Seminar 2

Deschideti fisierul **Seminar 2_Biostat.xls** (dedicat „Statisticii descriptive”) si salvati-l sub numele: "Nume_seminar2.xls" in directorul dedicat cursului. Fişierul conţine mai multe foi (worksheets)!

Partea 1: determinarea marimilor caracteristice **tendintei centrale**

Fisierul conţine o foaie (worksheet) numită "tendinta centrala"in gasiti numarul de pacienti care s-au prezentat la UPU in luna august, pe zile.

Calculati media aritmetică şi medianul (pentru numarul de pacienti) în două moduri: modul "de mână" şi modul “EXCEL”.

Etichetaţi viitoarele calcule (ex: celulele A34-A37) pentru a şti exact ce conţine fiecare celulă (etichetarea se poate face pentru fiecare celulă înainte sau după ce a fost introdusă formula):

	A	B
34	suma	
35	numar date	
36	medie	
37	medie (Excel)	

38

Media aritmetică

Pentru calculul mediei aritmetice trebuie să folosim formula pentru media aritmetică:

$$\bar{x} = \frac{\sum x}{n} \qquad (1)$$

I. se calculează suma pacientilor din luna August:

- click în celula destinata calculului sumei tuturor valorilor (B34), apoi tastati (ignorând ghilimelele) "=SUM("
- selectaţi celulele ce contin numarul de pacienti folosind mouse-ul - veţi observa referirea la celule (cell references) apărând după "=SUM(" ca parte a formulei.
- apoi apăsaţi ENTER şi programul EXCEL va completa formula: "=SUM(B2:B32)"

Nota: Acest calcul se poate face folosind icoana f_x *(function wizard), prin găsirea funcţiei SUM la capitolul "Math & Trig" şi urmarea paşilor indicaţi.*

II. se calculează numărul de date (n) cuprinse in setul de date:

- click pe celula in care doriti sa apara numarul de date (B35)
- introduceţi "=COUNT(domeniu)" unde selectaţi domeniul (range) in acelaşi mod cum aţi făcut la sumă,
- numarul de date se poate calcula şi folosind icoana f_x: se găseşte funcţia "COUNT" şi apoi se urmează paşii indicaţi.
-

III. se calculează media aritmetică folosind formula (1):

- în următoarea celulă (B36), tastaţi "="
- click pe celula B34 unde aţi calculat suma: formula trebuie să se schimbe în "=B34"
- introduceţi simbolul împărţit (/) folosind tastatura (keyboard)
- click pe celula B35 în care aţi calculat numărul de date: formula trebuie să se schimbe în "=B34/B35"
- apasaţi tasta ENTER: răspunsul va fi afişat în celulă, iar formula introdusă este vizibilă în bara de formule de sub meniu.
-

IV. se calculeaza media aritmetică folosind funcţia EXCEL dedicata"=AVERAGE(range)".

- click pe următoarea celulă goală (B37).
- introduceţi formula "=AVERAGE(range)" unde pentru domeniu (range) sunt selectate celulele cu datele a caror medie vrem sa o calculam. Funcţia "Average" se poate găsi folosind icoana , apoi se urmează paşii indicaţi.

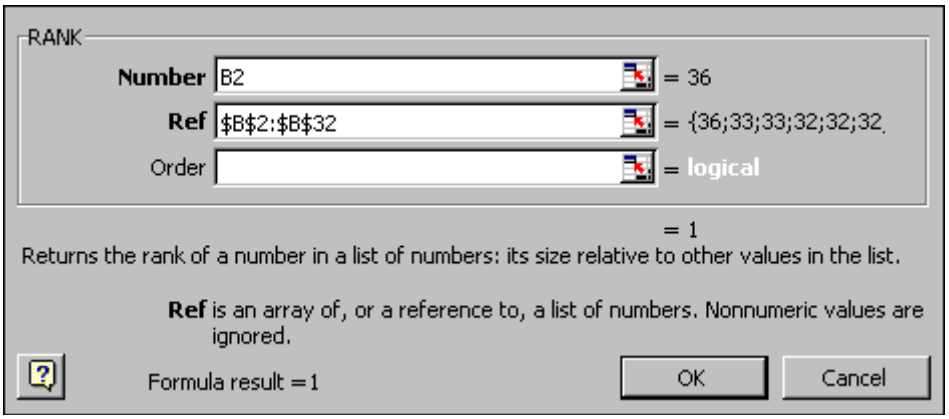
NOTA: folositi icoanele pentru a afişa rezultatul, din celulele B36 şi B37, cu precizia dorită (setează numărul de zecimale după virgulă): întâi se activează (click) celulele care ne interesează şi apoi se foloseşte una din icoanele indicate.

