

Travaux pratiques III : réponses

1. Test d'adéquation en génétique

Le génotype probable des cobayes achetés est vraisemblablement Gg Ll, car c'est le seul génotype qui permette l'apparition dans la descendance de 4 phénotypes différents. Dans ce cas, les fréquences relatives attendues des différents phénotypes sont (voir cours de génétique):

$$E_1 = 9/16 : \text{fréquence de gris lisses,}$$

$$E_2 = 3/16 : \text{fréquence de gris rudes,}$$

$$E_3 = 3/16 : \text{fréquence de blancs lisses,}$$

$$E_4 = 1/16 : \text{fréquence de blancs rudes.}$$

Les fréquences relatives observées correspondantes sont:

$$O_1 = 78/128,$$

$$O_2 = 19/128,$$

$$O_3 = 26/128,$$

$$O_4 = 5/128.$$

Calculons la statistique χ^2 pour cette situation:

$$\begin{aligned}\chi^2 &= n \sum_{i=1}^4 \frac{(E_i - O_i)^2}{E_i} \\ &= 128 \times \left[\frac{(9/16 - 78/128)^2}{9/16} + \frac{(3/16 - 19/128)^2}{3/16} \right. \\ &\quad \left. + \frac{(3/16 - 26/128)^2}{3/16} + \frac{(1/16 - 5/128)^2}{1/16} \right] \\ &\approx 2,83.\end{aligned}$$

Selon la théorie, si le modèle est bon, la distribution des valeurs possibles de la statistique χ^2 est (approximativement) une distribution χ^2 à $(4 - 1) = 3$ degrés de liberté. Dans la table de la distribution χ^2 , on trouve que le percentile 95% d'une distribution χ^2 à 3 degrés de liberté vaut 7.815. Comme $2.83 < 7.815$, l'hypothèse que le modèle est correct n'est pas rejetée avec un seuil de 5%.

2. Analyse d'une expérience de radiophysique

2.1 Nous sommes en présence de données appariées: deux mesures pour chaque sujet.

2.2 Comme énoncé dans la phrase à la fin de la page 4, la quantité à laquelle on s'intéresse est la "réduction" du taux de plaquettes. Pour mesurer cette réduction on peut utiliser la différence entre le nombre de plaquettes avant et après traitement ou bien le rapport entre ces deux nombres. Pour le Bexxar ces deux mesures de réduction sont D_1 et R_1 ; pour le Zevalin elles sont D_2 et R_2 .

2.3 Bien que les données originales soient appariées, pour tester l'hypothèse de la page 4, il faudra comparer la "réduction" obtenue avec le Bexxar à la réduction obtenue avec le Zevalin. Et comme ces deux produits ont été administrés à des sujets différents, les mesures de réduction ne sont pas appariées. Nous pensons donc au test de Student pour données non appariées. Ce test (dans sa forme classique présentée dans le cours) donne un résultat fiable seulement à condition que les distributions de la variable utilisée pour mesurer la "réduction" dans les deux populations soient de type gaussien et de même variance (condition C). L'analyse graphique des distributions empiriques de R_1 et R_2 (graphique (c)) suggère que leurs distributions sont asymétriques et de variances différentes. Nous écartons donc la possibilité d'un test de Student pour R_1 et R_2 . Par contre, les graphiques des distributions de D_1 et D_2 (graphique (d)) ne suggèrent pas de violations sévères de la condition C, malgré la présence d'un outlier modéré dans D_1 . Nous allons donc effectuer un test de Student en utilisant D_1 et D_2 . Nous utiliserons d'abord toutes les données, puis nous enlèverons l'outlier.

2.4 S'agissant d'une étude dont l'impact pratique ne semble pas être "dramatique", nous choisissons un seuil (niveau) de 5%. Nous pouvons aussi ne faire aucun choix et calculer la p-value. Nous déplacerons ainsi le problème du choix sur le dos du lecteur !

