Probabilităţi şi Statistică - Curs 11

Mai 2018

- Statistica inferenţială are scopul să trage concluzii relativ la o populaţie folosind rezultate ale teoriei probabilităţilor şi statistici obţinute din eşantionane.
- Fără utilizarea teoriei probabilităţilor (legea numerelor mari, teorema limită centrală ş. a.) am putea considera drept sistematic un comportament care este de fapt datorat hazardului sau, din contră, un comportament sistematic ar putea trece neobservat.
- Exemple de inferenţe statistice: *intervale de încredere* pentru estimarea parametrilor sau *teste de semnificaţie*.
- Tehnicile de inferenţă au la bază distribuţiile eşantioanelor.

- Distribuţia unei anumite populaţii poate fi necunoscută în întregime; de aceea ne putem dori să aflăm cel puţin media, dispersia sau alţi parametri ai ei.
- Aceşti parametri ai unei populaţii pot fi estimaţi folosind statistici calculate din eşantion.
- Există două tipuri de estimare: *estimare punctuală* şi *estimare cu intervale* .

Proposition 1

Estimarea punctuală a unui parametru *constă în determinarea unui număr, de obicei valoarea unei statistici corespunzătoare, desemnat sa estimeze acel parametru.*

Proposition 2

Estimarea cu un interval a unui parametru *constă în determinarea unui interval ale cărui limite sunt statistici calculate din eşantioane.*

Nivelul de încredere $(1-\alpha)$ *este proporţia acelor intervale (de încredere) care conţin parametrul estimat.*

Un **interval de încredere** *este un interval care are un anumit nivelul de încredere (prescris).*

- Exemple de estimatori punctuali sunt media de selecţie $\overline{x}_n$, dispersia eşantionului s^2, sau deviaţia standard a eşantionului s.
- Pentru un parametru dat putem avea mai mulţi estimatori punctuali: media populaţiee poate fi estimată prin media de selecţie, mediană, mod.
- Se pot ridica anumite întrebări legate de calitatea estimatorilor punctuali.
- Cât de exact este un estimator (i.e., *încrederea*) - este în mod frecvent mai mare (*supra-estimator*) sau mai mic (*sub-estimator*) în raport cu parametrul estimat?
- Din acest punct de vedere se preferă estimatorii *nedeplasaţi*.
- Care este variabilitatea unui estimator punctual (văzut ca o variabilă aleatoare) - aceasta este *acurateţea*.

- De exemplu dispersia mediei de selecţie, care este σ^2/n: cu cât este mai mare eşantionul cu atât mai mică este variabilitatea mediei de selecţie.

Definition 1

Fie θ un anumit parametru al unei populaţii, x un eşantion al acestei populaţii şi $\hat{\theta}_n = \hat{\theta}_n(x)$ un estimator punctual al lui θ.

$\hat{\theta}_n$ este o **statistică nedeplasată** *dacă $\mathbb{E}[\hat{\theta}_n] = \theta$. Altfel $\hat{\theta}_n$ este numită statistică* **deplasată**.

$\hat{\theta}_n$ este o **statistică consistentă** *dacă, pentru orice $\epsilon > 0$*

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$

Un estimator nedeplasat cu dispersie minimă, dacă există, se numeşte **statistică eficientă**.

- Fie $\mathcal{P}$ o populaţie formată din N indivizi ale căror valori ale atributului sunt $a_1, a_2, \ldots, a_N$.
- Variabila care reprezintă atributul populaţiei este notată cu X.
- Pentru această populaţie, deci şi pentru X, media şi dispersia sunt

$$\frac{1}{N}\sum_{i=1}^{N} a_i = \mu, \text{ respectiv } \frac{1}{N}\sum_{i=1}^{N}(a_i - \mu)^2 = \sigma^2.$$

- Pentru a estima μ şi σ^2folosim eşantioane de dimensiune n.

- Pentru un eşantion dat, $x^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right)$, media de selecţie este

$$\overline{x}_n^{(k)} = \frac{1}{n}\sum_{j=1}^{n} x_j^{(k)}.$$

- Fiecare $x_j^{(k)}$ este o valoare a unei variabile aleatoare cu aceeaşi distribuţie ca X.

Definition 2

Variabila aleatoare care are drept valori toate mediile de selecţie posibile, $\overline{x}_n^{(k)}$, pentru eşantioane de dimensiune n, se numeşte **distribuţia mediei de selecţie**.

- Se poate demonstra următorul rezultat

Theorem 1.1

Media şi dispersia mediei de selecţie, $\overline{x}_n$, sunt μ şi σ^2/n:

$$\mathbb{E}[\overline{x}_n] = \mu, D^2[\overline{x}_n] = \frac{\sigma^2}{n}.$$

În plus, pentru valori mari ale lui n ($\geqslant 30$), distribuţia mediei de selecţie este normală, i. e.

$$\overline{x}_n \sim N(\mu, \sigma^2/n).$$

- Bineînţeles, un estimator al mediei adevărate a populaţiei (μ) este media de selecţie, $\overline{x}_n$.
- Un estimator pentru adevărata dispersie a populaţiei σ^2) este dispersia eşantionului s^2 - o statistică nedeplasată.
- Pentru deviaţia standard a mediei de selecţie, $\sigma/\sqrt{n}$, un estimator este $s/\sqrt{n}$ numită **eroarea standard a mediei** (the standard error of the mean - SEM).

