Biostatistică – Medicină Generală

Lucrarea de laborator Nr.5

Scop: la sfârşitul laboratorului veţi şti:

- ***Să folosiţi foaia de calcul Excel pentru a executa calculele necesare găsirii intervalelor de încredere***
- ***Să efectuaţi teste statistice parametrice cu ajutorul pachetului Excel***

1. Intervale de încredere

În cele de mai jos, sunt prezentate metode de calcul a intervalelor de încredere în Excel.

Media unui eşantion pe care îl avem la dispoziţie este doar o aproximare a **mediei populaţiei** din care provine eşantionul, adică este doar o **aproximare a realităţii**, pe care nu o cunoaştem şi pe care de altfel, nu o să o cunoaştem niciodată. Intervalul de încredere este o aproximare în plus şi în minus a acestei medii necunoscute.

Intervalele de incredere se calculează pornind de la media de eşantionare şi deviaţia standard de eşantionare, care se obţin folosind funcţiile EXCEL, **Averag**e şi **Stdev**.

Formula: Intervalul de incredere de 95% pentru estimarea mediei **m** a populaţiei se calculează cu formula:

$$\overline{X} \pm t_{95\%} * \frac{\sigma}{\sqrt{n}}, \text{ sau cu formula } \overline{X} \pm t_{95\%} * \text{StErr}$$

unde:

- $\overline{X}$ = media eşantionului,
- σ = deviaţia standard a eşantionului
- $\frac{\sigma}{\sqrt{n}}$ = StErr =eroarea standard,
- n = volumul eşantionului (numărul de pacienţi),
- $t_{95\%}$ = pragul teoretic pentru intervalul de incredere de 95%(distribuţie t).

Pentru intervalele de încredere de 99%, avem după cum se ştie de la curs, o formulă cu totul analoagă, singura diferenţă fiind aceea că se schimbă pragul t, punându-se $t_{99\%}$ în locul lui $t_{95\%}$.

Dăm mai jos un tabel al distribuţiei Student, din care se iau pragurile $t_{95\%}$ sau $t_{99\%}$, sau dacă este nevoie, $t_{99,9\%}$. Pe coloana denumită N, sunt listate numărul de grade de libertate, iat pe coloanele $t_{95\%}$, $t_{99\%}$ şi $t_{99,9\%}$, sunt listate valorile pragurilor căutate de noi. De exemplu, pentru 10 grade de libertate, $t_{95\%}$ este 2,228, iar $t_{99\%}$ este 3,169. Pentru 120 de grade de libertate, $t_{95\%}$ este 1,98, iar $t_{99\%}$ este 2,617.

N	95%	99%	99,9%	N	95%	99%	99,9%
1	12,706	63,657	636,620	18	2,101	2,878	3,922
2	4,303	9,925	31,598	19	2,093	2,861	3,883
3	3,182	5,841	12,924	20	2,086	2,845	3,850
4	2,776	4,604	8,610	21	2,080	2,831	3,819
5	2,571	4,032	6,869	22	2,074	2,819	3,792
6	2,447	3,707	5,959	23	2,069	2,807	3,767
7	2,365	3,499	5,408	24	2,064	2,797	3,745
8	2,306	3,355	5,041	25	2,060	2,787	3,725
9	2,262	3,250	4,781	26	2,056	2,779	3,707
10	2,228	3,169	4,587	27	2,052	2,771	3,690
11	2,201	3,106	4,437	28	2,048	2,763	3,674
12	2,179	3,055	4,318	29	2,045	2,756	3,659
13	2,160	3,012	4,221	30	2,042	2,750	3,646
14	2,145	2,977	4,140	40	2,021	2,704	3,551
15	2,131	2,947	4,073	60	2,000	2,660	3,460
16	2,120	2,921	4,015	120	1,980	2,617	3,373
17	2,110	2,898	3,965	Mai mare	1,960	2,576	3,291

a. Calculul intervalului de încredere când datele nu sunt înregistrate deja. Să presupunem că în urma aplicării unor teste de inteligenţă verbală, au fost obţinute mai multe scoruri, ale unor indivizi dintr-un lot de studiu.

Lotul de 40 de indivizi este alcătuit din ofiţeri de resurse umane ai unor firme şi studiul urmăreşte să stabilească cu o aproximaţie cât mai bună nivelul de inteligenţă verbală al indivizilor ce ocupă astfel de posturi. Lotul a fost extras aleator din populaţia formată din ofiţerii de resurse umane ai firmelor de distribuţie a medicamentelor care operează pe teritoriul României. Această populaţie, o vom numi populaţie de referinţă.

Scorul maxim posibil este 48. Scorurile sunt înregistrate pe hârtie. Pentru calculul intervalului de încredere al scorului mediu, prima operaţie este introducerea scorurilor într-o foaie de lucru Excel. Vom încerca de fapt să calculăm dintr-o dată ambele intervale de încredere importante, cel de 95% şi cel de 99%.

Înregistraţi, ca în figura de mai jos, pe coloana A, numerotarea de la 1 la 40, iar pe coloana B scorurile. În imagine nu se văd decât primele 16 scoruri. Veţi înregistra toate cele 40 de scoruri, care sunt:

	A	B	C	D	E
1	Nr	Scor			
2	1	44			
3	2	42		Media	36.75
4	3	37		Deviatia standard	4.82
5	4	43		Eroarea standard	0.76
6	5	36		t 95%	2.02
7	6	47		t 99%	2.7
8	7	29			
9	8	46		Intervalul de incredere de 95 %	
10	9	42		Limita inferioara	35.21
11	10	36		Limita superioara	38.29
12	11	34			
13	12	36			
14	13	35		Intervalul de incredere de 95 %	
15	14	28		Limita inferioara	34.69
16	15	30		Limita superioara	38.81
17	16	35			

44 42 37 43 **36** 47 29 46 42 **36** 34 36 35 28 **30** 35 34 37 34 **27** 36 45 42 37 **38** 41 40 30 34 **32** 37 31 36 38 **36** 41 37 36 35 **36**

Apoi, la D3, D4 şi aşa mai departe, până la D16, scrieţi textele pe care le vedeţi în figură, care sunt texte explicative. Apoi, se fac calculele. La E3, se calculează media cu formula =Average(b2:b41). La E4, veţi calcula deviaţia standard cu formula =Stdev(b2:b41). La E5, veţi calcula eroarea standard cu formula =E4/SQRT(40).

