# Probabilitate și statistică în Data Science

## Partea IV. Statistică I

## Ce ne așteaptă?

1. Ce este statistica
2. Eșantionarea
3. Teorema limitei centrale
4. Eroarea standard
5. Testul z
6. Erori de tip 1 și de tip 2

# Notații

$n$ – ***dimensiunea esantionului***

$\mu$ – ***valorea medie a populatiei***

$\sigma$ – ***abatera standard a populatiei***

$\bar{x}$ – ***valoarea medie a esantionului***

$p$ – ***parametrul populatiei***

$\hat{p}$ – ***statistica esantionului***

$SE_{\bar{x}}$ – ***eroarea standard a valorii medii a esantionului***

$H_0$ – ***ipoteza nula***

$H_1$ – ***ipoteza alternativa***

$\alpha$ – ***pragul de semnificatie***

$Z$ – ***statistica testului z***

$z$ – ***valoarea critica a testului z***

$P$ – ***valoarea percentilei corespunzatoare statisticii testului z***

# 1. Ce este statistica

## Noțiune de statistică

- **Statistica – aplicarea a ceea ce se cunoaște asupra a ceea ce se dorește a se cunoaște**

- **Exemple de statistică:**
  - *Utilizarea sondajelor la ieșirea de la urnele de vot pentru a reflecta rezultatele alegerilor întregii populații*
  - *Analiza nivelului de trai a populației unei regiuni pentru a descrie nivelul de trai a populației întregii țări*

# Termeni ai statisticii

- Populația – oricare membru al unui grup care se dorește a fi studiat
- Eșantion – un set mai mic al membrilor populație selectați aliator
- Parametru – o caracteristică a populației ce se dorește a fi studiată
- Statistici – caracteristici ale eșantionului care pot fi determinată
- Interferențe statistice – instrumente aplicate asupra eșantionului în încercarea de a descrie populația
- Variabila – o caracteristică ce descrie un membru al eșantionului

# 2. Eșantionarea

## Erori de eșantionare

- **Principalul avantaj al unui model statistic constă în faptul că concluziile obținute pe un eșantion de dimensiuni rezonabile pot fi extrapolate la întreaga populație.**
- **Provocarea constă în selectarea eșantionului ca acesta să fie cât mai reprezentativ, evitându-se erorile**
- **La selectarea eșantionului se pot produce următoarele erori:**
  - **Erori de selecție:**
    - **Erori de neacoperire**
    - **Erori de auto-selecție**
    - **Erori a “utilizatorilor-sănătoși”**
  - **Erori de supraviețuire**

# Erori de neacoperire

- **Erorile de neacoperire - sunt erori în selectarea eșantionului și apar atunci când se realizează puține observații și se omite un întreg segment al populației**

- **Exemplu:**

  *Realizarea unui sondaj printre lucrătorii unui spital și omiterea personalului care lucrează în tura de noapte*

# Erori de auto-selecție

- **Erorile de auto-selecție - sunt erori în selectarea eșantionului și apar atunci când se realizează observații doar printre subiecții interesați de problema abordată și se omite părerea segmentului celor neinteresați**

- **Exemplu:**

  *Realizarea unui sondaj online referitor la o echipa de fotbal va include rezultatele doar a celor interesați de echipa respectiva și nu va include rezultatele celor ce nu sunt suporteri ai acesteia*

# Erori a “utilizatorilor sănătoși”

- **Erorile “utilizatorilor sănătoși” - sunt erori în selectarea eșantionului și apar atunci când se realizează observații doar pe un segment al populației cu implicații directe sau indirecte în rezultatele sondajului**

- **Exemplu:**

  *Realizarea unui sondaj printre vânzătorii de frunte și legume referitor la importanțe dietei și a modului sănătos de alimentare*

# Erori de supraviețuire

- **Erorile de supraviețuire – se datorează îmbunătățirii în timp a populației prin dispariția unor membri ce nu au supraviețuit încercărilor problemei analizate**

- **Exemplu:**

  *Pe timp de război s-au schimbat coifurile din plastic a soldaților în coifuri de fier. La realizarea unui sondaj referitor la numărul de răniți rezultatele au fost mai proaste comparativ cu perioada coifurilor din plastic. Acest fapt se datorează faptului că mulți dintre soldații cu coifuri de plastic mureau și nu ajungeau în statisticile răniților.*

