Statistique - Contrôle du 22 mai 2023 - Correction

L1 Mathématiques - L1 Informatique

Année 2022-2023

Avertissements. Même si tous les exercices portent sur la même thématique, ils peuvent être traités indépendamment. A 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. (6 points) Pour les besoins des hôpitaux, un centre de fabrication remplit des poches de nutriment. Dans le but de surveiller la qualité du résultat, elle contrôle régulièrement ses lots, en mesurant la quantité contenue dans les poches tirées aléatoirement de la fabrication. Soit X_1 la variable aléatoire qui mesure la quantité de nutriment dans une poche (on ne précise pas l'unité de mesure, qui n'a pas d'importance pour la résolution des exercices). On suppose que X_1 suit une loi normale de paramètres $\mu = 16$ et $\sigma^2 = 0.64$. *Rappel de cours.* Si X est une variable aléatoire qui suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, alors $\frac{X-\mu}{\sigma}$ suit une loi normale $\mathcal{N}(0,1)$.
Voici quelques commandes R avec leurs résultats pour vous aider éventuellement :

z	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2
pnorm(z)	0.57926	0.65542	0.72575	0.78814	0.84134	0.88493	0.91924	0.94520	0.96407	0.97725	0.98610

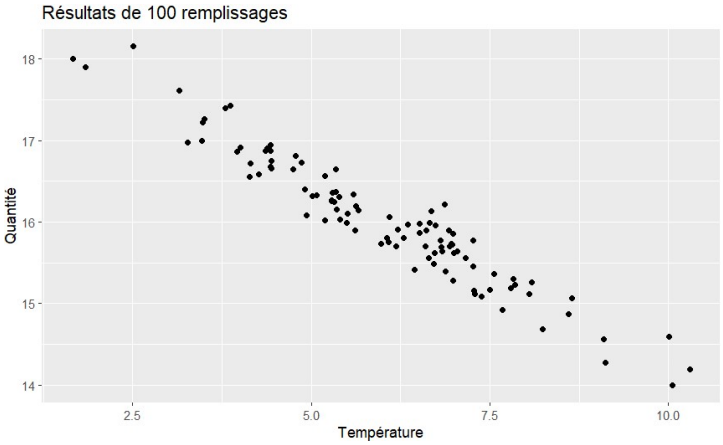
1.1. Déterminer $p_1 = p(X_1 > 17.6)$. On répondra par deux méthodes : la première méthode utilisera la table de la loi normale, la deuxième méthode utilisera les résultats R ci-dessus, en choisissant la valeur qui correspond. Il est important d'expliquer la démarche qui a permis de trouver ce résultat.
Correction 1.1. Puisque $X_1 \sim \mathcal{N}(16, 0.64)$, $\frac{X_1-16}{0.8} \sim \mathcal{N}(0,1)$. Par ailleurs, $p_1 = p(X_1 > 17.6) = p(\frac{X_1-16}{0.8} > 2)$.
Méthode a : avec la table. On cherche dans la table des probabilités de la loi $\mathcal{N}(0,1)$ la valeur de la probabilité pour $u = 2$ et on trouve 0.9772, qui égale $p(X < 2)$ lorsque $X \sim \mathcal{N}(0,1)$. Donc $p(\frac{X_1-16}{0.8} > 2) = 1 - 0.9772 = 0.0228$.
Méthode b : avec les résultats de R. Il suffit de lire l'avant dernier résultat, qui correspond à `1-pnorm(2)`, et on a 0.02275. Le résultat donné est précis à 5 décimales, celui de la table à 4 décimales.

1.2. En déduire, sur une production de 1000 poches, une estimation du nombre de poches qui dépassent la quantité de 17.6 (on arrondira à l'unité près).
Correction 1.2. Puisque la probabilité p_1 pour que la quantité dépasse 17.6 égale 0.02275, sur 1000 poches, on peut estimer à $1000 \times 0.02275 = 22.75$, qu'on arrondit à 23 le nombre de poches dépassant 17.6.