- Un estimator de tip interval al mediei populaţiei poate fi obţinut cu inegalitatea lui Cebâşev:

$$P(|\overline{x}_n - \mathbb{E}[\overline{x}_n]| \geqslant k \cdot D[\overline{x}_n]) \leqslant \frac{1}{k^2} \Leftrightarrow$$

$$P\left(|\overline{x}_n - \mu| \geqslant k \cdot \frac{\sigma}{\sqrt{n}}\right) \leqslant \frac{1}{k^2} \Leftrightarrow$$

$$P\left(|\overline{x}_n - \mu| < k \cdot \frac{\sigma}{\sqrt{n}}\right) \geqslant 1 - \frac{1}{k^2} \Leftrightarrow$$

$$P\left(\overline{x}_n - k \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + k \cdot \frac{\sigma}{\sqrt{n}}\right) \geqslant 1 - \frac{1}{k^2}.$$

- Astfel o estimare cu interval a lui μ este

$$\left(\overline{x}_n - k \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + k \cdot \frac{\sigma}{\sqrt{n}}\right).$$

Definition 3

Un interval de încredere pentru un parametru θ cu $(1-\alpha)$ nivel de încredere este definit cu ajutorul a două statistici L şi U astfel ca

$$P(L \leqslant \theta \leqslant U) = (\geqslant) 1 - \alpha.$$

- L şi U sunt de fapt variabile aleatoare ale căror valori sunt statisticia: pentru diferite eşantioane au diferite valori.
- De obicei nivelul de încredere este o probabilitate aproape de 1, cum ar fi 0.90, 0.95,sau 0.99 (care dau $\alpha \in \{0.10, 0.05, 0.01\}$).

- Să presupunem că avem un eşantion de dimensiune n şi un nivel de încredere $(1-\alpha)$ şi dorim să construim un interval de încredere pentru media μ.
- Ştim că media de selecţie, $\overline{x}_n$, urmează o distribuţie normală

$$N\left(\mu, \frac{\sigma^2}{n}\right).$$

- Prin standardizare variabila $Z = \dfrac{\overline{x}_n - \mu}{\sigma/\sqrt{n}}$ este distribuită $N(0,1)$.
- Căutăm o valoare z^*, numită *valoare critică*, astfel ca o variabilă normal standard să acopere sub grafic o arie de $(1-\alpha)$ pe intervalul centrat în medie şi de lungime $2z^*$ deviaţii standard.

- Fie $Z : N(0, 1)$, valoarea critică este aleasă aşa încât

$$P(-z^* \leqslant Z \leqslant z^*) = 1 - \alpha.$$

- Definiţii echivalente ale lui z^*:

$$P(Z \leqslant -z^*) = \alpha/2 \text{ sau } P(Z \geqslant z^*) = \alpha/2.$$

- Notâm cu $\Phi(a) = P(Z \leqslant a)$, funcţia de repartiţie normală standard.
- Astfel $z^* = -\Phi^{-1}(\alpha/2)$ - valoare care poate fi găsită în tabele sau aproximată în pachetele de prelucrare statistică uzuale (R, MiniTab, SPSS etc).
- Odată ce am determinat valoarea critică, ştim că

$$P\left(-z^* \leqslant \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} \leqslant z^*\right) = 1 - \alpha.$$

- Echiuivalent

$$P\left(\overline{x}_n - z^* \frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{x}_n + z^* \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- Am demonstrat astfel

Theorem 1.2

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

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

unde z^* este valoarea critică asociată cu $\alpha/2$. Mai mult, acest interval este exact pentru o populaţie distribuită normal şi aproximativ altfel, când eşantionul este suficient de mare ($n \geqslant 30$).

Intervale de încredere pentru media unei populaţii cu dispersia cunoscută

Nivel de încredere	$\alpha/2$	z^*
90%	0.05	1.645
95%	0.025	1.960
99%	0.005	2.576

- $z^* \dfrac{\sigma}{\sqrt{n}}$ se numeşte *eroarea marginală*.
- Lungimea unui interval de încredere pentru medie este $2z^* \dfrac{\sigma}{\sqrt{n}}$.
- Dacă dorim o lungime anume pentru acest interval, w, atunci ne trebuie un eşantion de dimensiune $n = \dfrac{(2z^*\sigma)^2}{w^2}$. În practică această valoare poate fi nerealistă (dacă n este prea mare).
- Pe măsură ce n creşte lungimea intervalului (sau eroarea marginală) scade, ceea ce este util, dar poate fi nepractic.

- Să ne amintim că $\frac{\sigma}{\sqrt{n}}$ este deviaţi standard a mediei de selecţie. Un intervale de încredere poate fi văzut astfel

$$estimatedMean \pm z^* estimatedStdDev$$

- *Exemplu.* Un anumit medicament este analizat măsurându-i-se substanţa activă de trei ori, rezultatele sunt 0.8403, 0.8636, şi 0.8447 g/l. Se ştie că această concentraţie urmează urmează o lege normală cu deviaţia standard $\sigma = 0.0068$ g/l. Să se determine un interval de încredere de 99% pentru adevărata medie a concentraţiei, μ.