  *Pe timp de război s-a analizat care este zona cea mai sensibilă a avioanelor pentru a fi protejată suplimentar. Rezultatele au fost eronate deoarece avioanele care erau lovite în zona cea mai sensibilă se prăbușeau și nu erau analizate.*

# Eșantionarea aliatoare

- În funcție de modalitatea de selectare a eșantionului, eșantionarea poate fi aliatoare, aliatoare stratificată sau clusterizată

- Eșantionarea aliatoare presupune faptul că orice membru al grupului are șanse egale de a fi selectat în cadrul eșantionului

- Neajunsul eșantionării aliatoare – întrucât eșantioanele sunt mult mai mica ca număr decât populație, exista probabilitate ca întregi segmente demografice să fie omise

- Exemplu: la selectarea aliatoare a unui eșantion relativ mic pentru un sondaj public există probabilitate de a fi omise persoanele în vârstă sau persoanele foarte tinere

# Eșantionarea aliatoare stratificată

- **Eșantionarea aliatoare stratificată asigura reprezentarea adecvată pentru toate segmentele populație**
- **Inițial populația se divizează în segmente pe baza unor caracteristici, iar un membru al populație poate fi inclus doar într-un singur segment**
- **Apoi se selectează aliator eșantioane din fiecare segment de dimensiuni proporționale dimensiunilor segmentului**
- **Exemplu: Sondaj referitor la satisfacție angajaților realizat pe 10% dintre angajați unei întreprinderi cu 1000 de lucrători.**

*Initial se divizează angajații pe categorii de vârstă și apoi se selectează 10% din fiecare categorie*

| 20-29 | 30-39 | 40-49 | 50+ | TOTAL |
|---|---|---|---|---|
| 1400 | 4450 | 3200 | 950 | 10,000 |
| **140** | **445** | **320** | **95** | **1,000** |

# Eșantionarea clusterizată

- Eșantionarea clusterizată presupune faptul că populația este divizată în anumite grupuri numite clustere, iar selecția eșantioanelor se realizează pe un singur cluster selectat aliator
- Asigură o precizie mai mică dar presupune costuri reduse
- Exemplu: Un supermarket a ales să distribuie materiale publicitare doar locuitorilor din vecinătatea acestuia și nu populație întregului oraș

