

Seminar 4

Deschideți fișierul "P4.xls" și salvați-l sub numele: "Nume_seminar4.xls" în directorul dedicat cursului. Fișierul conține mai multe foi (worksheets)!

A. Probabilitatea unei distribuții normale standard Această parte se concentrează pe folosirea tabelelor unei funcții cu distribuție normală.

1. Veti folosi programul Excel pentru a crea un tabel ce conține probabilitățile cumulative corespunzătoare unei distribuții normale standard.


"Distribuția normală standard" este o distribuție normală ce are media aritmetică egală cu 0 și deviația standard egală cu 1 (ex. o distribuție normală exprimată ca punctaj z).

○ În worksheet-ul "**Tabelul probabilitatilor**" creați în coloana A o serie care să varieze de la -3,0 la +3,0 cu incrementul de 0,10 (numerele obtinute reprezinta punctajul z, adica deviațiile standard de la media aritmetică (egală cu zero)):

- introduceți -3 în celula A2

- introduceți $=A2+0,1$ în celula A3

- selectați celula A3 și trageți de cruciulita din colțul din dreapta jos pentru a copia formula din A3 în următoarele 60 celule de mai jos (pana la celula A62).

- folosiți butoanele  din bara de meniu pentru ca numerele ce reprezintă valorile lui z să fie afișate cu două zecimale.

○ În coloana B, calculați probabilitatea cumulativă p a distribuției normale standard.

- în celula B2 folosiți funcția 'NORMSDIST' pentru a calcula probabilitatea cumulativă a primei valori z din coloana A: $=\text{NORMSDIST}(A2)$.

- selectați celula B2 și trageți-o de cruciulița neagră în jos pentru a umple coloana B. Valoarea p trebuie să se apropie de valoarea 1 atunci când z se apropie de valoarea 3.

2. Trasați graficul probabilității cumulative a unei distribuții normale standard, cu z pe axa x și p pe axa y (cel mai bine este să folosiți graficul tip "scatter", cu punctele unite cu o linie).

3. Folosind aceeași metodă ca în exemplele de la curs, (*inclusiv trasarea unei schițe de mana a distribuției normale*), folosiți tabelul creat pentru a calcula probabilitatea ca o măsurătoare dintr-un eșantion cu distribuție normală să aibă o valoare mai mare decât 2,3 deviații standard deasupra mediei aritmetice:

$$p(z > +2.3)$$

4. Răspundeți la următoarea întrebare: la câte deviații standard sub media aritmetică trebuie să fie o măsurătoare dintr-o distribuție normală pentru a avea numai 10% probabilitate de apariție?

Obs.: pentru a da un răspuns exact, va trebui efectuată o interpolare a valorilor din tabel.

5. În coloana C, folosiți funcția NORMSINV pentru a calcula inversul probabilității cumulative a unei deviații normale standard:

- în celula C2 introduceți formula $=\text{NORMSINV}(B2)$

- copiați C2 în jos pentru a completa coloana

6. Încercați să vă dați seama ce calculează funcția NORMSINV (comparați valorile din coloanele A și C). Acum folosiți funcția NORMSINV (în celula F3) pentru a da un răspuns precis la întrebarea 4.

Lucrul cu numere reale

Media aritmetică a maximelor zilnice înregistrate în luna August în Cluj este 21°C, iar deviația standard este 6°C.

NOTA: Media aritmetică a maximelor zilnice se definește ca media tuturor maximelor zilnice care s-au înregistrat vreodată în perioada respectivă – de ex. dacă sunt 10 ani de înregistrări, eșantionul va conține 310 temperaturi.

1. Folosind programul EXCEL se pot converti următoarele temperaturi (in °C) la deviații standard (punctaj z): 19; 15; 34; 25; 27; 13

- în worksheet-ul "Numere reale" introduceți temperaturile în coloana A (începând cu A2).

- În coloana B calculați punctajul z după cum urmează: introduceți în celula B2

formula pentru calculul punctajului z ($z = \frac{x - \bar{x}}{s}$) și apoi *copiați formula în celelalte celule ale coloanei B*.

2. Folosiți funcția NORMSDIST pentru a calcula probabilitatea cumulativă asociată fiecărui punctaj z în celulele corespunzătoare din coloana C.

3. În coloana D, folosiți funcția NORMDIST pentru a calcula probabilitatea cumulativă asociată fiecărei temperaturi direct, fără a mai calcula punctajul z. (pentru a afla cum să folosiți funcția NORMDIST accesați "help on this function").