2.5 Définissons μ_1 et μ_2 comme les moyennes de D_1 et D_2 dans les populations des sujets que l'on pourrait traiter avec Bexxar et Zevalin. Il s'agit donc des populations à partir desquelles les sujets ont été échantillonnés. L'hypothèse nulle est $\mathcal{H}_0 : \mu_1 = \mu_2$. Pour l'hypothèse alternative on peut considérer deux possibilités. Si nous pouvons écarter à priori l'hypothèse que la réduction du taux de plaquettes est plus importante avec le Bexxar qu'avec le Zevalin, nous posons $\mathcal{H}_1 : \mu_1 < \mu_2$ (alternative unilatérale). S'il n'y a pas d'informations préalables qui nous permettent

d'écarter cette hypothèse, il faudra considérer l'alternative bilatérale $\mathcal{H}_1 : \mu_1 \neq \mu_2$. (Ce choix peut lui aussi être laissé au lecteur en lui fournissant la p-value !).

2.6 En utilisant les moyennes et les écarts types de la table, on a:

$$m(D_1) = 116.11, \quad m(D_2) = 183.14, \quad s(D_1) = 88.09, \quad s(D_2) = 91.96$$

et donc, avec $m = 18$ et $n = 14$, $d = m(D_2) - m(D_1) = 67.03$, $s(D) = 32.00$, et $t = 2.10$. Le calcul exact (avec Excel) de la p-value $p = P(t > 2.10)$ donne $p = 0.022$. A l'aide de la table on peut faire un calcul approximatif: le percentile 97.5% de la distribution de Student avec $m + n - 2 = 30$ degrés de liberté est 2.042 tandis que le percentile 99% est 2.457. Comme $2.042 < 2.10 < 2.475$ on en déduit que $1\% < p < 2.5\%$. En conclusion, l'hypothèse \mathcal{H}_0 peut être rejetée au seuil de 5% tant par un test bilatéral que par un test unilatéral.

2.7 Sur le graphique des distributions de D_1 et D_2 on remarque un outlier modéré dont la valeur est 381 plaquettes.

2.8 En supprimant l'outlier on obtient les résultats suivants:

$$m(D_1) = 100.53, \quad m(D_2) = 183.14, \quad s(D_1) = 60.02, \quad s(D_2) = 91.96$$

et donc avec $m = 17$ et $n = 14$, $d = m(D_2) - m(D_1) = 82.61$, $s(D) = 27.43$, et $t = 3.01$. La valeur de t se modifie donc de façon importante. Elle dépasse maintenant le percentile 99.5% de la distribution de Student à 29 degrés de liberté. Le calcul exact donne $p = 0.003$.

2.9 Les graphiques des distributions de X_1 et X_2 (graphique (a)) pourraient donner l'impression que le nombre de plaquettes avant traitement est un peu plus élevé pour les sujets Zevalin que pour les sujets Bexxar. Toutefois, cette impression est due à la valeur la plus élevée. (Un test préliminaire de l'hypothèse que les moyennes de ces deux populations sont identiques n'est pas recommandé. La p-value globale de la combinaison du test préliminaire et du test principal serait difficile à calculer !)

2.10 Par analogie avec la formule de l'intervalle de confiance pour une moyenne, on peut deviner la formule suivante pour l'intervalle de confiance $1 - \alpha$ pour la différence moyenne $\mu_2 - \mu_1$:

$$(d - t_{\alpha/2, m+n-2} s(D), \quad d + t_{\alpha/2, m+n-2} s(D)).$$

Avec $\alpha = 5\%$, toutes les données et $t_{2.5\%, 30} = 2.042$ on obtient l'intervalle (1.69, 132.38). En supprimant l'outlier en D_1 , avec $\alpha = 5\%$, $t_{2.5\%, 29} = 2.045$, on obtient l'intervalle (26.51, 138.72).