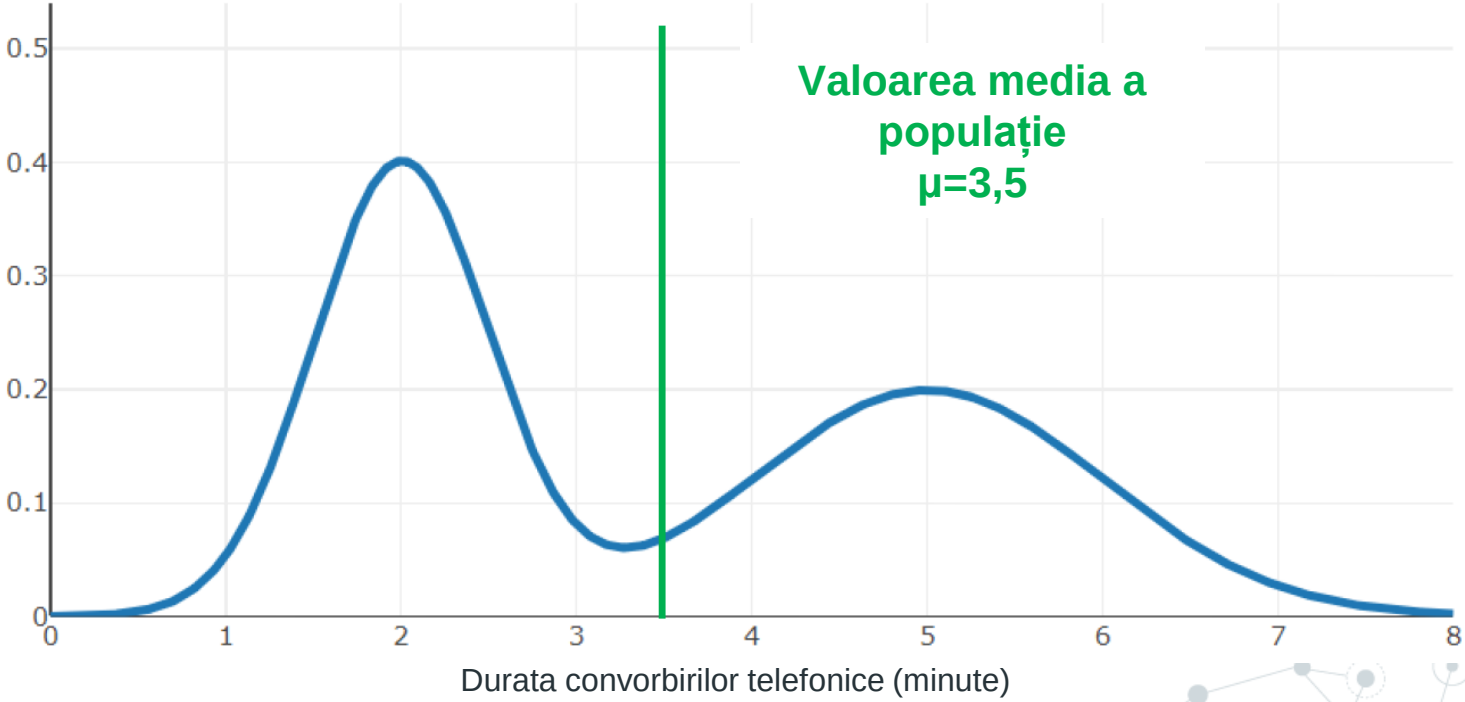
# 3. Teorema limitei centrale

## Esența teoremei

- Valoarea medie a eșantionului de cele mai multe ori diferă de valoarea medie a populație
- Teorema limitei centrale consideră un grup din mai multe eșantioane, adică crearea mai multor eșantioane în cadrul aceleiași populații
- Teorema limitei centrale: Valorile medii ale unui grup de eșantioane ale populației vor fi distribuite normal valorii medii a populație chiar dacă populația nu are o distribuție normală.
- Adică, 95% din valorile medii ale eșantioanelor se vor include în banda 2σ de la valoarea medie a populație

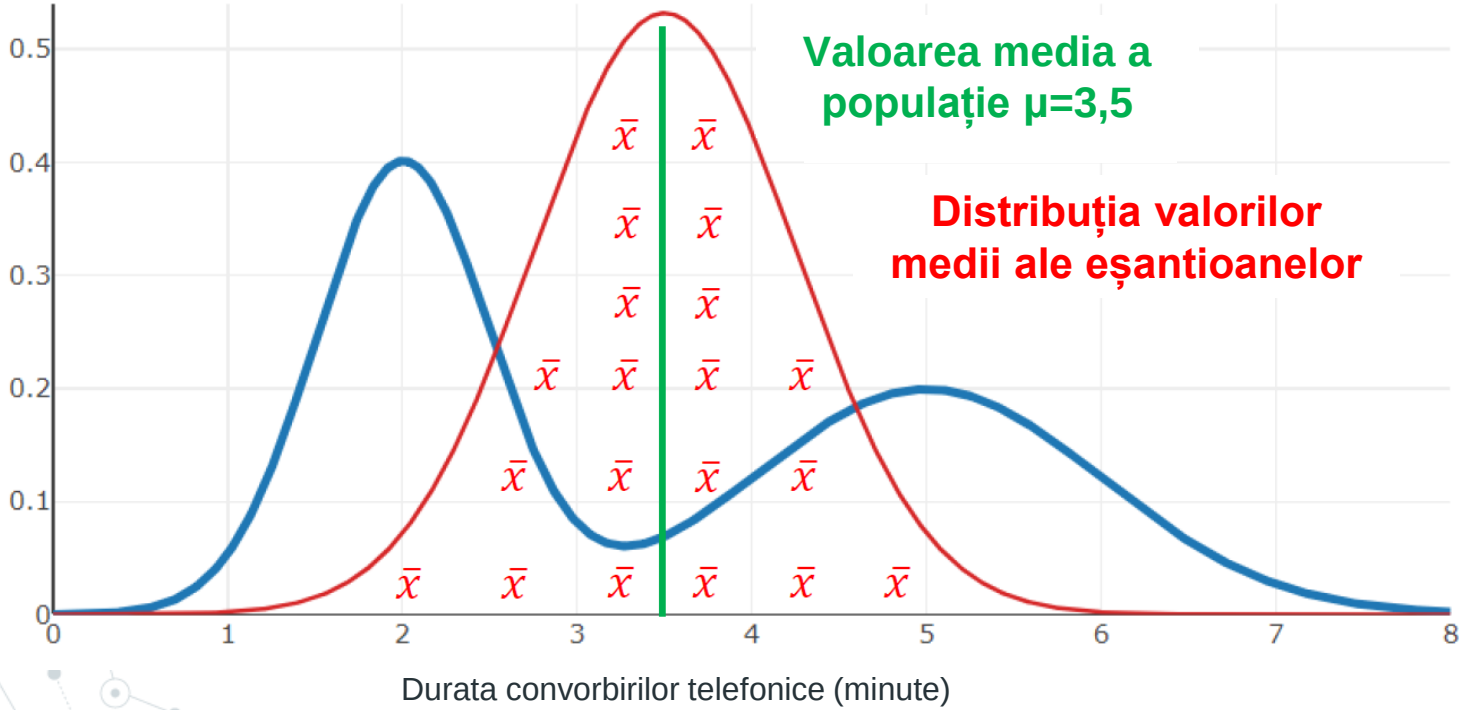
# Exemplu grafic (1)