4. Observați diferența dintre funcțiile: NORMSDIST și NORMDIST.

B. Intervalul de încredere pentru media aritmetică

Această activitate practică vă va ajuta să înțelegeți cum puteți folosi programul EXCEL pentru a calcula **intervalul de încredere** pentru media aritmetică folosind atât **statistica-z** cât și **statistica-t**. (EXCEL nu are o funcție care să permită calcularea directă a intervalului de încredere a mediei în cazul statisticii-t).

Dacă nu aveți opțiunea 'Data Analysis' în meniul programului EXCEL, va trebui să instalați Analysis ToolPak prin selectarea opțiunii 'Tools', apoi 'Add-Ins'. În fereastra care apare alegeți 'Analysis ToolPak', apoi 'OK'.

Partea 1: Teorema limitei centrale

Vom folosi programul EXCEL pentru a ilustra conceptul Teoremei limitei centrale și cum este aceasta legată de eroarea standard a mediei aritmetice. Pentru aceasta vom genera serii de numere aleatoare.

1. Deschideți worksheet-ul "Number generation". Deschideți 'Data Analysis' din meniul 'Tools' și apoi selectați funcția "Random Number Generation".

Pentru a obține 50 de seturi de date, cu câte 100 de variabile în fiecare set în fereastra "Number of variable" scrieți 50, iar în fereastra "Number of Random Numbers" scrieți 100.

Lăsați media aritmetică (mean) și deviația standard 0 respectiv 1 și 'Random Seed' gol.

Selectați 'Normal' pentru tipul distribuției.

Activați "output range" și selectați celula A1, apoi click OK.

2. Programului EXCEL îi trebuie câteva secunde pentru a genera numerele cerute. După ce aveți cele 50 de coloane a câte 100 de numere fiecare, calculați media aritmetică pentru fiecare coloană, folosind o formulă potrivită. (Indicație: introduceți formula pentru prima coloană în prima celulă de sub coloană (A101) și apoi copiați formula sub toate celelalte coloane).

3. Copiați rândul cu cele 50 de medii aritmetice în worksheet-ul "Teorema" (care este goală) în așa fel încât mediile să fie afișate pe coloana:

Selectați cele 50 celulele din worksheet-ul "Number generation" care conțin mediile aritmetice. Activați în worksheet-ul "Teorema".

Folosiți opțiunea 'Paste Special' din meniul 'Edit' pentru a transforma rândul copiat într-o coloană: în opțiunea 'Paste Special', bifați căsuțele 'Values' și 'Transpose', apoi Click OK.

Notă: încercați să înțelegeți ca fac opțiunile 'Values' și 'Transpose'.

4. Construiți o histogramă a acestor date folosind metoda învățată în activitățile anterioare:

- Mărimea claselor (Bin size): 0,1
- Domeniul (bin range): -1.0 to 1.0
- Folosiți opțiunea "Histogram" din "Data Analysis" pentru a calcula distribuția în frecvență.

5. Repetați pașii 2-4, folosind media primelor 10 coloane generate în worksheet-ul "Number generation". Copiați coloana cu noile medii într-o zonă liberă din worksheet-ul "Teorema" pentru a nu acoperii calculele anterioare.

6. Folosiți programul EXCEL pentru a calcula media aritmetică și deviația standard a celor 50 medii aritmetice calculate pentru seriile generate aleator.

7. Folosiți formula potrivită (se găsește în curs) pentru a calcula eroarea standard a mediei unui eșantion ce conține 100 date, respectiv 10 date (eșantion a unei populații ce are media aritmetică egală cu 0 iar deviația standard egală cu 1).

8. Salvați fișierul!

Histogramele pe care le-ați obținut ilustrează cantitativ distribuția mediilor aritmetice ale eșantioanelor, pentru două mărimi diferite ale eșantioanelor. Observați dacă mărimea eșantioanelor afectează distribuția mediilor aritmetice, deci și deviația standard.

Notă: Funcția EXCEL de generare aleatoare a numerelor a fost folosită numai pentru ilustrarea Teoremei limitei centrale. Nu este indicat ca această funcție să fie folosită pentru o aplicație unde rezultatele sunt importante, deoarece procesul folosit de programul EXCEL pentru această funcție nu este perfect aleator.