Medianul

Pentru a găsi medianul, mai întâi trebuie ordonate datele, iar apoi trebuie găsită data care are ordinul (n - 1) / 2.

I. pentru a ordona datele folosiţi funcţia "RANK" care se gaseşte cu ajutorul icoanei (function wizard).

- etichetati coloana in care veti calcula ordinul datelor: tastaţi "Ordin" in prima celula din dreapta etichetei "Cluj" (C1).
- selectaţi celula situată la dreapta primei valori din coloana cu numar de pacienti (C2)
- click pe icoana
- selectaţi funcţia RANK din subsetul "statistics"
- click pe câmpul "Number" şi apoi folosiţi mouse-ul pentru a selecta prima valoare din coloana corespunzatoare numarului de pacienti (B2).
- click pe câmpul "Ref", apoi folosiţi mouse-ul pentru a selecta toate valorile din domeniul dorit (de la B2 la B32).
- apăsaţi tasta "F4" pentru a face trimitere absolută la domeniul de celule (B2:B32).
- daca se doreste ordonare ascendenta se pune orice cifra diferita de 0 in campul "Order". Daca nu se pune nimic, sau se pune "0", se va face ordonare descendenta.
-



- click "OK" - celula trebuie să indice ordinul primei valori din setul de date dedicat numarului de pacienti.
- la final, copiaţi celula cu formula în celulele goale de desubt - click pe mica cruciulită neagră din partea de jos-stânga a celulei cu formula şi trageţi-o în jos până când este acoperit tot setul de date.

NOTA: Prin setarea, în prima celulă în care se introduce formula, a referirii absolute la domeniu de celule (B2:B32), partea "Ref" din formulă rămâne constantă în toate celulele de mai jos, permiţându-ne să raportăm fiecare celulă la acelaşi grup de referinţă (ex. numarul de pacienti). Click pe orice celulă din coloana C pentru a vedea cum s-a schimbat conţinutul formulei.

II. Pentru a determina medianul trebuie cunoscut ordinul median:

- formula "COUNT" ne dă valoarea lui n (ar trebui să se obţină n = 31 - numărul de date din set)
- ordinul median este dat de formula: (n - 1) /2.
- Efectuând calculul se obţine 30 / 2 = 15
- medianul este acea valoare care are ordinul 15, deci medianul este 29
-

III. Acum vom determina medianul folosind funcţia "MEDIAN":

- click pe celula dedesubt celei în care s-a calculat media aritmetică.
- introduceţi formula "=MEDIAN(range)", unde "range" este domeniul de celule în care se află datele de interes.
- nu uitaţi să etichetaţi acest rezultat înainte de a salva fişierul.

Modul

Calculul "de mână" a modului este dificilă în EXCEL, deci este mai simplu să se folosească funcţia "MODE".

- În aceeaşi manieră a funcţiilor "MEDIAN" şi "AVERAGE", se calculează "modul" datelor din coloana B (fie prin tastarea formulei "=MODE(range)", fie prin folosirea icoanei functie.

