

# Statistique - Contrôle Terminal du 13 mai 2024 - Correction

## L1 Mathématiques - L1 Informatique - LAS - L2 3PE

*Avertissements.* Les exercices peuvent être traités indépendamment, les données de l'exercice 1 pouvant aussi servir dans les exercices 2 et 3. À l'exception des valeurs exactes qui ne nécessitent pas d'arrondi à calculer, les résultats numériques seront donnés à  $10^{-3}$  près au minimum. Dans votre rédaction, le raisonnement compte au moins autant que le résultat final. Détaillez vos calculs.

**Exercice 1. (12 points)** Une usine normande produit du beurre sur 3 sites différents, que nous nommerons "site 1", "site 2" et "site 3". Elle contrôle le taux d'humidité régulièrement. La norme pour le taux d'humidité de ce type de beurre est d'être dans l'intervalle [16; 18].

1.1. Soit  $A_1$  l'événement "la mesure vient du site 1",  $A_2$  l'événement "la mesure vient du site 2",  $A_3$  l'événement "la mesure vient du site 3" et  $B$  l'événement "la mesure est hors de la norme". On sait que les 3 sites produisent du beurre dans des proportions respectives de 20%, 45% et 35%, c'est-à-dire que  $p(A_1) = 0.2$ ,  $p(A_2) = 0.45$  et  $p(A_3) = 0.35$ . De plus, en analysant les prélèvements hors norme sur une longue période, on a observé que  $p_1 = p(B/A_1) = 0.25$ ,  $p_2 = p(B/A_2) = 0.4$ ,  $p_3 = p(B/A_3) = 0.15$ . Déterminer  $p(B)$ .

**Correction 1.1.** On utilise la formule des probabilités complètes :  $p(B) = p(B/A_1)p(A_1) + p(B/A_2)p(A_2) + p(B/A_3)p(A_3) = 0.25 * 0.2 + 0.4 * 0.45 + 0.15 * 0.35 = 0.2825$ .

1.2. Soit un prélèvement, qui a donné une valeur  $x = 18.02$ . Quelle est la probabilité pour que ce prélèvement provienne du site 1 ?

**Correction 1.2.** L'observation est un événement  $B$ , puisqu'elle est hors norme. La question revient donc à déterminer  $p(A_1/B)$ . On utilise la formule de Bayes :  $p(A_1/B) = \frac{p(A_1 \cap B)}{p(B)} = \frac{p(B/A_1)p(A_1)}{p(B)} = \frac{0.25 * 0.2}{0.2825} = 0.17699$ .

La mesure de 12 prélèvements de beurre sur chacun des 3 sites a donné les résultats suivants :

no prélèvement	1	2	3	4	5	6	7	8	9	10	11	12	moyenne	moyenne des carrés
Humidité site 1	17.02	16.79	16.33	16.63	15.04	16.69	15.49	17.94	14.75	17.24	17.41	17.65	16.5816	275.8969
Humidité site 2	15.42	17.03	16.34	17.19	14.86	16.96	15.78	18.13	14.88	17.20	17.63	18.11	16.6275	277.7082
Humidité site 3	17.30	17.33	16.54	17.15	16.51	16.69	15.94	18.09	16.33	17.51	17.58	17.76	17.0608	291.4603

Ces valeurs sont supposées issues de l'observation de variables aléatoires indépendantes et identiquement pour chaque site. Celles du site 1 sont supposées issues d'observations d'une loi normale  $N(\mu_1, \sigma_1^2)$ , celles du site 2 d'une loi normale  $N(\mu_2, \sigma_2^2)$ , et celle du site 3 d'une loi normale  $N(\mu_3, \sigma_3^2)$ .

1.3. Calculer les proportions de prélèvements hors norme sur chaque site  $f_1$ ,  $f_2$  et  $f_3$ .

**Correction 1.3.** Il y a 3 valeurs hors de l'intervalle [16; 18] sur le site 1, 6 valeurs sur le site 2 et 2 valeurs sur le site 3, cela donne  $f_1 = \frac{3}{12} = 0.25$ ,  $f_2 = \frac{6}{12} = 0.5$  et  $f_3 = \frac{2}{12} = 0.1667$ .

1.4. On souhaite contrôler si la valeur nominale de  $p_1$  n'a pas bougé. Sur la base des valeurs observées sur le site 1, calculer les bornes de l'intervalle de confiance à 95% pour  $p_1 = p(B/A_1)$ .

**Correction 1.4.** On utilise la formule  $IC = [f_1 - me, f_1 + me]$ , avec  $f_1 = 0.25$  calculée à la question précédente, et  $me = z_{\alpha/2} \sqrt{\frac{f_1(1-f_1)}{n}} = 1.9600 * \sqrt{\frac{0.25*(1-0.25)}{12}} = 0.2450$ . D'où l'IC :  $[0.25 - 0.2450; 0.25 + 0.2450] = [0.0050; 0.4950]$ .

1.5. Calculer les bornes de l'IC pour  $\mu_1$  au niveau  $1 - \alpha = 0.95$ , en supposant que  $\sigma_1$  est connu et égal à 1.