La E6 şi la E7, depuneţi valorile lui $t_{95\%}$, care ne trebuie la calculul intervalului de încredere de 95%, şi a lui $t_{99\%}$, care ne trebuie la calcului intervalului de încredere de 99%. Aceste valori le luaţi din tabelul distribuţiei Student, de pe linia corespunzătoare la 40 de grade de libertate. Normal, trebuia să folosim 39 de grade de libertate, (n-1, numărul de valori minus 1] dar în tabel se observă că nu sunt prevăzute astfel de valori, deoarece diferenţele sunt prea mici şi nu mai contează prea mult în calcule.

La E10 şi E11 se calculează limitele inferioră şi superioară ale intervalului de încredere de 95%, iar la E15 şi E16, limitele intervalului de încredere de 99%, cu formulele următoare:

- La E10, formula =E3-E6*E5 (adică media, minus produsul dintre $t_{95\%}$ şi eroarea standard)
- La E11, formula =E3+E6*E5 (adică media, plus produsul dintre $t_{95\%}$ şi eroarea standard)
- La E15, formula =E3-E7*E5 (adică media, minus produsul dintre $t_{99\%}$ şi eroarea standard)
- La E16, formula =E3+E7*E5 (adică media, plus produsul dintre $t_{99\%}$ şi eroarea standard)

Dacă aţi introdus exact valorile şi calculele au fost făcute corect, intervalul de încredere de 95% va fi între 35,21 şi 38,29, iar intervalul de încredere de 99% va fi între 34,69 şi 38,81.

Interpretarea pe care o dăm acestor rezultate este următoarea: avem o siguranţă de 95% că media populaţiei de referinţă este între 35,21 şi 38,29. Este o aproximare destul de bună. Referitor la intervalul de 99%, interpretarea este: avem o siguranţă de 99% că media populaţiei de referinţă este între 34,69 şi 38,81. Aceasta este o aproximare ceva mai puţin bună, aşa cum ne aşteptam. Din teorie se ştie că intervalele de 95% sunt mai mici (mai înguste, sau mai scurte), iar cele de 99% mai mari (mai largi).

b. Calculul intervalului de încredere când datele nu sunt înregistrate deja

În tabelul HepRen, sunt înregsitrate printre altele şi diametrele lobilor prehepatic şi cardiohepatic ai ficatului, la pacienţii cu diferite afecţiuni hepatice. Cele două coloane pe care sunt înregistrate cele două dimensiuni sunt coloanele R şi S, şi au numele **DPrehep1** şi respectiv, **DCrdHep1**. Lotul este extras aleator din populaţia celor cu afecţiuni hepatice din judeţul Dolj (o vom numi populaţie de referinţă). Ne propunem să determinăm intervalul de încredere de 95% pentru diametrele ambilor lobi la lotul din tabel. În total lotul are 280 de pacienţi.

Deschideţi tabelul **HepRen.xls**. Vom merge cu cursorul de mouse în celula Q282 şi vom scrie Media, în Q283 vom scrie Deviaţia standard, în Q284, scriem Eroarea standard. La Q285 vom scrie t95%, iar la Q287 şi Q288 vom scrie Limita inferioară şi respectiv, Limita superioară.

La R282, se calculează media cu formula =Average(r2:r280). La R283, veţi calcula deviaţia standard cu formula =Stdev(r2:r280). La R284, veţi calcula eroarea standard cu formula =R283/SQRT(279).

La R285, depuneţi valoarea lui $t_{95\%}$, care ne trebuie la calculul intervalului de încredere de 95%. Această valoare o luaţi din tabelul distribuţiei Student, de pe linia corespunzătoare la "Mai mare" şi este 1,96. deci scrieţi acest număr la R285.

	P	Q	R	S	T
276	1.79	78.00	14.50	10.00	1
277	1.73	66.00	14.00	10.00	1
278	1.82	86.00	14.00	9.00	1
279	1.73	76.00	16.00	8.00	1
280	1.82	88.00	16.00	11.00	1
281					
282		Media	14.74	8.65	
283		Deviatia standard	1.65	1.19	
284		Eroarea standard	0.10	0.07	
285		t95%	1.96	2.96	
286					
287		Limita inferioara	14.55	8.44	
288		Limita suerioara	14.94	8.86	
289					

La R287 şi R288 se calculează limitele intervalului de încredere de 95% cu formulele următoare:

- La R287, formula =R282-R285*R284 (adică media, minus produsul dintre $t_{95\%}$ şi eroarea standard)
- La R288, formula = R282+R285*R284 (adică media, plus produsul dintre $t_{95\%}$ şi eroarea standard)

Pentru calculul intervalului de încredere al mediei valorilor de pe coloana S, copiaţi formuele pe coloana S astfel: puneţi cursorul pe r282 şi trageţi din colţul din dreapta jos, spre dreapta, când aveţi cursor în formă de cruce. Apoi faceţi aceeaşi operaţie cu celulele R283, R284, R285, R287 şi R288.