Partea 2. Crearea histogramelor de frecvenţă

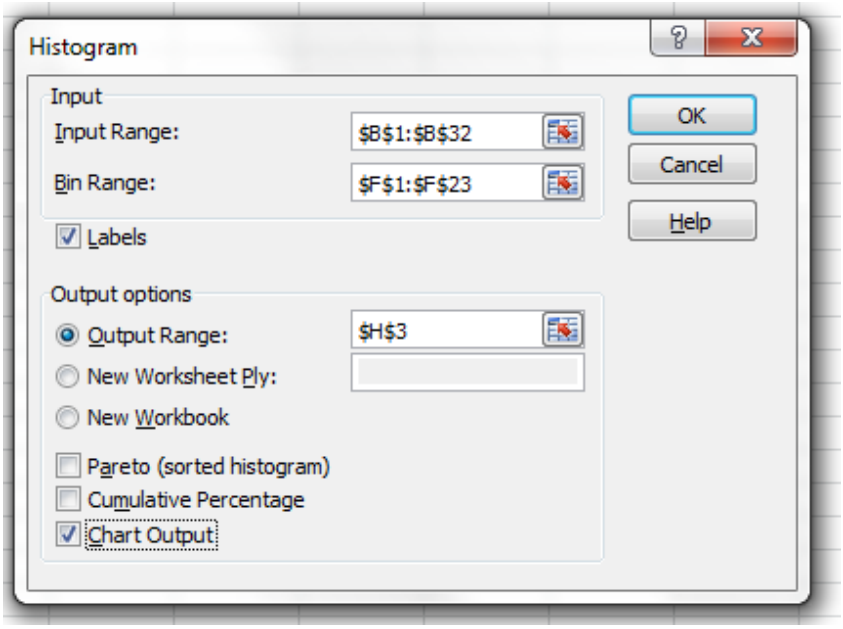
Programul EXCEL are o unealtă folositoare dar limitată pentru a crea histograme, pe care le vom folosi pentru a explora distribuţia în frecvenţă a datelor.

Pentru a crea histograma numarului de pacienti urmaţi indicaţiile următoare.

- Pentru început trebuie definit un set de categorii ("bins") pentru histograma de frecvenţe In acest scop trebuie determinate valoarea minima (MIN) si maxima (MAX) din setul de date.
  - într-o zonă goală a foii (ex. celula F1), se introduce textul "numar" - acesta este eticheta pentru setul de categorii (bins).
  - în celula imediat dedesubt (F2), introduceţi numărul corespunzator minimului setului de date - acesta este cea mai joasă categorie.
  - în celula F3, introduceţi formula "=F2+2" (unde "F2" este celula referinţă pentru cea mai joasă categorie).
  - copiaţi formula în jos pe următoarele celule pana cand ajungeti la valoarea maxima a setului de date- acum ar trebui să aveţi o listă cu categorii între 18 şi 36.
  - *NOTA: în acest caz s-a folosit o variaţie a categoriei (bin size) cu 2, aleasa aleator. O regulă generală (pentru determinarea variaţiei ideale de la o categorie la alta) este aceea că numărul de categorii (bins) trebuie să fie 1,0 + 3,3log(n), unde n este mărimea eşantionului; mărimea unei categorii este domeniul/(1,0+3,3log(n)), sau în EXCEL (referinţa la celulă este arbitrară): =(max(A1:A10)-min(A1:A10))/(1+3.3*log(count(A1:A10)))*
- În continuare folosiţi "unealta" Histogram din Analysis Toolpak al programului EXCEL pentru a crea o histogramă.
  - În bara de meniu EXCEL, click pe "Tools" apoi pe "Data Analysis..."

NOTA: Dacă nu găsiţi opţiunea "Data Analysis...", va fi nevoie să o instalaţi folosind opţiunea "Add Ins..." din meniul "Tools". Cereţi ajutor dacă nu sunteţi sigur ce trebuie făcut.

