Lab Nr. 5, Probability and Statistics

Statistical measures; Correlation and regression; Confidence intervals for means and variances

!!!!!!!!! PRECIZARE: la lab 5, se incepe cu partea B (intervale de incredere) si, doar daca ramane timp, se face si partea A (corelatie si regresie). Eu la unii am reusit, la altii, nu... Intervalele de incredere sunt mai importante, de acolo vor primi probleme la examenul practic. Formulele aferente sunt in fisierul

`conf_int.pdf`.

A. Correlation and Regression

Cateva chestiuni pe care insist eu la acest laborator.

In primul rand, le reamintesc pe scurt despre ce e vorba: **corelatia** ne spune daca exista vreo legatura intre doua variabile (sau si mai multe, dar noi ne rezumam la cazul a doua variabile), iar **regresia** reprezinta metoda statistica prin care se gaseste o astfel de legatura. E important de stiut daca doua variabile (cantitati, caracteristici) sunt sau nu corelate, pentru ca, in cazul ca sunt, informatii despre una dintre ele ne pot oferi informatii si despre cealalta. Legatura dintre doua variabile e data de **curba de regresie**, adica functia care ne spune in ce fel una dintre ele depinde de cealalta.

Au fisierul de statistica descriptiva pe server ($descr-stat.pdf$), le spui sa-l aiba deschis tot timpul (se afla in acelasi director unde sunt si enunturile laboratoarelor). In general, curba de regresie reprezinta felul in care valoarea medie a unei caracteristici depinde de valorile celeilalte caracteristici. Se pot cauta curbe de regesie de diferite forme, cea mai raspandita si mai la indemana o reprezinta o curba de regresie liniara, adica **dreapta de regresie**. Deci cautam sa vedem daca exista o legatura liniara intre cele doua caracteristici, fie pozitiva (caz in care cresc in acelasi timp), fie negativa (cand una creste, cealalta descreste). Daca o astfel de legatura exista, sau mai bine zis un astfel de *trend* liniar exista, acest lucru se va vedea grafic: Norul statistic (in engleza "scatterplot"), adica punctele de coordonate (x_i, y_i), se afla strans in apropierea dreptei de regresie, chiar daca nu toate punctele se afla chiar pe dreapta de regresie. Si le pori indica exemplele din fisier, cele 4 de la paginile 23-24. Statistica (virgula!) care masoara acest trend liniar, cat de puternica sau de slaba este legatura liniara dintre cele doua caracteristici, este **coeficientul de corelatie** (e vorba despre coeficientul de corelatie al lui Pearson, exista si altii, dar asta este cel mai uzual, cand nu se pomeneste numele, despre asta e vorba). Acesta are o anumita formula de calcul (in care intra mediile, variantele si covarianta), (le indici in text unde e definita) iar felul in care masoara trendul liniar este urmatorul: variaza intre $-1 \leq \rho \leq 1$, si cu cat valorile sunt mai mari, mai apropiate de extremele ± 1, cu atat legatura liniara e mai puternica, la $+1$ legatura perfect liniara pozitiva (dreapta de regresie cu panta pozitiva, deci crescatoare), la -1 legatura perfect liniara negativa (dreapta de regresie cu panta negativa, deci desccrescatoare). Daca valorile lui ρ sunt apropiate de 0, inseaman legatura liniara slaba. Daca $\rho = 0$, inseamna ca variabilele sunt *necorelate*, ceea ce nu inseamna ca sunt

independente. Nu exista legatura liniara, dar poate exista o altfel de legatura si deci dependenta exista. Independenta e mai puternica decat necorelarea. Din nou, pot revedea exemplele de acolo.

Bun, despre asta e vorba in prima parte din laboratorul de azi. Ni se dau doua caracteristici si vrem sa studiem corelatia dintre ele, respectiv daca exista legatura liniara sau nu. Vrem sa vedem asta teoretic (valoarea coeficientului de corelatie).