Interpretarea rezultatelor pe care le-am obţinut este următoarea: Avem o siguranţă de 95% că media diametrului lobului prehepatic la pacienţii cu afecţiuni hepatice (populaţia de referinţă), este între 14,55 şi 14,94. Este o aproximare foarte bună. La fel, media diametrului lobului cardiohepatic (în populaţia de referinţă), este aproape sigur (95% sigur) între 8,44 şi 8,86. Aceasta este de asemeni o aproximare foarte bună.

2. Testul t al lui Student

Testul t al lui Student, reprezintă de fapt o familie de teste statistice care pot fi aplicate în diferite situaţii practice. Astfel:

- *Se poate testa dacă o medie a unei serii de valori obţinute prin măsurători pe un lot de pacienţi, este compatibilă cu o medie teoretică dată, sau ştiută dinainte*
- *Se pot compara cu ajutorul acestui test mediile a două loturi diferite, atunci când se ştie că dispersiile sunt diferite*
- *Se pot compara cu ajutorul acestui test mediile a două loturi diferite, atunci când se ştie că dispersiile sunt egale*
- *Se pot compara mediile obţinute prin măsurători pe aşa –numitele loturi pereche*

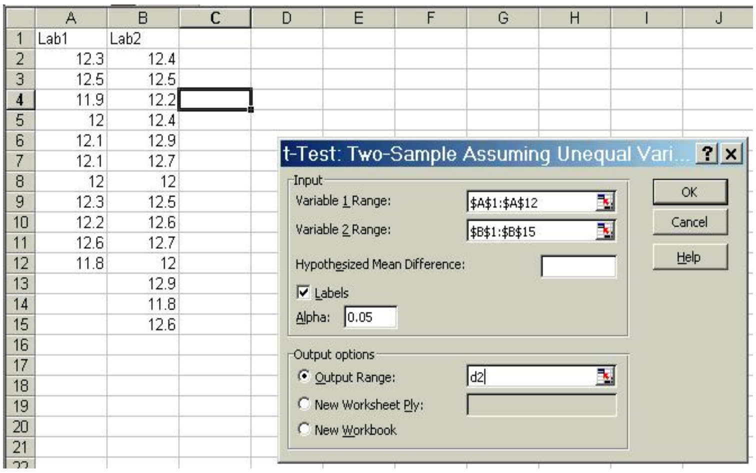
Pentru exemplificare, în cele de mai jos, se consideră că se compară serii de valori pe care le obţinem punând doi laboranţi să măsoare de mai multe ori eşantioane sau probe, sau analize pe care le-am preparat în condiţii identice. Eventualele diferenţe între rezultatele furnizate de ei s-ar datora în acest caz faptului că unul foloseşte metode mai bune sau mai precise, sau aparatură mai modernă, sau unul are tendinţa de a da rezultate exagerate în minus sau în plus, etc. Exemplul ales este întâmplător, în locul celor doi laboranţi putem să considerăm că trimitem probele la două laboratoare şi dorim să comparăm rezultatele acestora. Tot astfel, putem presupune că numerele din cele două serii de valori sunt rezultate obţinute făcând analizele la două loturi de pacienţi. Astfel, în exemplul de mai jos, cele 11 şi respectiv 14 valori, pot reprezenta hemoglobina la 11 pacienţi trataţi cu un tratament, şi respectiv la 14 pacienţi trataţi cu un alt tratament şi dorim să comparăm eventualele diferenţe.

În general, atunci când facem un test statistic de comparare a mediilor, avem două serii de valori obţinute prin măsurători pe pacienţii din două loturi diferite.

Presupunem deci că, pentru a verifica dacă doi laboranţi lucrează sau nu la fel, sau unul are tendinţa de a furniza valori semnificativ diferite, **se prepară un număr de 25 de probe identice** şi sunt trimise la analiză, 11 la unul şi 14 la celălalt. (am luat 25 de probe şi am trimis 11 la unul şi 14 la celălalt, dar testul merge şi pentru alte cazuri).

Se scriu pe două coloane valorile de la cei doi laboranţi, aşa cum vedeţi în figura de mai jos, care nu sunt egale nici între ele şi nici valorile de la un laborant nu sunt aceleaşi cu cele ale celuilalt, deoarece la ambii laboranţi avem erori de măsurare care sunt inerente.

A) Testul pentru dispersii inegale. După ce se introduc valorile de la tastatură, ca în figura de mai sus, se execută secvenţa **Tools → Data Analysis,** iar din fereastra care apare, se alege penultima opţiune: **t-test: Two Sample Assuming Unequal Variances.** Apoi, apăsaţi butonul OK.



În fereastra care apare, şi pe care o vedeţi în figura din stânga, executaţi următoarele:

- În caseta de dialog **Variable 1 Range**, scrieţi A1:A12, în caseta **Variable 2 Range**, scrieţi B1:B15.
- Apoi, bifaţi caseta de validare Labels şi butonul de opţiune Output Range, iar în caseta corespunzătoare, scrieţi D2.

- Apoi apăsaţi butonul OK.

Efectul, este apariţia tabelului pe care îl vedeţi în figură. (Pentru claritate, coloanele au fost redimensionate (lăţite), iar pe coloana H, sunt făcute unele comentarii explicative care nu sunt furnizate de program).

	A	B	C	D	E	F	G	H
1	Lab1	Lab2						
2	12.3	12.4		t-Test: Two-Sample Assuming Unequal Variances				
3	12.5	12.5						
4	11.9	12.2			*Lab1*	*Lab2*		
5	12	12.4		Mean	12.16363636	12.44285714		mediile
6	12.1	12.9		Variance	0.060545455	0.113406593		dispersiile
7	12.1	12.7		Observations	11	14		volumele
8	12	12		Hypothesized Mean Difference	0			
9	12.3	12.5		df	23			
10	12.2	12.6		t Stat	-2.393893333			
11	12.6	12.7		P(T<=t) one-tail	0.01260842			
12	11.8	12		t Critical one-tail	1.713870006			
13		12.9		P(T<=t) two-tail	0.02521684			p=2,52%
14		11.8		t Critical two-tail	2.068654794			

Pe linia **Mean**, sunt afişate cele două medii. Un laborant dă media 12,16, iar celălalt 12,44, deci valori medii sensibil apropiate.