1.3. Sur une deuxième chaîne de fabrication, la variable aléatoire X_2 qui mesure la quantité de nutriment mis dans chaque poche suit une loi normale de paramètres $\mu = 15.95$ et $\sigma^2 = 0.7$. Calculer $p_2 = p(X_2 > 17.6)$. Pour cela, on utilisera la table de la loi normale donnée en cours.
Correction 1.3. On raisonne comme en 1.1. $X_2 \sim \mathcal{N}(15.95, 0.7)$, $\frac{X_1-15.95}{\sqrt{0.7}} \sim \mathcal{N}(0,1)$. Donc $p_2 = p(X_2 > 17.6) = p(\frac{X_2-15.95}{\sqrt{0.7}} > \frac{17.6-15.95}{\sqrt{0.7}}) = p(\frac{X_2-15.95}{\sqrt{0.7}} > 1.972127)$. Sur la table du cours, $p(X < 1.97) = 0.9756$, donc $p(X > 1.97) = 1 - 0.9756 = 0.0244$.

1.4. On sait que 30% des poches sont remplies au poste 1, et 70% sont remplies au poste 2. On note P_1 l'évènement "la poche a été remplie au poste 1" et P_2 l'évènement "la poche a été remplie au poste 2". Soit D l'évènement "La poche a une quantité dépassant 17.6". On a mesuré la quantité d'une poche prise au hasard parmi l'ensemble des poches remplies aux postes 1 et 2. On a trouvé une quantité dépassant 17.6. Quelle est la probabilité pour que cette poche ait été remplie au poste 1 ?
Rappel de cours. Formule de Bayes. Si $(A_j)_{j\in J}$ est une famille d'évènements disjoints tels que $\cup_{j\in J}(A_j) = \Omega$, où Ω est l'évènement certain, c'est-à-dire l'univers, alors, pour tout évènement A, $P(A_k/A) = \dfrac{P(A_k \cap A)}{P(A)} = \dfrac{P(A/A_k)P(A_k)}{\sum_{j\in J} P(A/A_j)P(A_j)}$.
Correction 1.4. On sait que $p(D/P_1) = 0.0228$ et que $p(D/P_2) = 0.0244$. On cherche $p(P_1/D)$. D'après la formule de Bayes, $p(P_1/D) = \frac{p(P_1 \cap D)}{p(D)} = \frac{p(D/P_1)p(P_1)}{p(D/P_1)p(P_1)+p(D/P_2)p(P_2)} = \frac{0.0228\times 0.3}{0.0228\times 0.3+0.0244\times 0.7} = 0.28595$. Remarque : on n'est pas loin de $0.3 = p(P_1)$ car les deux probabilités $p(D/P_1)$ et $p(D/P_2)$ sont très proches aussi, et il est normal que le résultat soit < 0.3 car $p(D/P_1) < p(D/P_2)$.

Exercice 2. (2 points) On souhaite maintenant savoir s'il y a des variations de la quantité de remplissage X en fonction de la température de la pièce T où se produit l'opération. On a fait $n = 100$ essais de variation de température au poste 1. Voici les résultats sous forme de graphique.



Et voici quelques statistiques sur ces résultats : $\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 = 64.37254$, $\sum_{i=1}^{n}(t_i-\overline{t}_n)^2 = 283.3498$, $\sum_{i=1}^{n}(x_i-\overline{x}_n)(t_i-\overline{t}_n) = -130.0007$.

2.1. Calculer le coefficient de corrélation linéaire $cor(X,T)$.

Correction 2.1. $cor(X,T) = \frac{-130.0007}{\sqrt{64.37254*283.3498}} = -0.96257$.

2.2. Commenter le lien entre ces deux variables.

Correction 2.2. La corrélation est proche de -1, donc le lien est très fort. Plus la température est élevée, plus la quantité est faible.

Exercice 3. (4 points) On considère maintenant trois postes de remplissage de ces poches. On a mesuré pendant un même intervalle de temps les quantités contenues dans 10000 poches sur chaque poste (donc $n_1 = n_2 = n_3 = 10000$). Voici les premiers résultats statistiques : $\overline{x}_1 = 15.8$, $\overline{x}_2 = 16.4$, $\overline{x}_3 = 16.1$, $v_1^2 = 0.64$, $v_2^2 = 0.44$, $v_3^2 = 0.87$. La variance de l'ensemble est $V = 0.710$.