• **Introducerea datelor**

Sunt date doua caracteristici, X si Y si sunt date sub forma valoare-frecventa. Deci cifrele din randul doi reprezinta frecventele celor din prima linie, NU sunt si ele valori efective. Deci 20 apare de doua ori, 21 o data, 22 de 3 ori, etc. Le vom introduce fiecare ca un vector (linie sau coloana). Sigur, nu sunt repetitii multe, deci am putea da copy-paste, dar exista o modalitate mai simpla: folosind `ones`. `ones(m,n)` produce o matrice de $m \times n$ de "1", iar cu acestea putem produce matrici de alte valori:

```
x= [20*ones(1,2),21,22*ones(1,3),23*ones(1,6),...]
```

Aici am introdus valorile lui X ca vector linie - prima dimensiune la ones e 1 si, in continuare am mers pe aceeasi linie, deci cu virgula sau spatiu, daca puneam ";", se trecea pe o alta linie si nu se potriveau dimensiunile (dadea eroare). Te rog sa nu le scrii tu tot sirul pe tabla, ci doar inceputul (cum am scris eu mai sus), mai departe trebuie sa se primda singuri. Introducand datele in acest fel, putem verifica usor daca le-am introdus corect, pentru ca la orice prelucrare statistica, prima grija majora e introducerea CORECTA a datelor! O alta chestiune ce trebuie mentionata aici: cand e un sir (vector) lung de date si vrem sa le vedem mai bine (pe toate impreuna), se pot pune "..." la sfarsit si Matlab va sti automat ca urmatoarea e linie de continuare (o va si pune mai la dreapta).

```
X= [20*ones(1,2),21,22*ones(1,3),23*ones(1,6),...
    24*ones(1,5),25*ones(1,9),26*ones(1,2),...
    27*ones(1,2)];
```

Atentie doar ca sa fie lasat spatiu sau virgula, primul punct sa nu fie dupa o cifra, sa poata fi interpretat ca punct zecimal si sa dea eroare.

In continuare calculam toate statisticile necesare. Te rog sa atragi atentia si la notatii, sa se obisnuiasca cu ele, $\overline{X}$ e notatia pentru media variabilei X.

• **Calculul mediilor**

Se face cu `mean`, gasesc comanda in fisierul de teorie, s-a folosit la laboratorul anterior, sau pot sa-i dea *help mean*, daca nu mai tin minte. Pe urma, toate aceste valori trebuie afisate frumos, adica inteligibil, sa se inteleaga ce reprezinta, nu doar niste numere pe ecran.

```
fprintf('a) the means are: mx=%6.3f, my=%6.3f\n',mx,my)
```

Nu trebuie neaparat sub forma asta, sau cu formatul respectiv (evident sunt numere reale, nu intregi), dar sa apara si cuvinte explicative (in engleza)

• **Calculul variantelor**

Se face cu `var`. Aici trebuie facuta o observatie importanta: Varianta se calculeaza ca

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

atunci cand datele reprezinta o **populatie** (adica sunt datele in totalitate, in integritate), si cu

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

cand datele reprezinta o **selectie** ("sample"), adica doar o parte din intreaga populatie. Motivul se va vedea in capitolul urmator (teoria estimatiei), deocamdata trebuie doar retinuta mica diferenta in definitie. A se observa si notatia diferita! Acum, datele noastre ce reprezinta, populatie sau selectie? Pai zice astea sunt datele, nu spune ca provin din ceva mai mare. NU stiu ce reprezinta, dar astea sunt, in totalitate, deci reprezinta o populatie (de aceea am si folosit in enunt notatia potrivita). Acum, nu trebuie ei sa le calculeze, face asta `var`. Sa vedem cum. Ii dam un *help*.

```
var Variance.
    For vectors, Y = var(X) returns the variance of the values
    in X.  For matrices, Y is a row vector containing the
    variance of each column of X.  For N-D arrays, var operates
    along the first non-singleton dimension of X.

    var normalizes Y by N-1 if N>1, where N is the sample size.
    This is an unbiased estimator of the variance of the
    population from which X is drawn, as long as X consists of
    independent, identically distributed samples. For N=1,
    Y is normalized by N.

    Y = var(X,1) normalizes by N and produces the second moment
    of the sample about its mean.  var(X,0) is the same as var(X).
    ...
```