- **Distribuția duratei convorbirilor telefonice**



# Exemplu grafic (2)

- **Distribuția valorilor medii ale eșantioanelor**



# 4. Eroarea standard

## Noțiunea erorii standard

- Eroarea standard descrie cât de departe de valoarea medie a populație se află valoarea medie eșantionului
- Deosebirea dintre abaterea standard a populație și eroarea standard constă în faptul că prima descrie distanța dintre media populației și valorile individuale, iar a doua distanța dintre media populație și media eșantioanelor
- Dacă este cunoscută valoarea abaterii standard a populație σ atunci valoarea erorii standard se determină cu relația

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# Exemplu eroarea standard

- **Un test a fost conceput pentru a avea un scor mediu de 100 puncte cu abaterea standard de 15 puncte. Dacă un eșantion de 10 scoruri are o medie de 104 puncte se poate considera că acestea provin din populația generală**

$$n = 10 \quad \bar{x} = 104 \quad \sigma = 15 \quad \mu = 100$$

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4{,}743$$

Gama de valori in care se vor încadra 68% dintre mediile eșantioanelor

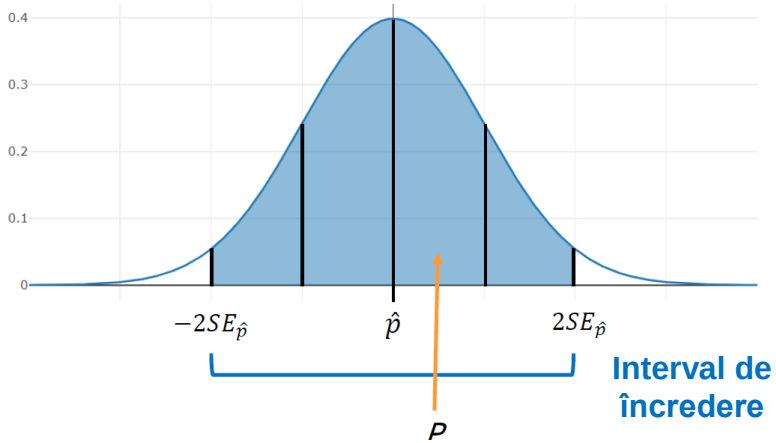
$$B = \mu - SE_{\bar{x}} \dots \mu + SE_{\bar{x}}$$

$$B = 100 - 4{,}743 \dots 100 + 4{,}743 = 95{,}257 \dots 104{,}743$$

Valoarea medie a 10 scoruri $\bar{x} = 104$ se încadrează în gama celor 68% și deci face parte din populația generală

# Intervalele de încredere

- **Intervalul de încredere reprezintă gama de plus-minus 2 erori standard de la valoarea statistici $\hat{p}$ a unui eșantion**
- **Nivelul de încredere presupune faptul că 95% dintre valorile parametrului $p$ al populație și situează în intervalul de încredere**
- **În acest caz valoarea statisticii $\hat{p}$ a eșantionului este punct estimator al valorile parametrului $p$ al populație**