Partea 2: Intervalul de încredere pentru meda aritmetica

Intervalul de încredere pentru eșantioane mari

Problema: Ca urmare a unui sondaj legat de venitul global anual a 50 de familii din Cluj, s-au obținut următoarele date: Media aritmetică a veniturii globale este 3.000 lei iar deviația standard este 550 lei. Folosind EXCEL, calculați **intervalul de încredere** pentru media aritmetică la un nivel de încredere de 95%

1. Introduceți informațiile de mai sus, cu etichetele corespunzătoare, în worksheet-ul "Interval Incredere").

2. Într-o celulă goală, introduceți nivelul de încredere la care doriți să faceți acest calcul (ex. 0,95) și etichetați celula.

3. Introduceți o formulă care calculează nivelul de semnificație (α) asociat nivelului de încredere. Nu uitați să etichetați celula!
4. Click pe o celulă goală, alegeți funcția 'CONFIDENCE' din subsecțiunea 'Statistics' a icoanei :
 - pentru "alpha" introduceți referința la celula în care ați calculat nivelul de semnificație [amintiți-vă că pentru a introduce o referință puteți să faceți pur și simplu click pe celula pe care doriți, referința va apare în câmpul activat];
 - pentru "Standard_dev" introduceți referința la celula unde ați calculat deviația standard;
 - pentru "Size", celula în care ați notat mărimea eșantionului;

NOTĂ: În câmpurile funcției CONFIDENCE se pot introduce și numere, dar atunci nu se mai pot calcula punctele 5 și 6 în modul indicat (trebuie să se aplice din nou funcția CONFIDENCE într-o altă celulă!).
5. Tastați un nivel de încredere diferit de cel dat și observați cum se schimbă valoarea CI.
6. Tastați o mărime diferită a eșantionului și o deviație standard diferită pentru a observa cum se modifică intervalul de încredere (CI).
7. Salvați datele!

Intervalul de încredere pentru eșantioane mici

Programul EXCEL nu are o formulă pentru calculul intervalului de încredere folosind statistica-t (deci pentru eșantioane mici). În acest caz intervalul de încredere se calculează "de mana" (în modul longhand):

Să presupunem că avem date despre venitul global anual numai de la 10 familii. O analiză preliminară ne dă următoarele date:

\bar{x} (media aritmetică): 2.700 lei

s (deviația standard): 720 lei (calculată folosind corecția Bessel)

n (mărimea eșantionului): 10

Calculați intervalul de încredere pentru media eșantionului pentru un nivel de încredere de 95%.

1. Introduceți datele de mai sus (cu etichete în casuțele adiacente) într-o zona liberă a worksheet-ului "Interval Incredere",.
2. Pentru eșantioane mici, formula pentru calculul intervalului de încredere este: $CI = \bar{x} \pm t \cdot \hat{\sigma}_{\bar{x}}$, deci trebuie determinate mărimile $\hat{\sigma}_{\bar{x}}$ și t:
 - introduceți singuri o formulă pentru calculul erorii standard $\hat{\sigma}_{\bar{x}}$, în funcție de deviația standard (s) și mărimea eșantionului (n) [*relatia o găsiți în curs*].
 - introduceți formule pentru calculul nivelului de semnificație (α) și a gradului de libertate [gradul de libertate este egal cu "n-1" iar $\alpha = 1 - (\text{nivelul de încredere})/100$, dacă nivelul de încredere este exprimat în procente].
 - folosiți formula TINV pentru a calcula probabilitatea-t tăiată la ambele capete (two-tailed probability of t) pentru un nivel de încredere de 95%, introducând referința la valoarea α în câmpul 'Probability' și referința la gradul de libertate în câmpul 'Deg_freedom'.
 - introduceți o formulă pentru calculul produsului $t \cdot \hat{\sigma}_{\bar{x}}$ [programul EXCEL nu poate opera cu semnul \pm]
 - Acum ați obținut domeniul (cu \pm) cu cât poate să varieze media aritmetică pentru un nivel de încredere de 95%.

Folosind formula CONFIDENCE calculați intervalul de încredere (CI) pentru aceleași date. Valoarea obținută este mai mică decât cea calculată folosind statistica-t: de ce au aparut diferențe?