N-am reprodus totul, doar partea care ne intereseaza si anume, `var(x)` va imparti suma la $n-1$, deci calculeaza s^2, varianta de selectie, iar `var(x,1)` imparte suma la n, adica σ^2, varianta de populatie. In cazul nostru, o folsim pe a doua.

Inca o observatie importanta. Notatia variantei cu patrat nu e intamplatoare, din moment ce e o cantitate pozitiva, deci patratul a ceva. Acel ceva este **deviatia standard**

(standard deviation). Aceasta e, de asemenea, implementata in Matlab. Functia `std` calculeaza pe σ, respectiv s, avand aceleasi optiuni (cu "1" sau fara) ca si `var`. Sa-i dea un help sa vada. Si inca ceva, legat de de asta. Si matalab-ul calculeaza deviatia standarad asa cum am face-o noi de mana, adica mai intai varianta si pe urma ia radicalul ei. Deci, in formulele statistice in care va aparea σ^2 sau s^2, NU se foloseste NICIODATA

`std(x)^2,`

cu sau fara "1", (pentru ca fac doua operatii inverse si introduc erori degeaba!), ci se foloseste functia `var`! Sa le atragi atentia! (eu le si spun ca daca vad asa ceva la examenul practic, ii pic sigur!!...)

• **Calculul covariantei**

Se face cu functia `cov`. Pe langa faptul ca are aceleasi optiuni (de selectie, fara "1", de populatie, cu "1"), mai are o particularitate. Cand calculeaza `cov(x,y,1)`, inainte de a tipari rezultatul, sa lase comanda fara ';". Vor observa ca e o matrice de 2×2. Asta pentru ca sunt 2 vectori. Deci `cov` e de fapt, matricea covariantelor, adica variantele fiecarui vector cu fiecare. Intrebare (pentru ei, ii lasi putin sa se gandeasca): ce va fi pe diagonala principala? Evident, vor fi chiar variantele vectorilor. Deci, variantele lui X si Y puteau fi luate de aici, fara a le mai calcula individual (asta nu inseamna sa schimbe). Oricum, deci, pentru covarianta, ii dau un nume matricii respective si de afisat (cu `fprintf`) afiseaza din matricea respectiva ce e pe pozitia $(1, 2)$ (sau $(2, 1)$, caci, evident, matricea e simetrica). Pe urma, pot pune ';" inapoi la matrice

Inca ceva: sa nu dea acestor variabile nume deja ocupate in Matlab! Sa nu-i spuna "mean", sau"var", sau "cov"!!!

• **Calculul coeficientului de corelatie**

Se face cu functia `corrcoef`. Aceasta nu mai are optiunea cu sau fara "1", pentru ca se observa din definitie

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

ca indiferent daca numitorul variantelor si covariantei e n sau $n - 1$, acesta se reduce.

Dar, e tot o matrice, functioneaza ca si `cov`. Ce va fi pe diagonala principala? Evident, vad si ei ca doar 1, dar de ce? pentru ca e coeficientul de corelatie al unei variabile cu ea insasi! Ce masoara ρ? Legatura liniara. Cat de liniara e legatura dintre X si el insusi? $X = 1 \cdot X + 0$, deci legatura perfect liniara pozitiva. La fel ca mai inainte, se afiseaza din matricea respectiva ce e pe pozitia $(1, 2)$ sau $(2, 1)$. Observam ca valoarea e $\rho = 0.964$, o valoare foarte mare, apropiata de 1, care sugereaza o puternica legatura liniara.