  - Selectaţi "Histogram" din opţiunile afişate în fereasra "Data Analysis", click OK şi se va deschide fereastre Histogram.
  - Completaţi următoarele câmpuri, făcând click în câmp şi apoi folosiţi mouse-ul pentru a selecta domeniul potrivit din foaia de lucru:
    - Input Range: datele "nr pacienti", *incluzând titlul*
    - Bin Range: categoriile pe care tocmai le-aţi creat, *incluzând titlul.*
    - Output Range: o celulă din partea de sus stânga a unei părţi goale din foaia de lucru (ex. H3)
  - Click pentru a selecta opţiunile "Labels" şi "Chart Output".
  - Când fereastra "Histogram" va arata ca imaginea de jos, click "OK":



EXCEL va produce:

- un tabel în care sunt trecute categoriile (sau bins) ce le-aţi definit, şi va lista frecvenţele de apariţie a datelor în fiecare categorie.
- un grafic ce reprezintă histograma frecvenţelor pentru setul de date.

Petreceţi ceva timp pentru a va juca cu graficul, pentru a vedea cum se poate schimba modul în care arată:

- click pe grafic pentru a-l selecta, mutaţi-l undeva unde il puteţi vedea şi apoi ajustaţi-i mărimea la o dimensiune rezonabilă facând click si apoi trageţi-l de unul din colţuri.
- click-dreapta oriunde pe grafic pentru a schimba opţiunile generale ale graficului.
- dacă doriţi puteţi să editaţi graficul:
  - click pe orice etichetă pentru a edita un text - ex. schimbaţi titlul din "Histogram" în "UPU - August"
  - click pe legendă şi deplasaţi-o în altă poziţie, sau chiar stergeţi-o de vreme ce aveţi un singur set de date.
  - doublu-click pe orice parte a graficului pentru a schimba optiunile: ex. pe axa OY (axa verticală) pentru a schimba scala, culoarea etc.

Partea 3: Determinarea marimilor ce descriu variabilitatea unui set de date

Pentru a determina dispersia unui set de date vom explora funcţiile grafice ale programului EXCEL, precum şi capacităţile programului EXCEL de a calcul măsurile dispersiei.

Graficele create în programul EXCEL sunt legate de setul de date, orice modificare adusă setului de date este vizualizată imediat prin modificarea graficului.

I Vom folosi acelasi set de date (numarul de pacienti tratati la UPU in luna august).

1. Calculaţi manual deviaţia standard a numarului de pacienti.

- Pentru a face acest calcul trebuie să determinati cateva valori intermediare de care aveti nevoie (abaterea, suma patratatelor abaterilor, etc.). Nu uitati sa puneti etichete acestor valori!
- Amintiţi-vă formulele şi funcţiile folosite în practica precedentă, precum şi tehnicile de trimitere la celule: "cell referencing".
- În plus, veţi avea nevoie de următoarele funcţii:
  - =SUM(A1:A10) pentru a calcula suma datelor ce se află în celulele A1-A10;
  - =SQRT(A1) pentru a calcula radicalul valorii din celula A1;
  - =A1^2 pentru a calcula pătratul valorii din celula A1.

(în exemplele de mai sus trimiterea la celule este arbitrară – pentru a obţine rezultatul corect, va trebui să folosiţi o trimitere la celule adaptată setului de date folosit).

2. Calculaţi deviaţia standard pentru setul de date **"Numar pacienti"** folosind următoarele două funcţii EXCEL prin activarea icoanei f_x.

- =STDEV(A1:A10)
- =STDEVP(A1:A10)
-

3. Comparaţi răspunsurile obţinute pentru paşii 1 şi 2. De ce credeţi că sunt două valori diferite pentru pasul 2 şi care este mai apropiat de valoarea reală?