  Soluţie:

  $$\overline{x}_3 = 0.8495, \alpha = 0.01, \alpha/2 = 0.005, z^* = 2.576, z^* \frac{\sigma}{\sqrt{n}} = 0.0101$$

  Intervalul de încredere este $(0.8394, 0.8596)$.

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

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

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

Proposition 3

Un interval de încredere cu nivelul de încredere $(1-\alpha)$ pentru pentru media unei populaţii normale 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 $\alpha/2$.

Exemplu.

- Într-un oraş există 10000 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 95% pentru chiria medie a tuturor celor 10000 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 $(519.9465, 616.0535)$.

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.)

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.)

- 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/informa 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 formulăm două *ipoteze* legate de ea.

Definition 4

Testarea ipotezelor statistice *este un proces prin care se ia o decizie (se alege) 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.

Definition 5

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).
- Într-un anumit fel ipoteza alternativă are rolul de a ataca ipoteza nulă şi observă o diferenţă acolo unde cea nulă nu găseşte nimic.

- 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 apă în lapte? Sau este doar o valoare datorată hazardului?

- 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 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; $X : N(-0.545, 0.008)$.

- 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 $P(Z > 1.95655948)$ 0.0250. Astfel, cu probabilitate 0.025, putem spune că distribuitorul adaugă apă în lapte.

- 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 diferă semnificativ de rezultatul aşteptat în condiţiile ipotezei 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.

- 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).

- 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)

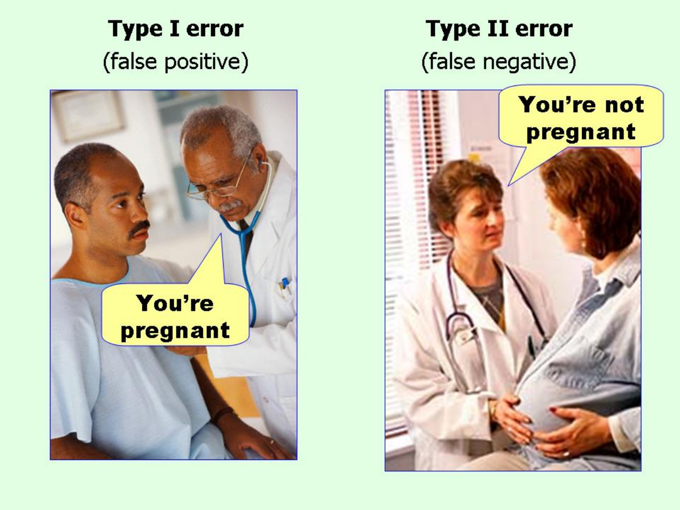

Figure: Erori de tip I şi II.

- 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 : "Nu există nicio diferenţă între cele două tipuri de detergent

H_a : "Detergentul cunoscut este mai bun decât celălalt."

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

	H_0 este adevărată	H_0 este falsă
se respinge	**Eroare de tipul I** **Situaţia reală:** Nu există diferenţă. **Concluzie:** Detergentul cunoscut e mai bun. **Acţiune:** Clientul plăteşte mult cu aceleaşi rezultate.	**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:** Detergentul cunoscut e mai bun. **Concluzie:** Nu există diferenţă. **Acţiune:** Clientul plăteşte puţin cu rezultate slabe.

- 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).

- Regula după care se ia decizia trebuie sa minimizeze erorile descrise.
- Exisă 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 6

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.*

- 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.

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

Definition 7

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 probabilitate mică, atunci putem respinge ipoteza nulă.

- 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.

Definition 8
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

$$P(Z > z^*) = 1 - \alpha = 0.99 \Rightarrow z^* = 2.32, \text{ unde } Z : N(0,1).$$

- 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.

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

Definition 9

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

- 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 10

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.*

- 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 o familie etc.

- Dacă $X : B(n, p)$ (numărul de succese), atunci parametrii populaţiei sunt

$$\mu = np, \sigma^2 = np(1-p).$$

- Dacă dimensiunea 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.

- Parametrii lui p' sunt

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

- Deci

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

- Astfel următoarea statistică urmează o distribuţie normală standard

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

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

- 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 adulţilor care au într-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).

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

- 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 $z \not< z^*$ nu putem respinge H_0, datele nu sunt suficient de semnificative pentru a accepta ipoteza alternativă.

- 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.

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.

- 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 a treia este ipoteză alternativă **bilaterală** sau **"two-tailed"**).
- Pentru fiecare din aceste cazuri valoarea critică se calculează în mod diferit.

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

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

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

$$\text{în R } z^* = qnorm(\alpha).$$

- Pentru $H_0 : p = p_0$ and $H_a : p \neq p_0$ - *bilateral sau simetric - two-tailed*.

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

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

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

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

Sfârşit

- 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-errors-can-you-show-me-somethin