# 5. Testul z

## Noțiunea ipotezei

- Testul ipotezei – aplicarea metodelor statistice în probleme a lumii reale
- Aplicarea începe cu o presupunere numită ipoteza nulă $H_0$
- Apoi se execută un experiment pentru testarea ipotezei nule
- În rezultatul experimentului ipoteza nulă fie se respinge fie se eșuează respingerea ei
- Dacă ipoteza nulă eșuează atunci se spune că datele suportă o altă ipoteză reciproc exclusivă ipotezei nule numită ipoteza alternativă $H_1$
- O ipoteză niciodată nu va fi “dovedită”

# Formarea ipotezei

- La inițierea experimentului ipoteza nulă se formulează pentru a fi adevărată
- Doar dacă în rezultatul experimentului ipoteza nulă nu este suportată de date atunci ce analizează ipoteza alternativă
- Dacă se testează ceva ce se presupune a fi adevărat, ipoteza nulă trebuie să reflecte presupunerea
- Dacă se testează ceva ce se dorește a fi adevărat dar nu se poate presupune, atunci ipoteza nulă trebuie să reflecte valoarea opusă

# Exemplu de formarea a ipotezei

- **Predicție: Masa medie de transportare a produsului este de 5,5 kg**

  *Ipoteza nulă $H_0$ :* *masa medie = 3,5 kg*

  *Ipoteza alternativă $H_1$ :* *masa medie ≠ 3,5 kg*

- **Predicție: Frecventarea unui curs va crește nivelul de cunoștințe**

  *Ipoteza nulă $H_0$ :* *nivelul vechi ≥ nivelul nou*

  *Ipoteza alternativă $H_1$ :* *nivelul vechi < nivelul nou*

- **Ipoteza nulă trebuie să conțină o egalitate (=, ≥,≤) iar ipoteza alternativă nu va conține egalitate (≠, >, <)**

# Tipul testelor

- **Dacă ipoteza alternativă conține semnul < (mai mic) atunci testul este unilateral stânga**
- **Dacă ipoteza alternativă conține semnul > (mai mare) atunci testul este unilateral dreapta**
- **Dacă ipoteza alternativă conține semnul ≠ (mai mare) atunci testul este bilateral**

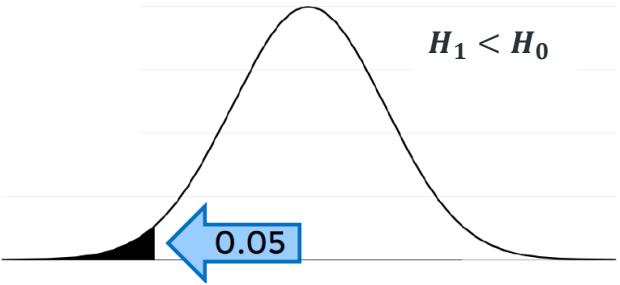
# Pragul de semnificație

- Ipoteza nulă este adevărată în mod implicit la inițierea experimentului
- Se realizează experimentul și se urmărește valoarea probabilității ipotezei nule în funcție de pragul de semnificație α.
- Pragul de semnificație este un nivel de probabilitate sub care ipoteza nulă se respinge
- Valoarea implicită a pragului de semnificație este α = 0,05 (5%)
- Pragul de semnificație este aria de la extremitățile curbei probabilității și depinde de tipul testului

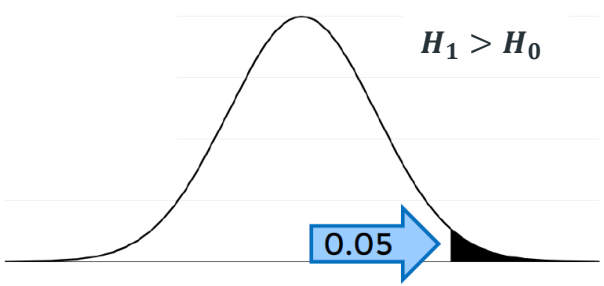
# Pragul de semnificație funcție de tipul testelor

- Se consideră valoarea implicită a pragului de semnificație α=0,05
- Aria de la extremitățile curbei probabilităților în funcție de tipul testelor se va considera