B. Confidence intervals

Cateva lucruri pe care le poti spune la acest laborator. In primul rand, sa ne reamintim pe scurt despre ce e vorba:

Cum se pune problema: avem o caracteristica de populatie, X, al carei pdf (probability density function), $f(x;\theta)$, depinde de un parametru necunoscut, θ. Acesta se numeste *target parameter*. Acest parametru trebuie estimat. Cum? Pe baza unei selectii de volum n (sample of size n). Avem o selectie de volum n, asta insemnand ca avem n *variabile de selectie*, adica $X_1, X_2, \ldots, X_n$, independente si identic distribuite (adica au acelasi pdf) cu X (deci toate au acelasi pdf, $f(x;\theta)$). Putem folosi un estimator punctual (*point estimate*), adica o functie de selectie, $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$, sau *an interval estimate*, adica gasim doua valori $\overline{\theta}_L, \overline{\theta}_U$ (de la lower and upper), astfel incat

$$P\left(\theta \in (\overline{\theta}_L, \overline{\theta}_U)\right) = 1 - \alpha.$$

Terminologie:
$-$ $\left(\overline{\theta}_L, \overline{\theta}_U)\right)$ e intervalul de incredere de $100(1-\alpha)\%$ (*confidence interval*);
$-$ $1-\alpha$ e *confidence level or confidence coefficient* (care se da);
$-$ α e *significance level* (va avea relevanta in capitolul urmator, la testarea ipotezelor statistice)

Un interval de incredere NU este unic determinat de conditia de mai sus! Poti eventual sa spui (foarte pe scurt!!!, ca le fac eu la curs) ca, in general (si asta facem si noi), se foloseste asa-numita metoda pivotala: pivotul e o variabila aleatoare (o statistica) W, care depinde de datele de selectie (si, eventual, de alte lucruri cunoscute) si de θ (deci, singura necunoscuta e θ) si al carei pdf e cunoscut si NU depinde de θ. Atunci se folosesc cuantilele potrivite astfel incat

$$P\left(w_{\alpha/2} < W < w_{1-\alpha/2}\right) = 1 - \alpha.$$

Pe urma, se inlocuieste expresia variabilei W si se rescrie inegalitatea pana ajunge sa fie

$$P\left(\overline{\theta}_L < \theta < \overline{\theta}_U\right) = 1 - \alpha.$$

Asa au fost obtinute formulele din fisierul $conf - int.pdf$.

1. Avem CI pentru estimarea unui parametru al unei populatii, si anume $\theta = \mu$ (media de populatie sau media teoretica), respectiv $\theta = \sigma^2$ (varianta de populatie). Pentru medie, sunt doua situatii, σ (deci de populatie) cunoscut sau nu. In fisier sunt date toate cazurile si pentru fiecare, e data variabila-pivot si distributia ei, casa poata fi calculate cuantilele si, deci, limkitele intervalului de incredere. Restul explicatiilor sunt in fisierul Matlab.

2. Acum vrem sa comparam parametrii de la doua populatii, si anume:
$-$ compararea mediilor de populatie, cand $\theta = \mu_1 - \mu_2$;
$-$ compararea variantelor de populatie, cand $\theta = \sigma_1^2/\sigma_2^2$.

Aici avem doua caracteristici de populatie (doua medii de populatie, doua variante de populatie), doua selectii, cu mediile lor, cu variantele lor, etc. (citeste putin din curs). Eventual spune putin despre varianta centralizata (*pooled variance*), care calculeaza o varianta considerand datele din ambele selectii (acea s_p^2). Exista 3 cazuri pentru $\theta = \mu_1 - \mu_2$ si unul pentru $\theta = \sigma_1^2/\sigma_2^2$ (de la teorie). Am pus exemple doar pentru cazurile 2 si 3 de la medii. Nu stiu daca va mai fi timp si pentru punctul c), de aceea l-am pus in paranteza. Chiar daca nu se mai face, se poate doar discuta cum s-ar face, similar cu celelalte.