Seminar 6 Corelatii si regresii

Deschideţi fişierul "P6.xls" si salvati-l sub denumirea: "Nume_P6.xls" in directorul dedicat cursului. Fişierul conţine mai multe foi (worksheets)!

0. În worksheet-ul 'Corelatie 0' determinati coeficientul de corelatie Pearson intre lungimea femurului si al humerusului introducand manual formule in celulele indicate:

femur	humerus	$(x-x_{med})$	$(y-y_{med})$	$(x-x_{med})^2$	$(y-y_{med})^2$	$(x-x_{med})\cdot(y-y_{med})$
38	41					
56	63					
59	70					
64	72					
74	84					
x_{med} =	y_{med} =					
n =				s_x =	s_y =	COV =

$$COV_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y}) \qquad r_{xy} = \frac{\frac{1}{n}\cdot\sum_{i}(x_i-\bar{x})\cdot(y_i-\bar{y})}{s_x\cdot s_y} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

I. În worksheet-ul 'Corelatie 1' aveţi date pentru nivelul hormonilor FSH (hormon de stimulare foliculara) si TSH (hormone de stimuilare tiroidiana) pentru un lot de 6 pacienti.

1. Folosiţi opţiunea "chart" pentru a crea un grafic tip"scatter" a nivelului TSH în funcţie de nivelul FSH. Daţi nume adecvate graficului şi axelor. Răspundeţi la urmatoarele întrebări privind natura relaţiei dintre cele două variabile:

   - există o corelaţie între cele două variabile?
   - corelaţia este liniară?

2. Folosind programul EXCEL, calculaţi coeficientul Pearson $\mathbf{r_{xy}}$ pentru datele indicate. *Există două funcţii: CORREL şi PEARSON, care dau exact acelaşi rezultat!*

3. Calculaţi coeficientul Spearman pentru aceleasi date. Pentru aceasta trebuie să urmaţi următorii paşi:

- Calculaţi ordinul "rank" pentru datele din fiecare set: într-o parte goală a worksheet-ului folosiţi funcţia RANK pentru a calcula ordinal corespunzator fiecarei valori FSH. Efectuaţi aceiaşi paşi pentru TSH.

Atentie: programul EXCEL atribuie greşit ordinul datelor prea apropiate!. (Ex: *pentru datele 3; 4; 5; 5; 6 - ordinul perechii de date "5" trebuie să fie 3,5!*)

Pentru a calcula corect coeficientul Spearman r_s, ordinele datelor apropiate trebuie să fie ajustate manual!

- Calculaţi pătratul diferenţelor dintre fiecare pereche de ordine.
- Folosiţi funcţia SUM pentru a însuma pătratele diferenţelor calculate. (Acest pas şi cel anterior se pot face într-unul singur dacă se foloseşte funcţia SUMXMY2).
- Calculaţi r_s folosind formula $r_s = 1 - \frac{6 \cdot \sum d^2}{n^3 - n}$

Indicaţie: puteţi folosi funcţia COUNT pentru a determina mărimea eşantionului automat.

4. Testaţi semnificaţia fiecărei statistici calculate pentru întrebările 1 şi 2 considerând 5% nivel de încredere. Enunţaţi cele două ipoteze clar, precum tipul de test-t folosit ("one-tailed" sau "two-tailed").

- Pentru coeficientul Pearson r_{xy}, calculaţi t_{calc} folosind formula $t_{calc} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$.

- Calculaţi probabilitatea asociată valorii t_{calc}, pentru gradul de libertate n-2, folosind funcţia TDIST.

- Deoarece mărimea eşantionului este mai mică decât 10, pentru testarea semnificatiei statistice a coeficientul Spearman r_s, trebuie folosit comparata valoarea r_{Scalc} cu r_{Scrit}, folosind tabelul adecvat (se găseşte în curs).

II. Determinarea valorilor FSH si TSH s-a facut pentru inca un lot de pacienti. Datele obţinute se găsesc în worksheet-ul 'Corelatie 2'.

- Trasaţi noul grafic şi repetaţi calculele de la punctele 1 - 4 pentru setul de date mai larg.

Avertizare: în noile date sunt ordine apropiate pe care EXCEL nu le calculează adecvat, deci trebuie ajustate ordinele calculate de EXCEL. Acest lucru se poate face în modul următor:

- Să presupunem că ordinul pentru FSH se găseşte în domeniul: D3:D20 (când urmaţi instrucţiunile de mai jos folosiţi referinţa adecvată datelor proprii).

- Într-o zonă goală a worksheet-ului tastaţi următoarea formulă:

=D3+ (COUNTIF(D$3:D$20,D3)-1)/2

Cu această formulă se adaugă 0,5 ordinelor corespunzătoare la două datelor apropiate, 1,0 pentru ordinele corespunzătoare la trei date apropiate, etc.

- Copiaţi celula în jos la următoarele 19 rânduri pentru a calcula ordinul *real* pentru fiecare valoare a debitului.
- Repetaţi paşii de mai sus pentru datele TSH.
- Continuaţi ca de obicei pentru a calcula coeficientul de corelaţie Spearman.

Comparaţi diferenţele dintre rezultatele celor două teste înainte şi după eşantionul adiţional, ţinând cont de puterea lor relativă şi de presupunerile ce stau la baza fiecărei statistici.

III.

A. Să se determine ecuaţia dreptei de regresie (prin metoda celor mai mici pătrate) pentru seturile de date ce reprezintă valoarea calciului seric (y) şi valoarea parathormonului (x) din worksheet-ul “regresie 1”.

B. Determinati valoarea coeficientului de regresie si apoi folosind formulele date la curs determinati ecuatia dreaptei de regresie.

IV. Pentru 9 pacienti cu anemie aplastica se evalueaza procentul de reticulocite si numarul limfocitelor (worksheet-ul “regresie 2”).

a) Trasati dreapta de regresie dintre procentul de reticulocite si numarul limfocitelor.

b) Calculati coeficientul de corelatie si coeficientul de determinare, si interpretati nivelul de corelatie.

c) Calculati variatia reziduala.

Aplicati analiza "Regression" din "Data analysis" si incercati sa intelegeti cat mai multe din datele afisate.

V. Folosind tabelul din worksheet-ul “regresie 3” calculati:

a) coeficientul de corelatie dintre durata spitalizarii si varsta

b) cea mai buna relatie liniara dintre durata spitalizarii si varsta.

c) verificati semnificatia statistica a acestei relatii

Pentru a raspunde la aceste intrebari puteti aplica analiza "Regression" din "Data analysis"