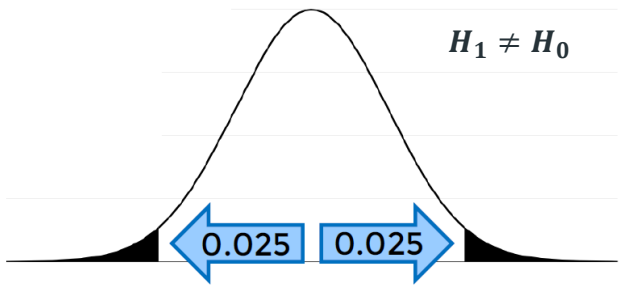
Test unilateral stânga




Test unilateral dreapta




Test bilateral




- Aceste arii permit stabilirea scorului z care se va consider valoarea critică

# Testul valorii medii vs testul proporționalității

- În funcție de tipul statisticii calculate testele pot fi test a valorii medii și test a proporționalității
- Testul valorii medie se utilizează atunci când se caută o valoarea media sau o valoarea specifică într-o populație
- Statistica testului valorii medii se determină cu relația

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- Testul proporționalității se utilizează atunci când se caută valori ce respectă o condiție și de obicei este însoțit de expresii de tip “35%”, “mai mare de”, “mai mult de”, etc.
- Statistica testului proporționalității se determină cu relația

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

# Testul tradițional vs testul valorii percentilei P

- **Algoritmul testului tradițional presupune:**
  - **Se specifică pragul de semnificație α**
  - **Se determină valoarea critică (scorul z) conform valorii α**
  - **Se determină aria extremității conform tipului testului și valorii critice**
  - **Se determină statistica testului Z în funcție de tipul testului**
  - **Se determină poziția statisticii testului Z față de aria extremități și dacă se află în interiorul acesteia ipoteza nulă se respinge**

- **Algoritmul testului valorii P presupune:**
  - **Se specifică pragul de semnificație α**
  - **Se determină statistica testului Z în funcție de tipul testului**
  - **Se determină valoarea perpentilei P conform valorii Z**
  - **Se compara valoarea P cu valoarea pragului de semnificație α**
  - **Dacă P < α ipoteza nulă se respinge**

# Exemplu test a valorii medii (1)

- **O companie încearcă să-și îmbunătățească performanțele site-ului web care are un timp mediu de încărcare de 3,125 s cu abaterea standard de 0,7s. Nivelul de încredere a fost stabilit la 99%. După îmbunătățire, la un eșantion de 40 de lansări a paginii web, timpul mediu a fost de 2,875s. Să se determine dacă noile rezultate sunt statistic mai bune decât cele anterioare**

$$\mu = 3{,}125 \quad \sigma = 0{,}7 \quad \alpha = 0.01 \quad n = 40 \quad \bar{x} = 2.875$$

# Exemplu test a valorii medii (2)

- **Implementarea algoritmului testului valorii medii prin metoda tradițională**
  - *Se stabilește ipoteza nulă – întrucât se dorește o îmbunătățire dar nu se poate prezice se testează valoarea opusă îmbunătățirii adică înrăutățirii performanțelor*

$$H_0: \quad \mu \geq 3{,}125$$

  - *Se stabilește ipoteza alternativă – valoarea reciproc exclusivă a ipotezei nule*

$$H_1: \quad \mu < 3{,}125$$

  - *Se determină poziția arie de la extremități – întrucât ipoteza alternativă conține semnul "<" testul este unilateral stânga*

  - *Se stabilește pragul de semnificație – în sarcină este specificat un nivel de încredere de 99% deci valoarea pragului de semnificație α va fi*

$$\alpha = 0.01$$

# Exemplu test a valorii medii (3)

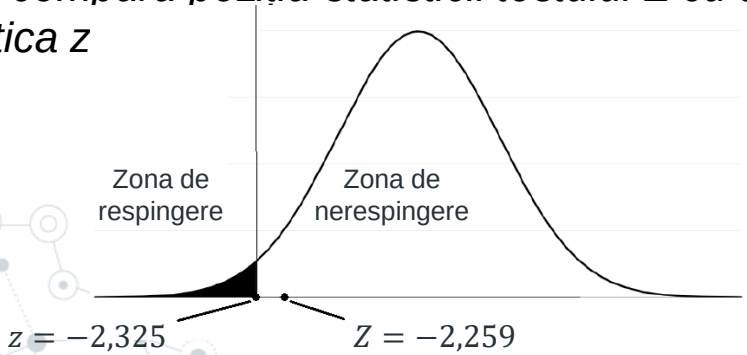
- *Se stabilește valoarea critică z – întrucât aria de la extremități se află în porțiunea stângă, cu ajutorul tabelului z de valori negative se determină scorul z negativ ce corespunde percentilei de valoarea pragului de semnificație α=0.01*

