci spunem că diferenţa este *semnificativă*.
- În acest caz putem fi tentaţi să *respingem ipoteza nulă*.
- Dacă statistica nu duferă semnificativ de rezultatul aşteptat în condiţiil eipotezei nule, atunci spunem că *am eşuat în respingerea ipotezei nule*.
- *Testarea ipotezelor* este o procedură care ne permite să verificăm dacă anumite statistici diferă semnificativ de valoarea aşteptată în condiţiile ipotezei nule.

Testarea ipotezelor statistice

- Dacă decizia este de a respinge H_0, atunci concluzia testului trebuie să fie: *"Există suficiente dovezi pentru nivelul de semnificaţie α pentru ca ..."* (şi aici urmează afirmaţiile din ipoteza alternativă).
- Dacă decizia este de a nu respinge H_0, atunci concluzia testului trebuie să fie: *"Nu există suficiente dovezi pentru nivelul de semnificaţie α pentru ca ..."* (şi aici urmează afirmaţiile din ipoteza alternativă).

Erori de tipul I şi de tipul II

- Evident că decizia aceasta, luată pe baza unor probabilităţi, poate fi greşită. Acesta este locul unde pot apărea erorile.
- Există două tipuri de erori: primul tip apare atunci când respingem o ipoteză care este adevărată, iar al doilea tip apare atunci când acceptă o ipoteză care este falsă şi ar trebui respinsă.

		validitatea ipotezei H_0	
		adevărată	falsă
Decizia asupra H_0	se respinge	**Eroare de tip I** (fals pozitiv)	Corect (adevărat pozitiv)
	nu se respinge	Corect (adevărat negativ)	**Eroare de tip II** (fals negativ)

Erori de tipul I şi de tipul II

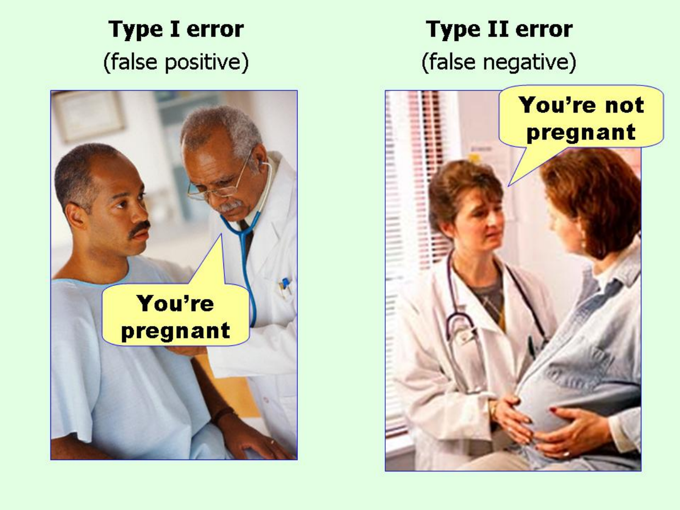

Figure : Erori de tip I şi II.

Testarea ipotezelor statistice - Exemplul II

- Se bănuie că un detergent foarte cunoscut este mai bun decât detergentul vândut sub marca de casă a unui hipermarket şi se doreşte să se testeze calitatea celor două tipuri de detergent, deoarece un cumpărător va dori să cumpere calitate la un preţ convenabil. Formulaţi cele două ipoteze.
- Bănuiala, "Detergentul cunoscut este mai bun decât detergentul de casă al hipermarketului," este motivul pentru care se întreprinde testul şi devine, deci, ipoteză alternativă.

$$H_0: \text{"Nu există nicio diferenţă între cele două tipuri de detergent."}$$

$$H_a: \text{"Detergentul cunoscut este mai bun decât celălalt."}$$

- Testul este conceput în speranţa de a respinge the ipoteza nulă, dar pentru consumator speranţa este de a nu respinde ipoteza nulă (din raţiuni bugetare).