Rappel de cours. Soit $(x_1, \ldots, x_n)$ une série statistique, la variance v_n^2 sans pondération de cette série se calcule par la formule $v_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$.

3.1. Calculer la moyenne $\overline{x}_n$ des quantités de l'ensemble des 30000 poches remplies.

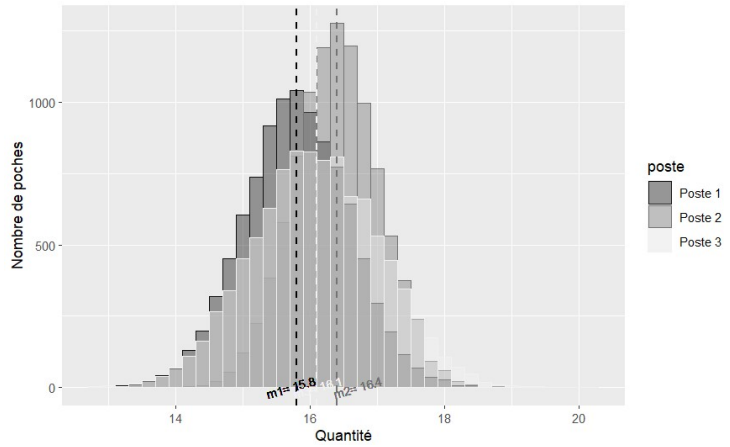
Correction 3.1. Comme les n_i sont tous égaux, la moyenne globale égale simplement la moyenne des trois moyennes : $\overline{x}_n = \frac{1}{3}(\overline{x}_1 + \overline{x}_2 + \overline{x}_3) = \frac{1}{3}(15.8 + 16.4 + 16.1) = 16.1$.

3.2. Calculer la variance intra W de X en tenant compte du poste Y de remplissage.

Correction 3.2. De même, la variance intra est la moyenne des variances dans chaque groupe. $W = \frac{1}{3}(v_1^2 + v_2^2 + v_3^2) = \frac{1}{3}(0.64 + 0.44 + 0.87) = 0.65$.

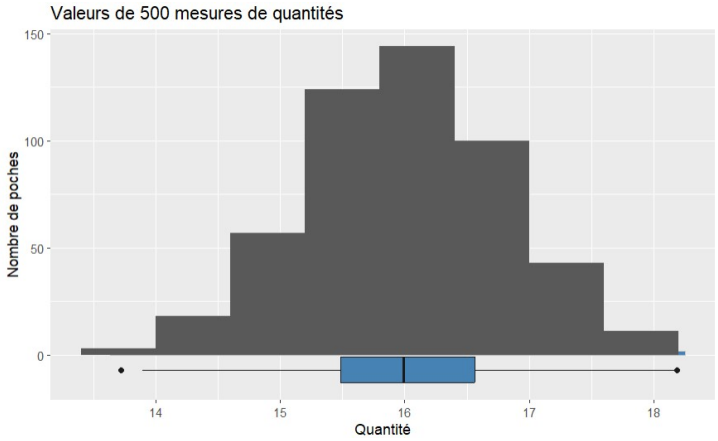
3.3. Calculer le rapport de corrélation entre la quantité X et le poste Y. Conclure, en vous aidant du graphique ci-dessous.

Rappel de cours. Étant donnée une variable quantitative X et une variable qualitative Y, le rapport de corrélation pour mesurer la liaison entre X et Y égale $\sqrt{\frac{B}{V}}$, où B est la variance inter de X/Y, et V est la variance totale de X.



Correction 3.3. $R_{X/Y} = \sqrt{\frac{0.71-0.65}{0.71}} = 0.2907$. Le rapport de corrélation est plus proche de 0 que de 1. Cependant, la liaison n'est pas nulle, surtout en considérant que chaque poste a un échantillon de 10000 poches mesurées, ce qui est beaucoup. Il y a une différence de moyenne entre les 3 postes, le poste 1 ayant la moyenne la plus faible, et le poste 2 la plus élevée.

Exercice 4. (8 points) On a effectué 500 mesures de quantité de poches venant du poste 1. Le diagramme suivant représente ces mesures.



4.1. Comment se nomment les parties de ce diagramme.

Correction 4.1. C'est un histogramme doublé d'un diagramme en boînte et moustaches (boxplot) pour la série des 500 mesures.