Pe linia **Variance**, sunt afişate dispersiile, primul având dispersia 0,06, iar celălalt 0,11, deci puţin diferite, primul dă valori mai constante, al doilea mai împrăştiate.

Rezultatul ***p***, al testului se află la linia pe care programul a scris **P(T<t) two-tail**, şi este 0,0252, adică 2,52%

Având în vedere regula de respingere a ipotezei de nul atunci când *p* este sub 5%, vom spune că între mediile măsurătorilor celor doi laboranţi este o diferenţă semnificativă. Laborantul al doilea are tendinţa de a furniza valori superioare celor furnizate de primul.

Acest test se poate aplica dacă nu se ştie despre cei doi laboranţi cât de dispersate sunt valorile furnizate de ei. Astfel, unul din ei ar putea furniza valori centrate în jurul mediei mai strâns, adică cu dispersie mică (sau abatere standard mică). În acest caz concret, s-a văzut că al doilea dă pe lângă o medie uşor crescută, şi o dispersie mai mare a valorilor.

Dacă veţi fi puşi în situaţia să raportaţi rezultatul unui astfel de test, pe lângă valoarea rezultatului p şi a interretării lui, trebuie raportate mediile şi deviaţiile standard la cele două loturi. Programul nu furnizează însă deviaţia standard, astfel că va trebui să o calculaţi. Acest lucru se face simplu. Mergeţi de exemplu în celula E16 şi scrieţi formula =sqrt(E5), iar în F16 scrieţi =sqrt(F5). Apoi raportarea rezultatului testului poate fi făcută precizând la ambele loturi mediile, deviaţiile standard, rezultatul p, precum şi interpretarea acestuia.

B) Testul pentru dispersii egale. Dacă din studii anterioare sau din alte surse de informaţie se ştie că cei doi laboranţi (sau cele două laboratoare), dau măsurători care au dispersii egale (deoarece folosesc aparate de aceeaşi precizie, de exemplu), se poate face testul t Student pentru două eşantioane în cazul dispersiilor egale

	A	B	C	D	E	F	G	H	I
1	Lab1	Lab2							
2	12.3	12.4		t-Test: Two-Sample Assuming Equal Variances					
3	12.5	12.5							
4	11.9	12.2			*Lab1*	*Lab2*			
5	12	12.4		Mean	12.16363636	12.44285714		mediile	
6	12.1	12.9		Variance	0.060545455	0.113406593		dispersiile	
7	12.1	12.7		Observations	11	14		volumele	
8	12	12		Pooled Variance	0.09042349			dispersie comuna	
9	12.3	12.5		Hypothesized Mean Difference	0				
10	12.2	12.6		df	23				
11	12.6	12.7		t Stat	-2.304610092				
12	11.8	12		P(T<=t) one-tail	0.015278815				
13		12.9		t Critical one-tail	1.713870006				
14		11.8		P(T<=t) two-tail	0.03055763			p=3,05%	
15		12.6		t Critical two-tail	2.068654794				
16									

Refaceţi testul ca la punctul **a**, şi în locul alegerii **t-test: Two Sample Assuming Unequal Variances**, alegeţi **t-test Two Sample Assuming Equal Variances**

În ferestra care apare, completaţi ca la punctul **a**. Observaţi că rezultatele sunt dispuse într-un tabel asemănător, totuşi, sunt unele diferenţe.

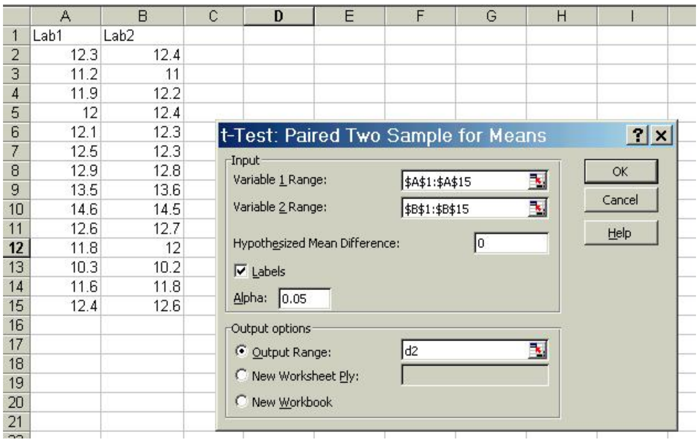
Rezultatul ***p***, este puţin mai mare, dar diferenţa între cele două medii este tot semnificativă, deoarece valoarea lui *p* este tot sub 5%.

*În practică, bineînţeles că trebuie dinainte stabilit care din cele două variante va fi cea corectă şi va fi folosită numai acea variantă. De aceea, sau se află din literatura de specialitate în ce caz suntem, sau trebuie dinainte calculate dispersiile pentru cele două serii de valori şi, când se suspectează că este o diferenţă semnificativă, se preferă varianta **t-test: Two Sample Assuming Unequal Variances,** iar dacă suntem suficient de siguri că diferenţa între dispersiile celor două serii de valori este întâmplătoare, putem efectua varianta **t-test Two Sample Assuming Equal Variances.***

De fapt, chiar când avem impresia că varianţele diferă semnificativ, numai un test statistic poate decide acest lucru suficient de sigur. Excel pune la dispoziţia utilizatorului testul F de comparare a dispersiilor ca test care să ne ajute în a decide care din cele două variante de mai sus este mai bine să o folosim. (Vezi mai jos, Testul F de comparare a dispersiilor). În plus, nu trebuie să se uite că acest test se aplică numai în ipoteza că valorile furnizate de laboranţi sunt distribuite normal.