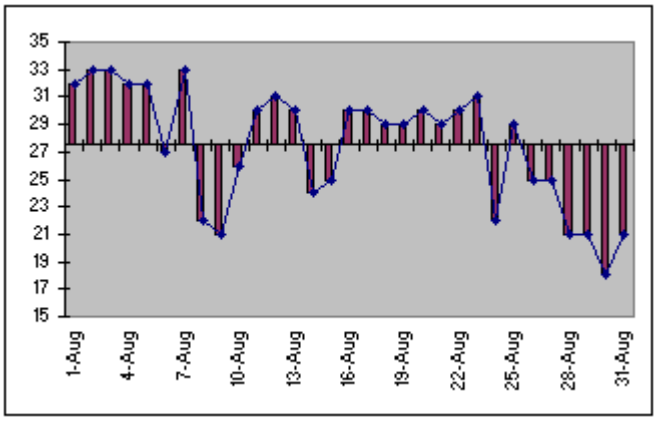
Indiciu: citiţi descrierea celor două funcţii în fereastra "EXCEL help".

4. Folosiţi "Chart Wizard" pentru a reprezenta grafic setul de date:

- Folosiţi mouse-ul pentru a selecta numarul de pacienti din setul de date.
- Deschideţi "chart wizard" făcând click pe icoana din bara de meniu şi creaţi un grafic tip coloane (column chart).
- Adăugaţi un titlu, precum şi numele celor două axe, click next.
- Click pe opţiunea "as object în worksheet 1" pentru a creea graficul în acelaşi worksheet cu datele, şi apoi click "Finish".
- Acum aveţi o reprezentare grafică elementară a datelor, reprezentare ce poate fi îmbunătăţită în diferite moduri:
  - Adăugaţi informaţiile din prima coloană (datele din luna August pentru care s-au înregistrat acele valori) pe axa x făcând click-dreapta cu mouse-ul aşezat pe suprafaţa graficului şi selectând "Source Data".
  - Selectaţi fereastra "Series" şi introduceţi trimiterea la domeniul celulelor din prima coloană în căsuţa "Category (x) axis labels".
  - Schimbaţi formatul datelor: click-dreapta pe axa x, selectaţi "Format Axis", apoi "Number". Alegeţi ce varianţă doriţi.

5. Pentru a avea o mai bună ilustrare a modului în care deviaţia standard derivă din date, vom schimba tipul graficului.

- Selectaţi "Source Data", apoi selectaţi "Add a new series", apoi vom selecta aceleaşi date (nr. pacienti) după ce casuţa "Values" a fost activată
- Click-dreapta pe una din coloanele noii serii pe care aţi adăugat-o, apoi selectaţi "Chart Type":
  - în "Custom Types", selectaţi graficul "line-column".
- Acum formataţi axa y astfel încât axa x să o intersecteze la media aritmetică a numarului de pacienti (calculată anterior):
  - click-dreapta pe axa y a graficului, alegeţi "Format Axis", "Scale" şi introduceţi în căsuta "Category (X) Axis Crosses At" media aritmetică.
- La final graficul vostru trebuie să arate ca cel de jos.
- În continuare puteţi să modificaţi caracteristicile graficului, încercând cât mai multe opţiuni, pentru a dobândi abilităţi diverse de operare a graficelor. De exemplu puteţi adăuga liniaturi orizontale sau verticale (gridlines), puteţi să faceti fundalul graficului alb, sau altă culoare, puteţi schimba culoarea barelor, puteţi schimba domeniul axei y, etc. O procedură utilă uneori este mutarea etichetelor corespunzătoare marcajelor (tick-mark labels) folosind meniul de formatare al axei x.
- Indiciu: dacă doriţi să printaţi numai graficul, click pe grafic pentru a-l selecta, apoi alegeţi "File"şi "Print" din meniul EXCEL şi selectaţi opţiunea "Selected Chart" în secţiunea "Print What".



