Probabilitati & Statistica

Documentatie set de date „stackloss”

In setul de date „stackloss” gasim inregistrari pentru *Air Flow, Water Temp., Acid Conc.* si *Stack Loss,* masurate in cele 21 de zile ale experimentului. Acesta a avut scopul de a observa oxidarea amoniacului in acid nitric.

1. Air Flow – rata de lucru a plantei
2. Water Temp. – temperatura apei la intrarea in turnul de absorbtie
3. Acid Conc. – concentratia acidului in turn (-50, inmultit cu 10; de exemplu valoarea 89 corespunde 58.9% acid)
4. Stack Loss (variabila dependenta de celelalte) – de 10 ori procentul amoniacului consumat de planta care nu ajunge in turnul de absorbtie (altfel reprezinta inversul eficientei generale a plantei)

Problema 1

Calculam media si varianta pentru fiecare variabila in parte folosind functiile *mean,* respectiv *var*, care iau ca parametru masuratoarea corespunzatoare.

Pentru a calcula quartilele unei variabile folosim functia *quantile*(). Nu este necesara sortarea valorilor deoarece este facuta automat de catre functie.

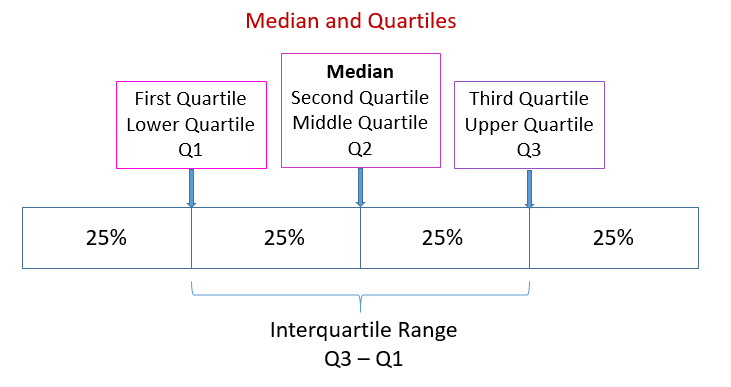
Boxplot-urile pentru fiecare dintre cele 4 masuratori:

* Air Flow – mediana = 58.00
* Water Temp. – mediana = 20
* Acid Conc. – mediana = 87
* Stack Loss – mediana = 15

→ Justificari:

Putem observa uitandu-ne la boxplot-uri ca:

* Air Flow, Stack Loss: in comparatie cu Water Temp si Acid Concentration, valorile sunt mai apropiate, dar observam si ca sunt singurele cu outlier-e (ambele au 2, dintre care 1 extrem)
* Prin pozitia medianei, in cazul variabilei Air Flow, valorile sunt mai apropiate de prima quartila.



Problema 2

La aceasta cerinta am avut de facut o regresie simpla si o regresie multipla. Apoi, am avut de adaugat la setul de date initial una sau mai multe variabile pe care sa le consideram potrivite a fi incluse in cel puţin un model de regresie.

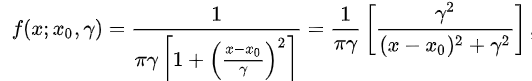
Am ales ca variabila raspuns "stack.loss" deoarece reprezinta scopul experimentului, si evident depinde de celelalte trei variabile inregistrate.

stack.loss = β0 + β1\*Air.Flow + β2\*Water.Temp + β3\*Acid.Conc. + ε

Am introdus variabila noua "Water.vol", cu date generate prin repartitie uniforma randomizata runif(21, 0.2, 0.8), si am folosit-o pentru o alta regresie in care stack.loss va depinde de aceasta.

Dintre cele 2 modele de regresie este mai potrivita cea multipla, deoarece intra in calcul un set mai mare de date, astfel obtinand o acuratete mai mare si o reprezentare optima a regresiei.

Problema 3

 Am ales sa utilizam repartitia Cauchy. In limbajul R, pentru functia de densitate folosim in acest caz dcauchy, iar pentru functia de repartitie, pcauchy. Functia de densitate are forma:

unde x0 este pozitia „varfului” repartitiei, iar γ este parametrul de proportie.

Repartitia Cauchy este un exemplu de repartitie „patologica” (are proprietati neregulate), deoarece media si varianta sunt nedefinite. Este una dintre repartitiile pentru care functia de densitate poate fi reprezentata analitic, celelalte fiind repartitia normala si repartitia Levy.

 Observam cazul special, pentru x0 = 0 si γ = 1, repartitia se mai numeste repartitia standard Cauchy, iar functia de densitate devine:

Bonus

La problema au participat toti 3 membrii echipei.