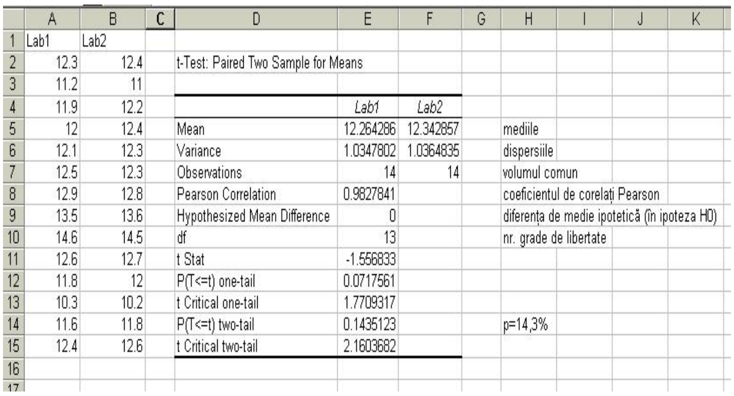
C). Testul t pentru măsurători pereche

Pentru a exemplifica modul de folosire a acestui test să revenim la exemplul cu cei doi laboranţi.



*Putem face compararea între laboranţi şi în alt fel: luăm 14 probe diferite, şi trimitem din **fiecare probă, câte o mostră** la fiecare din cei doi laboranţi. Acum ar trebui ca la mostrele din proba 1 ei să dea acelaşi rezultat, la mostrele din proba 2 să dea tot acelaşi rezultat, dar nu acelaşi cu rezultatul de la proba 1, etc. Totuşi, ei nu vor da rezultate chiar identice între ele, din cauza erorilor. În acest caz, trebuie efectuat pentru compararea mediilor testul t – Student pentru măsurători pereche.*

Introduceţi datele pe care le vedeţi în imaginea de mai sus pe coloanele A şi B, apoi se execută secvenţa **Tools → Data Analysis**, iar din fereastra care apare, se alege opţiunea: **t-test: Paired Two Sample for Means.** Completaţi ca în figura de mai sus, apoi apăsaţi butonul OK.



	A	B	C	D	E	F	G	H
1	Lab1	Lab2						
2	12.3	12.4		t-Test: Paired Two Sample for Means				
3	11.2	11						
4	11.9	12.2			*Lab1*	*Lab2*		
5	12	12.4		Mean	12.264286	12.342857		mediile
6	12.1	12.3		Variance	1.0347802	1.0364835		dispersiile
7	12.5	12.3		Observations	14	14		volumul comun
8	12.9	12.8		Pearson Correlation	0.9827841			coeficientul de corelaţi Pearson
9	13.5	13.6		Hypothesized Mean Difference	0			diferenţa de medie ipotetică (în ipoteza H0)
10	14.6	14.5		df	13			nr. grade de libertate
11	12.6	12.7		t Stat	-1.556833			
12	11.8	12		P(T<=t) one-tail	0.0717561			
13	10.3	10.2		t Critical one-tail	1.7709317			
14	11.6	11.8		P(T<=t) two-tail	0.1435123			p=14,3%
15	12.4	12.6		t Critical two-tail	2.1603682			

Rezultatele sunt listate în figura de mai sus fiind asemănătoare celor de la celelalte teste t.

Rezultatul p al testului este p=0,143, sau **p=14,3%.** Respectând regula de decizie de la acest tip de test, ipoteza de nul nu se respinge, datele furnizate de cei doi laboranţi nu au medii care să difere semnificativ.

3. Testul ANOVA

Este un test care testează dintr-o dată medii mai multor loturi. Aceasta înseamnă că avem mai mult de două loturi, pe care am făcut măsurători şi am obţinut tot atâtea serii de valori. În practică, acesta este cazul atunci când cele 3 sau mai multe loturi sunt supuse la tratamente diferite, şi se urmăreşte să se stabilească dacă efectele tratamentului sunt diferite la unul sau altul din loturi.

De exemplu, pentru a stabili eficacitatea unui pansament gastric folosit în tratamentul ulcerului, la diferite tipuri de ulcer, se folosesc trei loturi de pacienţi, fiecare lot cuprinzând pacienţi care au aceeaşi formă de ulcer (să le notăm cu A, B şi C). Se măsoară la fiecare pacient numărul de zile de tratament necesar pentru vindecare completă, iar cele trei serii de valori care se obţin trebuie supuse unui test statistic de comparare a mediilor pentru a stabili dacă eventualele diferenţe de medie (a duratei tratamentului) sunt semnificative. Evident, pentru a aputea aplica testul, este nevoie ca loturile să fie mari, deoarece la loturi mici, numărul de zile de tratament nefiind distribuit Gauss, testul ANOVA nu poate fi aplicat.