6. Incercaţi să răspundeţi la următoarele întrebări:

- În ce mod este graficul, pe care l-aţi creat, similar cu calculul de mână (longhand) a deviaţiei standard la pasul 1?
- Care este media aritmetică a tuturor valorilor (x - $\bar{x}$) pe care le-aţi calculat? De ce?
- De ce calculul deviaţiei standard necesită ridicarea la pătrat a acestor valori?
- Ce diferenţă ar fi dacă s-ar folosi valoarea absolută a diferenţelor (x - $\bar{x}$) în loc de pătratul lor pentru a calcula deviaţia standard? (Acest calcul ne dă media aritmetică a deviaţiilor - folosiţi funcţia EXCEL "=AVEDEV()" şi comparaţi rezultatul cu deviaţia standard.
- Încercaţi să descrieţi în cuvintele voastre ce reprezintă deviaţia standard. *Indiciu:* citiţi explicaţiile din "EXCEL help"
-

7. Calculaţi deviaţia cvartală pentru setul de date:

- Folosiţi funcţia QUARTILE pentru a calcula valorile medii ale cvartalelor de jos LQ (sau Q1) şi de sus UQ (sau Q3) corespunzătoare setului de date.
- Introduceţi propria voastră formulă pentru a calcula devaţia cvartală: (Q3-Q1)/2.
- Etichetaţi celulele pentru a şti mai tarziu ce reprezintă (ex. "Q1", "Q3" and "QD")
-

8. Calculaţi domeniul de dispersie:

- determinaţi în două moduri valorile maxime şi minime ale setului da date folosind:
  - funcţia QUARTILE
  - funcţiile MAX şi MIN
- introduceţi o formulă proprie pentru a calcula domeniul de dispersie, folosind celulele ce conţin maximul si minimul setului de date.
-

9. Calculaţi coeficientul de variaţie (CV) pentru setul de date creând propria formulă în EXCEL. Formula pentru calculul coeficientului de variaţie o găsiţi în partea de teorie.

10. Folosiţi funcţiile "SKEW" şi "KURT" pentru a calcula parametrii de asimetrie şi de formă corespunzători datelor .

Tema:

P1. Pentru evaluarea efectului anticonceptionalelor asupra tensiunii arteriale sistolice, s-a realizat un studiu pe un esantion de 10 persoane, care a furnizat urmatoarele date:

TAS (inainte de administrare)	TAS (dupa administrare)
115	128
112	115
107	106
119	128
115	122
138	145
126	132
105	109
104	102
115	117

a) Calculati media diferentei tensiunilor arteriale sistolice obtinute inainte si dupa administrarea de anticonceptionale.
b) Calculati varianta si deviatia standard a TAS (inainte si dupa) precum si a diferentei.
c) Calculati mediana diferentei TAS.
d) Stabiliti gradul de omogenitate al esantionului din punctul de vedere al nivelului Tensiunii arteriale sistolice inainte de administrare.

P2: Urmatoarele date reprezinta varsta inbolnavirii (in ani) de o boala "A" in 25 cazuri de aparitie a acestei boli (selectate aleator):
39, 50, 26, 45, 71, 51, 33, 40, 40, 51, 63, 55, 36, 57, 41, 61, 47, 44, 48, 59, 42, 47, 53, 54, 47.
a) Calculati cu o zecimala urmatoarele statistici: mediana, modul, media aritmetica, domeniul de dispersie, deviatia cvartala, varianta, deviatia standard, coeficientul de variatie
b) Cate din observatii cad in afara urmatoarelor intervale: $\{\bar{x} \pm 1 \cdot s\}$, $\{\bar{x} \pm 2 \cdot s\}$
c) Determinati nivelul de omogenitate al esantionului din punctul de vedere al varsei de imbolnavire.

P3: Calculaţi următoarele valori pentru datele ce reprezintă numarul mediu de pacienti ce au avut boala” A” la cabinetele de familie din orasul “B”

- medianul
- media aritmetică
- deviaţia standard
- deviaţia cvartală
- clasa modală (folosiţi "histogram wizard" pentru a construi un tabel al frecvenţelor, folosind o mărime pentru intervalul dintre frecvenţe "BIN"de 35 şi pornind de la 0)
- coeficientul de variaţie (creaţi propria voastră formulă).
- parametrul de asimetrie (skewness).

2358, 4980, 4140, 8506, 4472, 3155, 4070, 2364, 6069, 5578, 9693, 10721, 2044, 12166, 4643, 7153, 15193, 6405, 5122, 5760, 5354