**Correction 1.5.** On utilise la formule  $IC = [\bar{x}_1 - me, \bar{x}_1 + me]$ , avec  $me = z_{\alpha/2} \frac{\sigma_1}{\sqrt{n}}$ , avec  $\sigma_1 = 1$  et  $n = 12$ . Donc  $me = 1.9600 \frac{1}{\sqrt{12}} = 0.56579$ . D'où l'IC :  $[16.5816 - 0.56579; 16.5816 + 0.56579] = [16.0159; 17.1475]$ .

1.6. Calculer une estimation  $s_1^2$  de la variance  $\sigma_1^2$  basée sur la variance sans biais, grâce aux mesures faites sur le site 1 (on pourra s'aider des calculs des moyennes et des moyennes des carrés données dans le tableau), puis calculer les bornes de l'IC pour  $\mu_1$  au niveau  $1 - \alpha = 0.95$ , en supposant que  $\sigma_1$  est inconnu.

**Correction 1.6.**  $s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2 = \frac{n}{n-1} (\frac{1}{n} \sum_{i=1}^n x_{1,i}^2 - \bar{x}_1^2) = \frac{12}{11} (275.8969 - 16.5816^2) = 1.0336$ .

IC pour  $\mu_1$  au niveau  $1 - \alpha = 0.95$  :  $[\bar{x}_1 - me, \bar{x}_1 + me]$ , avec  $me = t_{\alpha/2, n-1} \frac{s_1}{\sqrt{n}} = 2.200985 * \sqrt{\frac{1.0336}{12}} = 0.64518$ , ce qui donne  $IC = [16.5816 - 0.64518; 16.5816 + 0.64518] = [15.9365; 17.2268]$ .

1.7. Avec un même niveau  $1 - \alpha$ , la marge d'erreur de l'IC pour  $\mu_2$  avec  $\sigma_2$  inconnu serait-elle plus petite ou plus grande que celle de l'IC pour  $\mu_1$ ? Même question avec l'IC pour  $\mu_3$ : sa marge d'erreur serait-elle plus petite ou plus grande que celle de l'IC pour  $\mu_1$ ? Justifier les réponses sans faire de calcul.

**Correction 1.7.** Comme les effectifs des échantillons sur les sites 2 et 3 sont les mêmes que l'effectif du site 1, la seule valeur qui change dans le calcul de la  $me$  pour chaque site est la variance. Or,  $v_1^2 = 275.8969 - 16.5816^2 = 0.9474$ ,

$v_2^2 = 277.7082 - 16.6275^2 = 1.23444$  et  $v_3^2 = 291.4603 - 17.0608^2 = 0.3894$ . Donc la *me* pour l'IC2 est plus grande que celle pour l'IC1, et la *me* pour l'IC3 est plus petite que celle pour l'IC1.

*Remarque* : ce n'est pas parce que la moyenne et la moyenne des carrés sont respectivement plus grandes dans un échantillon que dans l'autre que cela donne une variance plus grande.

### Exercice 2. (6 points)

On a aussi relevé, pour chacun de ces prélèvements, la température de fusion. Voici le tableau de ces relevés.

no prélèvement	1	2	3	4	5	6	7	8	9	10	11	12	moyenne	moyenne des carrés
t_fusion site 1	28.5	29.0	29.0	30.0	31.0	30.0	31.0	28.0	30.5	29.5	29.0	30.0	29.6250	878.4792
t_fusion site 2	31.0	29.5	29.5	29.0	31.0	30.0	30.0	29.0	31.5	29.5	29.0	28.5	29.7917	888.3542
t_fusion site 3	29.5	30.0	29.5	30.0	30.0	30.0	30.0	29.0	30.5	29.5	29.0	29.0	29.6667	880.3333

On note  $(x_1, \dots, x_{36})$  la série des valeurs de l'humidité sur les trois sites, et  $(y_1, \dots, y_{36})$  celle des valeurs des températures de fusion.

2.1. Calculer la moyenne de toutes les températures de fusion (les trois sites confondus), ainsi que la variance avec la formule  $\frac{1}{36} \sum_{i=1}^{36} (y_i - \bar{y})^2$  ou une formule équivalente. Calculer la moyenne de tous les taux d'humidité, les trois sites confondus. On se servira à chaque fois des moyennes données sur chaque site, et des moyennes des carrés pour les températures.