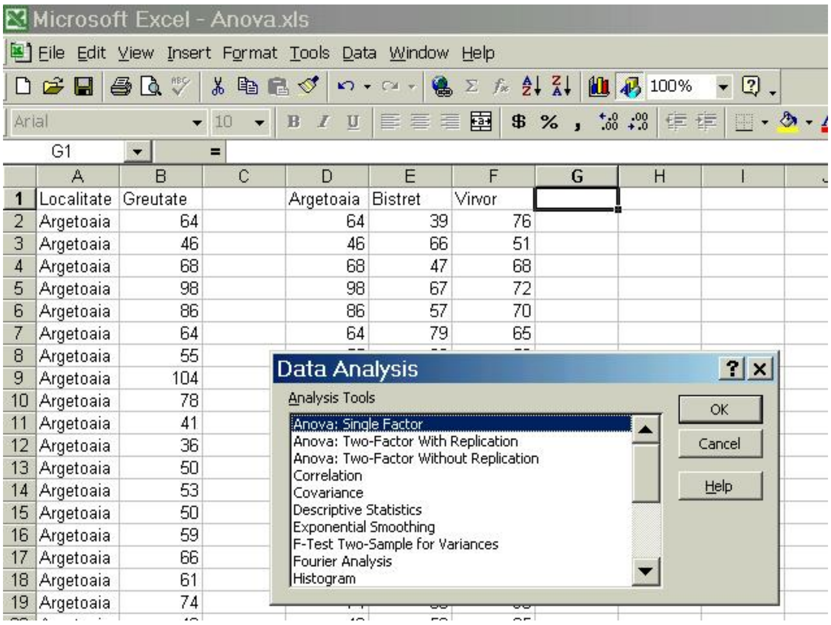
Testul ANOVA, face compararea celor trei medii dintr-o dată. Dacă nu găsesc diferenţe semnificative, tratamentul folosit nu poate fi considerat ca fiind mai eficace la una dintre formele de ulcer (A, sau B sau C). Dacă însă testul găseşte diferenţe semnificative, înseamnă că pansamentul este mai eficace la una din formele de ulcer şi mai puţin eficace la altele, sau este mai eficace la două din cele trei forme şi mai puţin eficace la cea de-a treia. Nu se poate stabili sigur în care din cele două situaţii de mai sus suntem. Oricum în acest caz, este nevoie ca studiul statistic să fie continuat prin aplicarea unor teste de comparare a două loturi.

Se pune întrebarea de ce nu se folosesc de la început teste care compară mediile loturilor câte două odată. Cauzele sunt două:

- Dacă testul ANOVA dă un rezultat nesemnificativ, atunci comparările câte două sunt inutile.
- Dacă numărul de loturi este mai mare, trebuie făcute foarte multe teste de comparare câte două. Pentru fiecare test Student de comparare, se face alt studiu clinic, pe alte loturi. De exemplu pentru 6 loturi, trebuie făcute 15 comparări luate câte două.

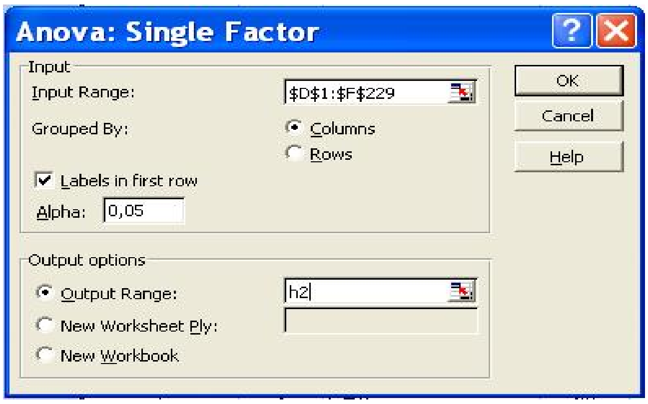
Pentru a efectua un exemplu, deschideţi tabelul OBEZ. Pe coloanele D şi M sunt înscrise localitatea şi respectiv greutatea fiecăruia din cei 510 subiecţi înregistraţi în tabel. Transferaţi cele două coloane pe o altă foaie de calcul, astfel:

- Inseraţi o nouă foaie de calcul folosind secvenţa **Insert->Worksheet**. Se va deschide o nouă foaie de calcul, goală
- Reveniţi pe foaia **Obez**, cu un clic pe numele ei, în partea stânga-jos a ecranului
- Executaţi clic pe litera **C**, care este numele coloanei a treia **(Localitate).** Coloana va fi selectată
- Executaţi clic pe butonul dreapta de mouse şi alegeţi din meniul care se deschide, **Copy**
- Executaţi clic pe numele foii de calcul goale, în partea din stânga jos a ecranului
- Executaţi clic pe litera **A**, numele primei coloane a foii goale
- Executaţi clic pe butonul drept de mouse şi alegeţi din meniu comanda **Paste**. Coloana **Localitate** a fost copiată
- Executaţi clic pe numele foii **Obez**
- Executaţi clic pe litera **L**, care este numele coloanei a 12-a **(Greutate**). Coloana va fi selectată
- Executaţi clic pe butonul dreapta de mouse şi alegeţi din meniul care se deschide, **Copy**
- Executaţi clic pe numele noii foi de calcul, în partea din stânga jos a ecranului
- Executaţi clic pe litera **B**, numele celei de-a doua coloane a foii
- Executaţi clic pe butonul drept de mouse şi alegeţi din meniu comanda **Paste**. Coloana **Greutate** a fost copiată
- Executaţi clic în orice celulă de pe coloana **A** sau **B**
- Cu secvenţa **Data->Sort**, sortaţi după localitate



Scrieţi pe coloanele D, E şi F, numele celor trei localităţi din care provin subiecţii, aşa cum vedeţi în figură. Apoi, copiaţi greutăţile subiecţilor din Argetoaia pe coloana D, cele ale celor din Bistreţ pe coloana E iar ale celor din Vîrvor pe coloana F.

Apoi executaţi secvenţa **Tools → Data Analysis** şi alegeţi din fereastra care apare testul ANOVA, care este primul din listă. Apoi apăsaţi OK.