$$\alpha = 0.01 \Rightarrow \quad z = -2{,}325$$

- *Se determină statistica testului Z – întrucât este un test al valorii medii statistica testului se determină cu relația*

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2{,}875 - 3.125}{0{,}7/\sqrt{40}} = -2{,}259$$

- *Se compară poziția statisticii testului Z cu aria extremității limitată de valoarea critica z*




*Întrucât Z se află în zona de nerespingere, ipoteza nulă nu se respinge și deci nu se poate spune statistic că noul site web e mai performant*

# Exemplu test a valorii medii (4)

- **Implementarea algoritmului testului valorii medii prin metoda valorii P**
  - *Se stabilește ipoteza nulă – întrucât se dorește o îmbunătățire dar nu se poate prezice se testează valoarea opusă îmbunătățirii adică înrăutățirii performanțelor*

$$H_0: \quad \mu \geq 3{,}125$$

  - *Se stabilește ipoteza alternativă – valoarea reciproc exclusivă a ipotezei nule*

$$H_1: \quad \mu < 3{,}125$$

  - *Se stabilește pragul de semnificație – în sarcină este specificat un nivel de încredere de 99% deci valoarea pragului de semnificație α va fi*

$$\alpha = 0.01$$

# Exemplu test a valorii medii (5)

- *Se determină statistica testului Z – întrucât este un test al valorii medii statistica testului se determină cu relația*

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2{,}875 - 3.125}{0{,}7/\sqrt{40}} = -2{,}259$$

- *Se stabilește valoarea percentilei P – cu ajutorul tabelului z de valori negative se determină valoarea percentilei conform scorului egal cu statistica testului z = -2,259 (https://www.socscistatistics.com/pvalues/normaldistribution.aspx)*

$$z = -2{,}325 \quad \Rightarrow \quad P = 0{,}0119$$

- *Se compară valoarea percentile P cu valoarea pragului de semnificație*

$$P = 0{,}0119 > \alpha = 0.01$$

*Întrucât P > α ipoteza nulă nu se respinge și deci nu se poate spune statistic că noul site web e mai performant*

# Exemplu test a probabilității (1)

- **O companie de jocuri video a constata că 58% dintre cei 400 de clienți chestionați sunt adolescenți. Sa se determine dacă se poate spune statistic că majoritatea clienților sunt adolescenți.**

$$\hat{p} = 0{,}58, \qquad n = 400, \qquad p = 0{,}5$$

- **Implementarea algoritmului testului proporționalității prin metoda tradițională**
  - *Se stabilește ipoteza nulă – întrucât se dorește a se stabili dacă majoritatea sunt adolescenți dar nu se poate prezice se testează valoarea opusă acestei afirmații*

$$H_0: \quad p \leq 0{,}5$$

  - *Se stabilește ipoteza alternativă – valoarea reciproc exclusivă a ipotezei nule*

$$H_1: \quad p > 0{,}5$$

# Exemplu test a probabilității (2)

- *Se determină poziția arie de la extremități – întrucât ipoteza alternativă conține semnul “>” testul este unilateral dreapta*

- *Se stabilește pragul de semnificație – în sarcină nu este specificat un nivel de încredere, valoarea pragului de semnificație α va fi valoarea implicită*

$$\alpha = 0.05$$

- *Se stabilește valoarea critică z – cu ajutorul tabelului z se determină scorul z ce corespunde percentilei de valoarea pragului de semnificație α=0.05*