Testarea ipotezelor statistice - Exemplul II

	H_0 is true	H_0 is false
se respinge	**Eroare de tipul I** **Situaţia reală:** Nu există diferenţă. **Concluzie:** Detergentul cunoscut e mai bun. **Acţiune:** Clientul plăteşte mult cu rezultate slabe.	**Decizie corectă** **Situaţia reală:** Detergentul cunoscut e mai bun. **Concluzie:** Detergentul cunoscut e mai bun. **Acţiune:** Clientul plăteşte mult cu rezultate bune.
nu se respinge	**Decizie corectă** **Situaţia reală:** Nu există diferenţă. **Concluzie:** Nu există diferenţă. **Acţiune:** Clientul plăteşte puţin cu rezultate similare.	**Eroare de tipul II** **Situaţia reală:** The brand name is better. **Concluzie:** Nu există diferenţă. **Acţiune:** Clientul plăteşte puţin cu rezultate slabe.

Testarea ipotezelor statistice - Exemplul II

- Am descris cele patru tipuri de rezultate posibile şi acţiunile corespunzătoare pentru un test statistic.
- Situaţia de fapt (reală) nu este cunoscută înainte de a lua decizia, de a trage concluziile şi de a întreprinde acţiunea. Adevărul despre ipoteza nulă, H_0, poate rămâne necunoscut pentru totdeauna.
- Rezultatul unei erori de tipul II este adesea ceea ce se numeşte o "oportunitate pierdută"; se pierde şansa de a utiliza un produs care are o mai bună calitate (în cazul expus).

Dimensiunea efectului, nivelul de semnificaţie şi puterea testului

- Regula după care se ia decizia trebuie sa minimizeze erorile descrise.
- Există patru factori care influenţează o decizie luată pe baza unui test statistic (ignorând metoda de eşantionare): *dimensiunea efectului*, *dimensiunea eşantionului*, *nivelul de semnificaţie* şi *puterea testului*.

Definition 2.1

Dimensiunea efectului *este magnitudinea diferenţei descoperite în eşantionul aleator (dacă există).*

Nivelul de semnificaţie, α, *este probabilitatea (condiţionată) maximă pe care ne-o asumăm drept risc de face o eroare de tipul I.*

Puterea testului, *este* 1 *minus probabilitatea de a face o eroare de tipul II.*

Dimensiunea efectului, nivelul de semnificaţie şi puterea testului

- Nivelul de semnificaţie este specificat de obicei înaintea eşantionării, astfel rezultatele testului nu vor influenţa eşantionarea.
- De obicei nivelul de semnificaţie este 0.05 sau 0.01; de exemplu 0.05 (sau 5% nivel de semnificaţie) este utilizat pentru o regula de luare a deciziei care dă cel mult 5 şanse din 100 de a respinge ipoteza nulă când ea este adevărată.
- Puterea testului este probabilitatea ca testul să detecteze o diferenţă atunci când există cu adevărat o asemenea diferenţă de detectat.
- Dacă puterea testului este mare, atunci probabilitatea de a face o eroare de tipul II (probabilitatea ca testul să detecteze o diferenţă atunci când nu există vreuna) este scăzută.

Nivelul de semnificaţie şi valoarea P

- Există două moduri de a desfăşura un test de semnificaţie: folosind *scorul* şi *valoarea critică* sau folosind *valoarea P*.

Definition 3.1

Valoarea P *este probabilitatea de a obţine un rezultat cel puţin la fel de neobişnuit (extrem) ca rezultatul obţinut din eşantion. Când valoarea P este "mică" putem respinge H_0.*

- Revenim la primul exemplu. Am calculat $P(X > -0.538) \sim 0.0250$; aceasta este valoarea P a testului.
- Dacă 2.5% este considerată o probabilitatea mică, atunci putem respinge ipoteza nulă.

Nivelul de semnificaţie şi valoarea P

- Dacă însă 2.5% nu este de ajuns de mică, atunci nu respingem H_0.
- De exemplu $2.5\% < 5\%$, astfel, cu 5% nivel de semnificaţie H_0 este respinsă şi este acceptată H_a.
- Cu 1% nivel de semnificaţie încercarea de a respinge ipoteza nulă eşuează, datele noastre nu sunt suficient de semnificative pentru a accepta ipoteza alternativă, H_a.

