Expression Data Analysis

Guedj Odélia

January 2020

1 Exercice 1 : Rappels de statistiques

Les distributions des intensités moyennes des spots d'ADNc correspondant aux gènes exprimés et non exprimés peuvent être modélisées par deux gaussiennes. Supposons que la première distribution a une moyenne $\mu_p = 1000$ et un écart-type $\sigma_p = 100$ et la seconde une moyenne $\mu_p = 400$ et un écart-type $\sigma_n = 150$. Chaque gène correspond à quatre spots répliqués. L'expression d'un gène est définie comme la moyenne des réplicats.

1.1 Question 1

Soit Z_p, Z_n deux variables aléatoires tel que :

$$Z_p \sim \mathcal{N}(\mu_p, \sigma_p^2) \; et \; Z_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

Alors, la probabilité qu'un spot correspondant à un gène exprimé possède une valeur inférieure ou égale à 700 est :

$$\mathbb{P}(Z_p \leq 700) = \mathbb{P}(Z \leq \frac{700 - \mu_p}{\sigma_p}) = \mathbb{P}(Z \leq \frac{700 - 1000}{100}) = \mathbb{P}(Z \leq -3) = 1 - \mathbb{P}(Z \leq 3) = 0.0013$$

où

$$Z \sim \mathcal{N}(0, 1)$$

1.2 Question 2

Puisque chaque gène correspond à quatre spots répliqués et que l'expression d'un gène est définie comme la moyenne des réplicats alors un gène exprimé suit la loi : $\mathcal{N}(\mu_e, \sigma_e^2)$ avec :

$$\mu_e = \mathbb{E}\left[\sum_{i=1}^{4} \mu_p^i\right] = \frac{1}{4}.4\mu_p = \mu_p$$

et :

$$\sigma_e^2 = \mathbb{V}\left[\sum_{i=1}^{4} (\sigma_p^i)^2\right] = \frac{1}{4^2}.4(\sigma_p)^2 = \frac{(\sigma_p)^2}{4}$$

Ainsi :

$$\mathbb{P}(gene_{expr} \leq 700) = \mathbb{P}(\mathcal{N}(1000, \frac{100^2}{4}) \leq 700) = 9.87e-10$$

1.3 Question 3

Soit $t \in \mathbb{R}$ tel que :

$$\begin{aligned}
\mathbb{P}(Z_p \leq t) = \mathbb{P}(Z_n \geq t) &\Leftrightarrow \mathbb{P}(Z_p \leq t) = \mathbb{P}(Z_n \geq t) \\
&\Leftrightarrow \mathbb{P}(Z \leq \frac{t-1000}{100}) = \mathbb{P}(Z \geq \frac{t-400}{150}) \\
&\Leftrightarrow -\frac{t-1000}{100} = \frac{t-400}{150} \\
&\Leftrightarrow -15(t-1000) = 10(t-400) \\
&\Leftrightarrow -15t + 15000 = 10t - 4000 \\
&\Leftrightarrow 25t = 19000 \\
&\Leftrightarrow t = 760
\end{aligned}$$

1.4 Question 4

Appelons t^* le seuil optimal trouvé à la question préédente alors la probabilité d'avoir un gène exprimé dont l'expression est inférieur à t^* est :

$$\mathbb{P}(gene_{expr} \leq t^*) = \mathbb{P}(\mathcal{N}(1000, \frac{100^2}{4}) \leq t^*) = 7.93e-07$$

1.5 Question 5

En faisant la même chose qu'à la question 2, un gène non exprimé suit la loi : $\mathcal{N}(\mu_{ne}, \sigma^2_{ne})$ avec :

$$\mu_{ne} = \mathbb{E}\left[\sum_{i=1}^{4} \mu_n^i\right] = \frac{1}{4}.4\mu_n = \mu_n$$

et :

$$\sigma^2_{ne} = \mathbb{V}\left[\sum_{i=1}^{4} (\sigma_n^i)^2\right] = \frac{1}{4^2}.4(\sigma_n)^2 = \frac{(\sigma_n)^2}{4}$$

$$\mathbb{P}(gene_{NonExpr} \geq t^*) = \mathbb{P}(\mathcal{N}(400, \frac{150^2}{4}) \geq t^*) = 1 - \mathbb{P}(\mathcal{N}(400, \frac{150^2}{4}) \leq t^*) = 7.93e-07$$

1.6 Question 6

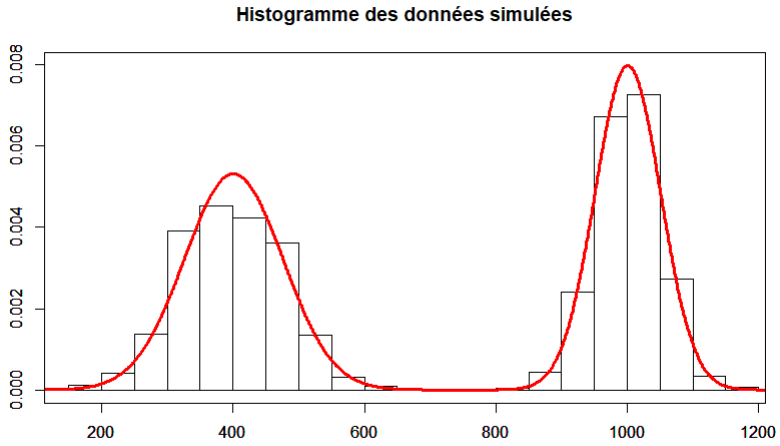

FIGURE 1 – Histogramme et densité des données simulées

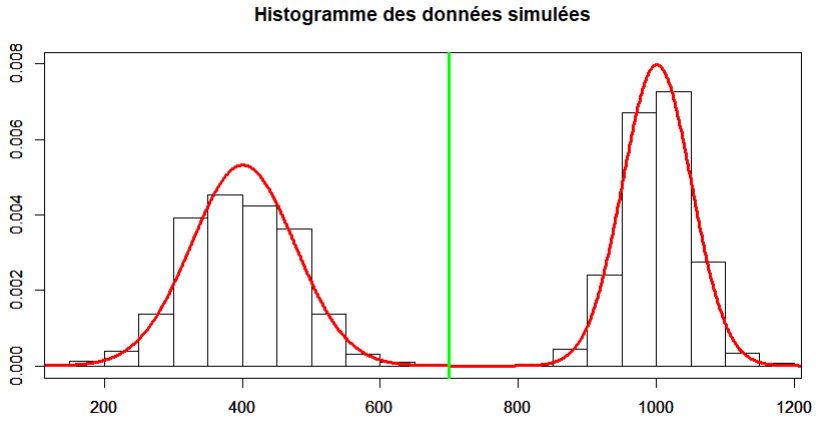

FIGURE 2 – Estimation de la probabilité qu'un gène exprimé soit inférieur ou égal à 700

La probabilité estimée est trés faible (proche de 0.001 comme trouvé à la question 1).

2 Exercice 2 : Tests d'hypothèse

Voir fichier R svp.