Completaţi fereastra care apare aşa cum vedeţi în figura de mai sus.
Observaţi că în caseta de dialog **Input Range**, este dată zona acoperitoare pentru toate cele trei coloane, din celula D1, de unde încep datele, până în celula F229. Numai pe una din coloane găsim date la linia 229. În caseta Output Range, s-a precizat h2, cu scopul ca programul să depună rezultatele acolo.
La final, apăsaţi butonul OK.
Testul furnizează multe date numerice ca răspuns. Importante sunt mediile (**Average**), în cazul nostru 60,78 Argetoaia, 60,65 Bistreţ şi 67,55 Vîrvor. Pe coloana L, în dreapta mediilor, sunt listate dispersiile. De obicei este nevoie să extrageţi radicalul din aceste valori, pentru aflarea deviaţiilor standard.

H	I	J	K	L	M	N
Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Argetoaia	228	13860	60.78947368	189.7087874		
Bistret	124	7521	60.65322581	393.2039732		
Vîrvor	143	9661	67.55944056	250.762287		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	4728.460687	2	2364.230344	9.1564536	0.00012463	3.014051231
Within Groups	127036.2282	492	258.2037159			
Total	131764.6889	494				

Rezultatul pe baza căruia se ia decizia este P-value, care se observă că este 0,000124.

Interpretare: ipoteza de nul se respinge, există cel puţin o diferenţa între mediile greutăţii la subiecţii din cele trei comune care este foarte înalt semnificativă. Urmărind valorile medii, se observă ce subiecţii din comuna Vîrvor au greutatea mai mare.

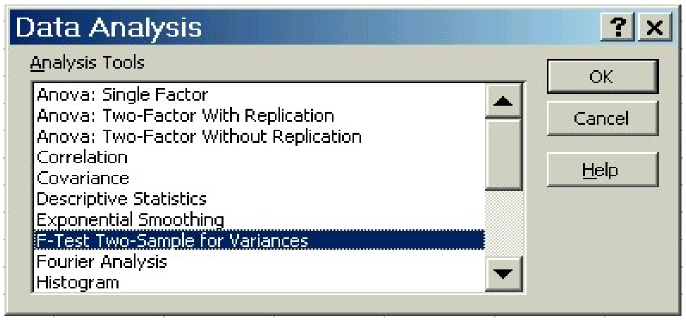
Dacă este nevoie să raportaţi rezultatul acestui test, veţi preciza mediile de greutate la fiecare din cele trei categorii, deviaţiile standard pe care le calculaţi prin extragerea radicalului din dispersii, valoarea lui p, precum şi interpretarea acestuia.

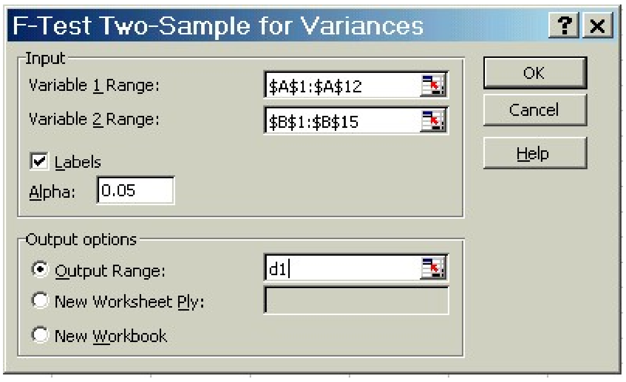
4. Testul F de comparare a dispersiilor a două loturi

Pentru a decide dacă dispersiile seriilor de valori obţinute prin măsurători pe două loturi de pacienţi sau de probe, diferă semnificativ, se poate folosi testul **F**, al lui Fisher, de comparare a dispersiilor. Pentru aceasta, să presupunem că avem valorile din exemplul descris la testul t-Student (sau altele, depinde de situaţie), listate în figura de mai jos, stânga. Le veţi introduce pe o foaie de EXCEL şi veţi proceda apoi astfel:

- Se execută **secvenţa Tools → Data Analysis**
- Din ferestra din figura de mai jos, se alege opţiunea **F-Test Two-Sample for Variances**
- Se apasă butonul OK.

	A	B
1	Lab1	Lab2
2	12.3	12.4
3	12.5	12.5
4	11.9	12.2
5	12	12.4
6	12.1	12.9
7	12.1	12.7
8	12	12
9	12.3	12.5
10	12.2	12.6
11	12.6	12.7
12	11.8	12
13		12.9
14		11.8
15		12.6
16		





D	E	F	G
F-Test Two-Sample for Variances			
	Lab1	*Lab2*	
Mean	12.1636	12.4429	
Variance	0.0605	0.1134	
Observations	11	14	
df	10	13	
F	0.5339		
P(F<=f) one-tail	0.1624		p=16,24%
F Critical one-tail	0.3464		

Ferestra care a apărut (vezi figura de sus), se completează aşa cum vedeţi că s-a completat pe figură, după care se apasă butonul OK. Efectul este apariţia tabelului din figura de mai sus (dreapta), în care observaţi că au fost listate începând de la D1, aşa cum s-a cerut la pasul anterior, mediile, dispersiile, numărul de valori din serii (Observations), şi alte câteva informaţii.