Nivelul de semnificaţie şi valoarea P

Definition 3.2
Scorul *testului este statistica ce corespunde valorii P;* **valoarea critică**, *pe de altă parte, coresponde nivelului de semnificaţie. Concluzia testului depinde de rezultatul comparării celor două valori.*

- Calculăm statistica
$$z = \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} = \frac{(-0.538) - (-0.545)}{0.008/\sqrt{5}} = 1.95655948.$$
- z este scorul testului.
- Valoarea critică, z^*, pentru 1% nivel de semnificaţie îndeplineşte condiţia
$$P(Z > z^*) = 1 - \alpha = 0.99 \Rightarrow z^* = 2.32, \text{ unde } Z : N(0,1).$$

Nivelul de semnificaţie şi valoarea P

- Cu 1% nivel de semnificaţie $z < z^*$, deci H_0 nu poate fi respinsă
- Pentru 5% nivel de semnificaţie, valoarea critică, z^*, se calculează astfel ca

$$P(Z > z^*) = 1 - \alpha = 0.95 \Rightarrow z^* = 1.64,$$

- Cu 5% nivel de semnificaţie $z < z^*$, deci H_0 poate fi respinsă; acceptăm H_a.
- Acest exemplu arată că pentru diferite nivele de semnificaţie concluziile testului pot fi diferite.

Teste parametrice şi neparametrice

- În exemplul anterior am presupus că populaţia urmează o distribuţie normală.
- Câteodată este dificil de aflat dacă populaţia urmează o distribuţie aume (normală cel mai adesea).
- Relativ la distribuţia populaţiei avem două tipuri de teste de semnificaţie: parametrice şi neparametrice.

Definition 4.1

Un **test parametric** *presupune că populaţia urmează o anumită distribuţie şi infereează asupra parametrilor acelei distribuţii.*

Teste parametrice şi neparametrice

- Testele neparametrice inferează mai degrabă asupra distribuţiei decât asupra parametrilor populaţei.
- De obicei testele parametrice presupun că populaţia urmează o distribuţie normală. (Unii autori consideră că orice altfel de test este neparametric).

Definition 4.2

Un **test neparametric** *numit şi* **distribution-free** *sau* **parameter-free** *se bazează pe puţine fapte - de obicei distribuţia şi parametrii săi (medie, dispersie) nu sunt cunoscuţi.*

Testul proporţiilor

- Una dintre cele mai utilizate inferenţe este cea asupra *parametrului binomial p*, probabilitatea succesului.
- În multe situaţii ne interesează doar ceea ce se "realizează" sau "nu se realizează"; există doar două rezultate posibile - aceasta este condiţia de bază a unui experiment binomial.
- Proporţia indivizilor dintr-o populaţie care au o anumită trăsătură poate fi considerată probabilitatea succesului.
- Exemple: proporţia indivizilor care fumează, a cetăţenilor care votează cu un anumit politician, a şoferilor care accelerează la lumina galbenă a semaforului, a tinerilor care îşi întemeiază un cămin etc.

Testul proporţiilor

- Dacă $X : B(n,p)$ (numărul de succese), atunci parametrii populaţiei sunt
$$\mu = np, \sigma^2 = np(1-p).$$
- Dacă dimensiune eşantionului este n şi x este numărul de succese din eşantion, atunci **frecvenţa** sau **probabilitatea binomială a eşantionului** este
$$p' = \frac{x}{n}.$$
- Pentru $n \geqslant 20$ şi $np \geqslant 5$, p' poate fi aproximată cu o distribuţie normală.

Testul proporţiilor

- Parametrii lui p' sunt

$$M[p'] = \frac{M[X]}{n} = p, D^2[p'] = \frac{D^2[X]}{n^2} = \frac{p((1-p)}{n}.$$

- Deci

$$p' \sim N\left(p, \frac{p((1-p)}{n}\right).$$

- Astfel următoarea statistică urmează o distribuţie normală

$$z = \frac{p'-p}{\sqrt{p(1-p)/n}}.$$

- Aceasta va fi statistica sau scorul testului (unde $p' = x/n$).

Testul proporţiilor - Exemplu

- Conform unui sondaj Harris din din august 2008, 68% dintre americanii adulţi au un abonament la o bibliotecă. Să presupunem că se alege un eşantion aleator de 1000 de adulţi pentru a testa $H_0 : p = 0.68$ versus $H_a : p < 0.68$, unde $p$reprezintă proporţia of adulţilor care au intr-adevăr abonament.
- 651 din cei 1000 de indivizi au un abonament la o bibliotecă. Folosiţi $\alpha = 0.01$.

  a. Calculaţi valoarea statisticii (scorul) testului.

  b. Aplicaţi testul folosind valoarea P.

  c. Aplicaţi testul folosind abordarea clasică (cu valoarea critică).