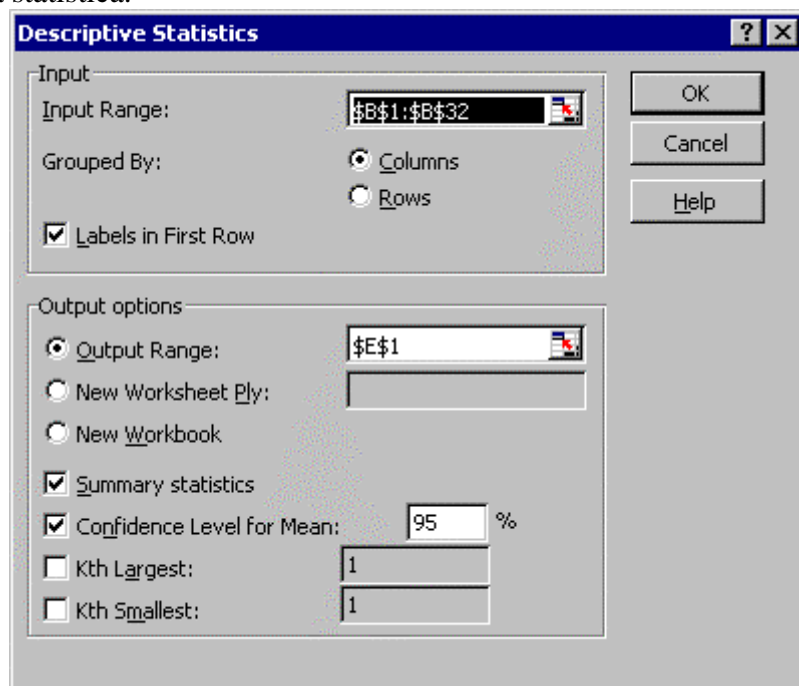
Folosiți funcția TDIST pentru a calcula probabilitățile asociate diferitelor intervale din

jurul mediei aritmetice.

Funcția "Statistica descriptivă"

Majoritatea calculelor pe care le-ați efectuat până acum se pot face folosind funcția 'Descriptive Statistics':

1. Worksheet-ul "LDH" conține valorile lactat dehidrogenazei (LDH) pentru un lot de 31 pacienți.
2. Folosiți 'Data Analysis' pentru a determina un sumar al statisticii acestor date.
 - Selectați 'Tools', apoi 'Data Analysis', apoi 'Descriptive Statistics'.
 - Completați câmpurile într-un mod similar imaginii de mai jos, apoi click OK pentru a genera statistica.



3. Pentru a înțelege ce valori sunt afișate folosind funcția "Descriptive statistics" puteți reface calculele (pentru toate mărimile) în modul pe care le-ați învățat anterior, în scopul comparării valorilor.

[Notă: nivelul de încredere ('Confidence Level') afișat de funcția "Descriptive statistics" este ceea ce noi am numit *interval de încredere*, și a fost calculat folosind funcția CONFIDENCE.]

Salvați datele

C. Probleme

În worksheet-uri noi (numite **P1. P2, .P3. P4**) introduceți datele din problemele următoare (indicând în celulele aditionale ce reprezintă) și folosiți funcțiile corespunzătoare pentru a le rezolva:

P1. Rezultatele unui sondaj efectuat în rândul elevilor din ciclul primar arată că numărul de ore pe săptămână cât aceștia se uita la televizor este normal distribuit, având media 10 și deviația standard 2. Ce procent din totalul elevilor din ciclul primar se uita la televizor:

- A. Sub 4 ore pe săptămână
- B. Peste 12 ore pe săptămână
- C. Între 8 și 14 ore pe săptămână

P2. Media glicemiei unui esantion de adulti este 100 mg/dl cu o variatie de 16 mg/dl.

A. Determinati valoarea maxima a glicemiei sub care se incadreaza 97,5 % din persoanele cuprinse in esantion.

B. Care este punctajul z asociat acestei probabilitati?

P3. Urmatoarele date reprezinta perioada de incubatie (in zile) pentru un esantion aleator de cazuri de hepatita A, provenite din orasul X in anul Y: 34, 20, 23, 28, 31, 25, 30, 29, 22, 32, 27, 28, 33, 24, 27, 26, 31, 26, 39, 38, 28, 30, 23, 25, 29, 31, 30, 34, 27, 24, 26 (gasiti datele in fisierul seminar 4/ worksheet "perioada incubatie").

A. calculati intervalul de incredere de 95% pentru media perioadei de incubatie

B. calculati intervalul de incredere de 90% pentru media perioadei de incubatie.

P4. Ca urmare a unui sondaj legat de venitul global anual a 25 familii din Dej, s-au obținut următoarele date: media aritmetică a veniturii globale este 2.750 lei iar deviația standard este 250 lei. Folosind programul EXCEL, calculați intervalul de încredere pentru media aritmetică cu nivel de încredere de 95%

Nu uitati sa puneti etichete pentru datele introduse si sa salvati fisierul!