ENSEM 2ème année 2020

Devoir de Statistiques

Ce travail est à rendre pour le 27 mai 2020. Si vous utilisez Matlab, vous inclurez, dans votre rapport, les lignes de codes utilisées. Vous travaillerez par groupe, de 2 (minimum) à 5 (maximum) étudiants. Une seule copie par groupe, me sera retournée, via ARCHE. Vous indiquerez aussi la ou les filière(s) d'appartenance du groupe (EMSYS, GENESE, ISN, SINERGIE).

Présentation du problème

Le phosphore est un élément chimique qui fait perdre à l'acier sa ductilité, le rendant fragile. L'acier perd alors tout intérêt, en devenant extrêmement cassant. Il existe deux méthodes pour doser le phosphore dans l'acier d'une coulée :

— méthode 1 : dosage par voie chimique,

— méthode 2 : dosage par méthode spectrographique.

On notera

— Y la quantité réelle de phosphore dans l'acier d'une coulée,

— Y_1 la quantité de phosphore dans l'acier d'une coulée, mesurée par la méthode 1,

— Y_2 la quantité de phosphore dans l'acier d'une coulée, mesurée par la méthode 2.

Pour toutes les coulées effectuées pendant une certaine période, on a réalisé les deux mesures. Mais, les méthodes de mesures altèrent la coulée : après une mesure, par l'une ou l'autre des méthodes, la coulée n'est plus utilisable. Ainsi, sur une coulée, on mesure le phosphore soit par la méthode 1, soit par la méthode 2, mais pas par les deux. C'est pour cela que vous n'avez pas nécessairement le même nombre de mesures Y_1 que de mesures Y_2.

On suppose aussi que

$$\begin{aligned} Y_1 &= Y + \varepsilon_1 \\ Y_2 &= Y + \varepsilon_2 \end{aligned}$$

où $\varepsilon_1, \varepsilon_2$ sont des termes d'erreurs, aléatoires et indépendants de Y, de lois $\mathcal{N}(0, \sigma_1^2)$ et $\mathcal{N}(0, \sigma_2^2)$.

Données

Les données se trouvent sur ARCHE, dans le fichier 'TPStat.xlsx'. Les unités des données sont arbitraires.

Pour le calcul de quantiles des lois gaussienne, de Student, χ^2 et Ficher, vous pourrez vous aider des fonctions suivantes (ou bien prendre vos tables du formulaire) :

— Pour $X \sim \mathcal{N}(\mu, \sigma^2)$ une variable aléatoire de loi normale de moyenne μ, et écart-type σ, si on veut x_α tel que $\mathbb{P}(X < x_\alpha) = 1 - \alpha$, on trouve x_α en faisant `icdf('norm', 1-α, μ, σ)`.

— Pour $T \sim \mathcal{T}(\nu)$ une variable aléatoire de loi de Student à ν degrés de liberté, si on veut t_α tel que $\mathbb{P}(T < t_\alpha) = 1 - \alpha$, on trouve t_α en faisant `icdf('t', 1-α, ν)`.

— Pour $X \sim \chi^2(\nu)$ une variable aléatoire de loi du χ^2 à ν degrés de liberté, si on veut χ_α tel que $\mathbb{P}(X < \chi_\alpha) = 1 - \alpha$, on trouve χ_α en faisant `icdf('chi2', 1-α, ν)`.

— Pour $F \sim \mathcal{F}(\nu_1, \nu_2)$ une variable aléatoire de loi de Fisher à ν_1 et ν_2 degrés de liberté, si on veut f_α tel que $\mathbb{P}(F < f_\alpha) = 1 - \alpha$, on trouve f_α en faisant `icdf('f', 1-α, ν1, ν2)`.

Questions

1. Estimations. Donner l'estimation des moyennes et des écart-types des variables Y_1 et Y_2.
2. Lois.
   (a) Supposons que $Y \sim \mathcal{N}(m, \sigma^2)$, quelles devraient alors être les lois de Y_1 et Y_2 ? Pourquoi ?
   (b) Vérifiez votre réponse à l'aide d'un histogramme et d'un qqplot (à commenter), puis à l'aide d'un test (pour le test, compte tenu que cela peut être long, ne mettre dans le compte-rendu, que votre travail pour Y_1). *PS : le code Matlab* `[n,xout] = hist(Y1,7)` *permet de couper l'échantillon des* Y_1 *en 7 classes,* `n` *représente les effectifs observés dans chaque classe, et* `xout` *les milieux des classes.*
   (c) A quoi cela sert-il de savoir les lois de Y_1 et Y_2 ?
3. Comparaison des 2 méthodes.
   (a) Donner les intervalles de confiance à 95% des moyennes et écart-types de Y_1 et Y_2.
   (b) Au vu des résultats précédents, peut-on considérer que les deux méthodes donnent des mesures égales ? Confirmez votre avis (ou pas), avec un test. On prendra un risque de 5%.
   (c) Si il n'y a pas de différence entre les méthodes de mesure, donner une nouvelle estimation de la moyenne et de l'écart-type des quantités de phosphore mesurées. Sinon, on continuera de travailler avec les mesures de la méthode 1 seulement.
4. Une coulée est de bonne qualité si sa quantité de phosphore est inférieure à m_0=0.055. Peut-on considérer que les coulées sont, en moyenne, de bonne qualité ? Répondre d'abord par l'examen des données, graphiquement et avec un intervalle de confiance. Puis confirmer (ou pas) votre avis avec un test. On prendra un risque de 5%.
5. Ajout de chaux.
   (a) Au cours du procédé de fabrication de l'acier, on rajoute de la chaux afin de réduire la quantité de phosphore. Plusieurs mesures on été faites (avec la méthode 1), pour différentes quantités x de chaux (3ème série de données). Représenter les points sur un graphique. Une relation linéaire semble se dessiner :
   $$Y_1(x) = \alpha + \beta x + \varepsilon$$
   entre la quantité de chaux x, et la quantité correspondante de phosphore $Y_1(x)$. Estimer les paramètres α et β. Sur le graphique, ajouter la droite de régression, l'intervalle de confiance de la droite, et des observations.
   (b) La régression est-elle de bonne qualité ? Le modèle de régression peut-il être utilisé pour faire de la prédiction ?
   (c) Pour une quantité de chaux $x = 100$, on a obtenu une nouvelle mesure $Y_1 = 0.01$. Cette mesure est-elle en accord avec la régression de la question précédente ?
   (d) Dans cette question, on supposera que α et β de l'équation $Y_1(x) = \alpha+\beta x+\varepsilon$ sont remplacés par leurs estimations, notées a, b. Une quantité x de chaux coûte à l'entreprise : $r \cdot x$ euros. Lorsque la coulée a une quantité de phosphore supérieure à $m_1 = 0.07$, la coulée est inutilisable et jetée. Sinon, elle est vendue à un prix $p(1 - Y_1(x))$. A x fixé, exprimer le gain de l'entreprise, et la probabilité que le gain soit positif (en fonction de m_1, r, x, p et des paramètres de la régression). Pour quelle valeur de x cette probabilité est-elle maximale ? On prendra $r = 1; p = 5$. On peut utiliser la fonction `cdf` de Matlab : pour calculer la probabilité $\mathbb{P}(Z < z)$ où $Z \sim \mathcal{N}(\mu, \sigma^2)$, on écrit `cdf('norm', z, `$\mu$`, `$\sigma$`)`.