Testul proporţiilor - Exemplu

$$z = \frac{p' - p}{\sqrt{p(1-p)/n}} = \frac{0.651 - 0.68}{\sqrt{0.68 * 0.32/1000}} = -1.965$$

- Cu valoarea P:

$$P = P(Z < z|H_0) = P(Z < -2.60) = 0.0246$$

- Cu $\alpha = 1\%$ nu putem respinge H_0, datele nu sunt semnificative pentru aceasta.
- Cu valoarea critică:

$$z^* = -2.326, (P(Z < z^*) = 0.01).$$

- Deoarece $z \nless z^*$ nu putem respinge H_0, datele nu sunt suficient de semnificative pentru a accepta ipoteza alternativă.

Testul proporţiilor - Exemplu

- Reluăm acum exerciţiul cu $\alpha = 5\%$.
- Cu P-value: $P = 0.0246 < 0.05$, deci putem respinge ipoteza nulă şi să acceptăm H_a, adică faptul că proporţia adulţilor care au abonament la bibliotecă este mai mic decât 0.68.
- Cu valoarea critică:$z^* = -1.644 > z$ cea ce înseamnă că H_0 poate fi respinsă.

Testul proporţiilor - Exerciţii

I. O monedă este aruncată de 10000 de ori şi de 5167 ori apare stema. Probabilitatea de a apărea stema este 0.5? Sau dimpotrivă, este semnificativ numărul mare de apariţii ale stemei? Aplicaţi un test statistic corespunzător pentru a răspunde acestor întrebări.

II. Rezolvaţi din nou exerciţiul I când stema apare de 5067 ori.

III. În 2009, într-un articol din USA Today se afirma că 58% dintre adulţii americani accelerează la lumina galbenă a semaforului. Într-un oraş se desfăşoară un sondaj cu 150 de adulţi aleşi aleator; se află că 71 dintre aceştia recunosc că accelerează la lumina galbenă a semaforului. Se poate trage concluzia că în acest oraş există o rată mai mică decât aceea de la nivel naţional? Folosiţi 0.05 nivel de semnificaţie.

Teste unilaterale şi bilaterale

- Există trei tipuri de ipoteze alternative relativ la probabilitatea binomială p.

$$H_a : p < p_0, H_a : p > p_0 \text{ sau } H_a : p \neq p_0.$$

- Primele două se numesc ipoteze alternative **unilaterale** sau (**"one-tailed"**), iar al treia este ipoteză alternativă **bilaterală** sau **"two-tailed"**).
- Pentru fiecare din aceste cazuri valoarea P şi valoarea critică se calculează în mod diferit.

Testul unilateral la proporţilor

- Pentru $H_0 : p = p_0$ and $H_a : p > p_0$ - *unilateral la dreapta - right tail*.

$$\text{valoarea } P : P(Z > z),$$

$$\text{valoarea critică } z^* > 0, \text{ a. î. } P(Z > z^*) = \alpha = 1 - P(Z < -z^*).$$

$$\text{în R } z^* = qnorm(1 - \alpha).$$

- Pentru $H_0 : p = p_0$ and $H_a : p < p_0$ - *unilateral la stânga - left tail*.

$$\text{valoarea } P : P(Z < z),$$

$$\text{valoarea critică } z^* < 0, \text{ a. î. } P(Z < z^*) = \alpha = 1 - P(Z > -z^*).$$

$$\text{în R } z^* = qnorm(\alpha).$$

Testul bilateral la proporţilor

- Pentru $H_0 : p = p_0$ and $H_a : p \neq p_0$.

$$\text{valoarea } P : P(Z > |z|) + P(Z < -|z|),$$

$$\text{valoarea critică } z^* > 0, \text{ a. î. } P(Z < -|z^*|) = \alpha/2 =$$

$$= 1 - P(Z > |z^*|).$$

$$\text{în R } z^* = -qnorm(\alpha/2) = qnorm(1 - \alpha/2).$$

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.
- http://effectsizefaq.com/2010/05/31/i-always-get-confused-about-type-i-and-ii-