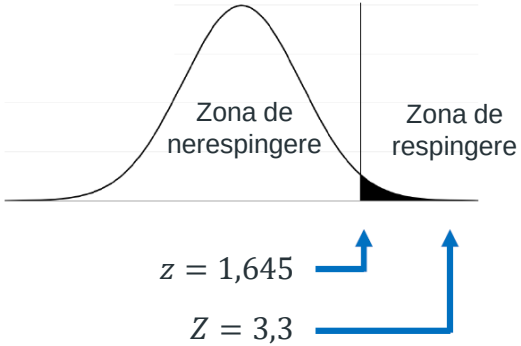
$$\alpha = 0.05 \Rightarrow \quad z = 1{,}645$$

# Exemplu test a probabilității (3)

- *Se determină statistica testului Z – întrucât este un test al proporționalității statistica testului se determină cu relația*

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0{,}58 - 0{,}5}{\sqrt{\frac{0{,}5(1-0{,}5}{400}}} = 3{,}3$$

- *Se compară poziția statisticii testului Z cu aria extremității limitată de valoarea critica z*




*Întrucât Z se în zona de nerespingere, ipoteza nulă se respinge și deci se poate spune statistic că majoritatea clienților companiei sunt adolescenți*

# Exemplu test a probabilității (4)

- **Implementarea algoritmului testului proporționalității prin metoda valorii P**

  - *Se stabilește ipoteza nulă – întrucât se dorește a se stabili dacă majoritatea sunt adolescenți dar nu se poate prezice se testează valoarea opusă acestei afirmații*

$$H_0: \quad p \leq 0{,}5$$

  - *Se stabilește ipoteza alternativă – valoarea reciproc exclusivă a ipotezei nule*

$$H_1: \quad p > 0{,}5$$

  - *Se stabilește pragul de semnificație – în sarcină nu este specificat un nivel de încredere, valoarea pragului de semnificație α va fi valoarea implicită*

$$\alpha = 0.05$$

# Exemplu test a probabilității (5)

- *Se determină statistica testului Z – întrucât este un test al proporționalității statistica testului se determină cu relația*

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0{,}58 - 0{,}5}{\sqrt{\frac{0{,}5(1-0{,}5}{400}}} = 3{,}3$$

- *Se stabilește valoarea percentilei P – cu ajutorul tabelului z se determină valoarea percentilei conform scorului egal cu statistica testului z = 3,3 (https://www.socscistatistics.com/pvalues/normaldistribution.aspx)*

$$z = 3{,}3 \quad \Rightarrow \quad P = 0{,}00048$$

- *Se compară valoarea percentile P cu valoarea pragului de semnificație*

$$P = 0{,}00048 < \alpha = 0.05$$

*Întrucât P < α ipoteza nulă se respinge și deci se poate spune statistic că majoritatea clienților companiei sunt adolescenți*

# 6. Erori de tip 1 și de tip 2

## Esența erorilor

- În unele domenii rezultatele testului ipotezei poate și comparat cu valoarea “adevărată” deja cunoscută

- În rezultatul comparării valorii ipotezei nerespinse de test și valoarea “adevărată” se poate constata prezența erorii testului

- În funcție de respingerea sau nerespingerea incorectă a ipotezei nule erorile pot fi de tipul 1 și de tipul 2

# Erorile de tip 1

- **Erorile de tip 1 apar atunci când se respinge ipoteză nulă deși aceasta este suportată de date (respingere greșită)**
- **Erorile de tip 1 se mai numesc erori fals-pozitiv**
- **Exemple:**
  - *Pornirea alarmei de incendiu*

    $H_0$: $nu\ există\ incendiu$

    *În rezultatul testului a fost pornită alarma deși în realitate nu exista incendiu*
  - *Testul sarcinii*

    $H_0$: $nu\ ești\ însărcinată$

    *Doctorul anunță un bărbat precum că e însărcinat*

# Erorile de tip 2

- **Erorile de tip 2 apar atunci când nu se respinge ipoteză nulă deși aceasta nu este suportată de date (nerespingere greșită)**
- **Erorile de tip 1 se mai numesc erori fals-negativ**
- **Exemple:**
  - *Pornirea alarmei de incendiu*

    $H_0$: $nu\ exită\ incendiu$

    *În rezultatul testului nu a fost pornită alarma deși în realitate exista incendiu*

  - *Testul sarcinii*

    $H_0$: $nu\ ești\ însărcinată$

    *Doctorul anunță o femeie vizibil însărcinată precum că nu e însărcinată*