4.2. Calculer, à partir de ce diagramme, la proportion de valeurs dépassant 17.6. La détermination graphique est approximative, c'est le raisonnement qui sera compté, pas la valeur exacte qui ne peut être déterminée précisément.

Correction 4.2. Il y a environ 12 mesures qui sont supérieures à 17.6 (hauteur du dernier intervalle), cela représente, sur les 500 mesures, une proportion de 12/500=0.024.

4.3. La machine du poste 1 remplissant les poches s'est un peu déréglée. On ne connaît plus la moyenne et la variance de X_1. On a relevé les quantités pour 200 poches, ce qui a donné le tableau de résultats suivant :

quantité	effectif
14	7
14.5	14
15	17
15.5	26
16	45
16.5	42
17	29
17.5	14
18	6

Calculer une estimation $\overline{x}_1$ de la moyenne de la nouvelle v.a. X_1, ainsi qu'une estimation de sa variance s_1^2.

Correction 4.3. Avec R, le résultat des commandes
quant=seq(14,18,0.5) ; eff=c(7,14,17,26,45,42,29,14,6) ; sum(quant*eff)/sum(eff) donne 16.09. C'est $\overline{x}_1$.
De plus, les commandes R
moy=sum(quant*eff)/sum(eff) ; quantm=quant-moy ; sum((quantm*quantm*eff)/(sum(eff)-1)).
donnent 0.886. C'est notre s_1^2.

4.4. On suppose toujours que X_1 suit une loi normale. Calculer les limites de l'intervalle de confiance (IC) à 95% pour μ_1. On commencera par calculer la marge d'erreur. Voici des résultats de R qui peuvent vous aider.

ν	195	196	197	198	199	200
qt(0.95, ν)	1.652705	1.652665	1.652625	1.652586	1.652547	1.652508
qt(0.975, ν)	1.972204	1.972141	1.972079	1.972017	1.971957	1.971896

Correction 4.4. On a $n = 200$. Pour $1-\alpha = 0.95$, $t_{\alpha/2,n-1} = 1.971957$. La marge d'erreur égale $t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} = 1.971957 * \sqrt{0.886/200} = 0.13127$. D'où l'IC : $[16.09 - 0.13127; 16.09 + 0.13127] = [15.959; 16.221]$.

4.5. On fixe pour le calcul, avec le niveau de confiance 95%, $t_{\alpha/2,n-1} = 1.970$ et $z_{\alpha/2} = 1.960$. On suppose que la moyenne $\overline{x_1}$ et la variance s_1^2 ne changent pas quand n varie. Calculer l'effectif minimum à prélever pour que l'amplitude de l'IC à 95% pour μ_1 ne dépasse pas 0.2.

Correction 4.5. L'amplitude de l'IC égale $2me = 2 * t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}$. Si on fixe $t_{\alpha/2,n-1} = 1.970$, on résout l'équation $2 * 1.970 * s_n/\sqrt{n} < 0.2$, soit $\sqrt{n} > 1.97 * s_n/0.1$, ou encore $n > (1.97/0.1)^2 * 0.886 = 343.848$. L'effectif minimum trouvé est donc de 344.
Remarque. Avec R, si on laisse varier $t_{\alpha/2,n-1}$ en fonction de n, on a men=function(n) return(2*sqrt(s12/n)*qt(0.975,n-1)) ; men(342) ; men(343) donne [1] 0.20027 [1] 0.19997. Cela signifie qu'on passe au-dessous de 0.2 à $n = 343$. Avec t approché, on trouve un résultat très voisin du résultat exact.

4.6. Un nouveau prélèvement de poches issues de la machine du poste 1 a donné une proportion de 22 sur 100 quantités dépassant 17.6. Calculer les limites de l'IC à 95% pour la proportion p_1 de quantités dépassant 17.6.

Correction 4.6. On est dans le cas où la population est infinie. On a $n = 100$, $\widehat{p} = 22/100 = 0.22$, et $z_{\alpha/2} = 1.960$ (table). Donc $me = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.96 * \sqrt{\frac{0.22*0.78}{100}} = 0.08119$. D'où l'IC : $[0.22 - 0.08119; 0.22 + 0.08119] = [0.1388; 0.3012]$.