Rezultatul testului se culege ca **P(F<=f) one–tail**, are valoarea 0,1624, adică **p=16,24%.** (Atenţie, în imagine coloanele au fost ajustate, valorile sunt afişate cu 4 zecimale, adică s-a folosit secvenţa **Format → Cells → Number** şi s-a ales 4 zecimale). Valoarea de pe coloana G nu a fost furnizată de program ci este adăugată ulterior, pentru a pune în evidenţă locul şi valoarea rezultatului p.
Folosind metoda de interpretare a rezultatului unui test statistic, deducem că, deoarece ***p*** este peste 5%, ipoteza de nul nu se respinge (diferenţa dintre dispersiile celor două loturi este nesemnificativă). De exemplu, dacă trebuie să comparăm şi mediile celor două loturi, putem folosi testul *t*, varianta **t-test Two Sample Assuming Equal Variances.** Dacă diferenţa ar fi fost semnificativă, adică rezultatul ***p*** ar fi fost sub 5%, eram obligaţi pentru compararea mediilor celor două loturi să folosim varianta **t-test: Two Sample Assuming Unequal Variances.**

5. Exerciţii şi chestiuni de examen

1. Comparaţi bilirubina totală pacienţilor de sex masculin faţă de cea a acelor de sex feminin. Veţi proceda ca la testul ANOVA, adică veţi transfera coloanele **Sex** şi **Brt** pe o foaie goală, veţi sorta după sex, veţi copia valorile bilirubinei la femei pe o coloană separată, iar cele ale bărbaţilor pe o altă coloană. Apoi, veţi face testul t de comparare pentru dispersii inegale.

2. În tabelul ASTM, pe coloana **Vârsta** este înregistrată vârsta fiecărui pacient, iar pe coloana SEX, sexul fiecărui pacient. Comparaţi mediile de vârstă la cele două sexe. Veţi proceda ca la ex1.

3. Într-un experiment, 22 subiecţi au fost împărţiţi aleator în două grupe, una formată din 13 pacienţi care au urmat un tratament de recuperare obişnuit (T_1) şi a doua formată din 9 pacienţi care au urmat un tratament alternativ (T_2). Scopul a fost de a determina dacă fiecare din cele două tratamente determină o creştere semnificativă a scorului şi dacă între cele două tratamente există o diferenţă semnificativă de scor înainte şi după tratament. S-au evaluat toţi cei 22 de pacienţi pe scala LDP atât înainte de efectuarea tratamentelor cât şi după. Din studii anterioare se ştie că scorul LDP are o distribuţie apropiată de o distribuţie Gauss şi deci se poate folosi testul ***t*** al lui Student. S-au obţinut următoarele scoruri:

Nr	Grupa	Înainte	După
1	T1	22	24
2	T1	14	17
3	T1	24	22
4	T1	15	12
5	T1	8	12
6	T1	18	20
7	T1	14	16
8	T1	22	22
9	T1	21	19
10	T1	17	20
11	T1	16	18
12	T1	14	17
13	T1	21	22
14	T2	12	18
15	T2	23	26
16	T2	14	21
17	T2	9	18
18	T2	17	22
19	T2	16	22
20	T2	22	22
21	T2	22	24
22	T2	18	16

Se cere să se testeze cu testul **t-Student**, dacă:

1. Există diferenţă semnificativă de scor mediu între cele două grupe, înainte de tratament
2. Există diferenţă semnificativă de scor mediu între cele două grupe, după tratament
3. Există diferenţă semnificativă de scor mediu la pacienţii din grupa T_1, înainte şi după tratament
4. Există diferenţă semnificativă de scor mediu la pacienţii din grupa T_2, înainte şi după tratament

Indicaţie. Se introduc datele din tabelul de mai sus într-o foaie Excel, pe coloanele A, B, C şi D de la liniile 1 la 23. Pentru 1 şi 2 se foloseşte testul **t-Student** eşantioane cu dispersie inegală, iar pentru 3 şi 4 se foloseşte testul **t-Student** măsurători pereche.

De fiecare dată, trebuie avut grijă să se precizeze corect intervalele de pe foaia Excel unde se găsesc valorile necesare testului. De exemplu, la punctul 1, valorile se găsesc în intervalele C2:C14 şi C15:C23. La punctul 3, intervalele sunt C2:C14 şi D2:D14.

4. Două laboratoare primesc 10 şi respectiv 12 mostre din acelaşi material la care trebuie să determine concentraţia unei substanţe. Concentraţia reală a substanţei este de 43%. Ele dau următoarele rezultate: L1: 43.5, 42.8, 43.5, 42.6, 44.1, 42.3, 42.5, 43.4, 44.2, 43.1, iar L2: 43.6, 42.7, 43.4, 42.5, 44.2, 42.4, 42.4, 43.5, 44.3, 43.2, 43.8, 44.2. Stabiliţi cu ajutorul unui test statistic dacă există o diferenţă semnificativă între mediile obţinute de cele două laboratoare.

5. Calculaţi intervalul de 99% pentru tensiunea pacienţilor din tabelul CARDIO (TAMAXI), luând $t_{99\%}$=2.30

6. Fie seria de numere: 32.1, 33.2, 34.4, 33.1, 35.7, 32.9, 36.5, 31.4, 33.9, 33.7, 34.4, 33.7, 35.4, 34.8, 35.3, valorile fiind extrase dintr-o populaţie cu distribuţie Gauss. Să se calculeze media, deviaţia standard şi coeficientul de variaţie al seriei. Să se calculeze intervalul de încredere de 95% şi intervalul de încredere de 99%. Cum interpretaţi cele două intervale?

7. Fie seriile de numere:
X: 32.1, 33.2, 34.4, 33.1, 35.7, 32.9, 36.5, 31.4, 33.9, 33.7, 34.4, 33.7, 35.4, 34.8, 35.3, 33.7, 33.8, 33.4.
Y: 33.2, 34.2, 35.5, 33.7, 35.2, 33.1, 37.1, 31.3, 33.7, 34.7, 35.4, 34.8, 35.3, 34.5, 35.7, 34.6, 34.7, 34.8, 34.6, 35.6, 36.7, 36.4.

Să se calculeze mediile, deviaţiile standard şi coeficienţii de variaţie pentru cele două serii. Care din cele două serii este mai împrăştiată? Dacă cele două serii de valori sunt extrase din populaţii cu distribuţie Gauss, să se calculeze pentru fiecare intervalul de încredere de 95% şi să se deducă dacă pot proveni din populaţii cu medii egale.