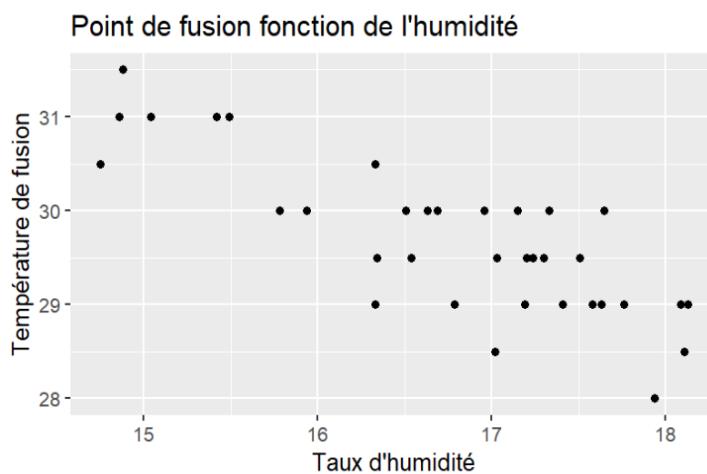
**Correction 2.1.** On a déjà les moyennes par site, il suffit de faire la moyenne pondérée par les effectifs des sites des trois moyennes (mais comme les trois sites ont les mêmes effectifs, c'est comme si on ne pondérait pas) :  $(29.6250 + 29.7917 + 29.6667)/3 = 29.6944$ . Quant à la variance, comme on a les 3 moyennes des carrés, on peut calculer la moyenne de tous les carrés  $\frac{1}{36} \sum_{i=1}^{36} y_i^2 = (878.4792 + 888.3542 + 880.3333)/3 = 882.3889$ . D'où la variance :  $\frac{1}{36} \sum_{i=1}^{36} y_i^2 - (\frac{1}{36} \sum_{i=1}^{36} y_i)^2 = 882.3889 - 29.6944^2 = 0.62886$ .

Moyenne des taux d'humidité de tous les sites :  $(16.5816 + 16.6275 + 17.0608)/3 = 16.7567$ .

2.2. On donne  $\frac{1}{36} \sum_{i=1}^{36} x_i^2 - (\frac{1}{36} \sum_{i=1}^{36} x_i)^2 = 0.9026$  et  $\sum_{i=1}^{36} x_i y_i = 17890.81$ . Calculer le coefficient de corrélation linéaire entre les séries  $X = (x_1, \dots, x_{36})$  et  $Y = (y_1, \dots, y_{36})$ .

**Correction 2.2.** C'est  $\frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{17890.81/36 - 16.75667 * 29.6944}{\sqrt{0.9026 * 0.62886}} = \frac{-0.6123}{\sqrt{0.9026 * 0.62886}} = -0.8137$ .

2.3. Le nuage de points correspondant à ces deux variables est le suivant.



Ce nuage est-il en conformité avec le coefficient de corrélation ? Interpréter cette corrélation.

**Correction 2.3.** La corrélation est fortement négative (assez proche de -1), ce qui signifie que plus le taux d'humidité dans le beurre est élevé, plus la température de fusion est basse. Le nuage montre bien des points répartis autour d'une droite de pente négative, donc oui, le nuage est en conformité avec la corrélation, dont on vient d'interpréter le sens.

### Exercice 3. (2 points)

Reprenez les mesures des taux d'humidité données dans l'exercice 1.

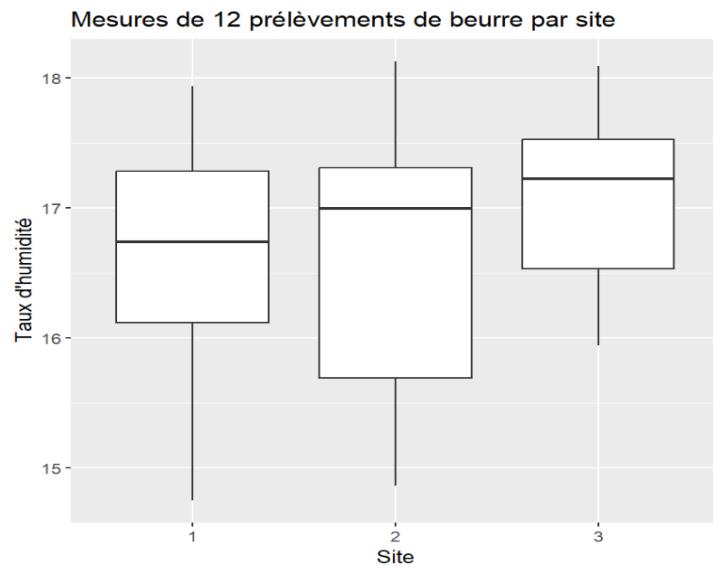
3.1. Calculer le rapport de corrélation entre la variable site et la variable Humidité.

**Correction 3.1.** On a déjà tous les calculs intermédiaires, sauf la variance inter. Variance totale =  $V = 0.9026$ .

Variance inter =  $B = ((16.5816 - 16.75667)^2 + (16.6275 - 16.75667)^2 + (17.0608 - 16.75667)^2)/3 = 0.04661$ .

D'où le rapport de corrélation :  $R = \sqrt{\frac{B}{V}} = 0.2272$ .

3.2. Interpréter ce rapport de corrélation, en vous aidant du graphique suivant, auquel vous donnerez le nom dans le lexique du statisticien (exemple : un diagramme en secteurs est le nom d'un diagramme dans ce lexique).



**Correction 3.2.** Le rapport de corrélation est assez faible mais pas nul. Il est conforme au diagramme en boîtes et moustaches parallèles (c'est le nom du graphique), qui montre que globalement, les mesure sont plus basses sur le site 1, intermédiaires sur le site 2, et plus hautes sur le site 3, mais sans nette distinction du site quand on prend une mesure au hasard.