BOUZIAT Elodie
GENNUSO Tàzio

Projet de MST 2016-2017

Rapport

L'objectif de ce projet est d'étudier un jeu de données de 150 élèves d'une école maternelle et primaire sur 3 variables que sont le nombre de jours d'absence, le sexe et l'âge de l'élève. L'étude est menée sous R, l'ensemble du code est disponible dans le fichier projet_bouziat_gennuso.R. Ce projet permet de montrer l'utilité de R dans la manipulation statistique sous ses différents aspects.

SOMMAIRE

PARTIE I : statistiques descriptives

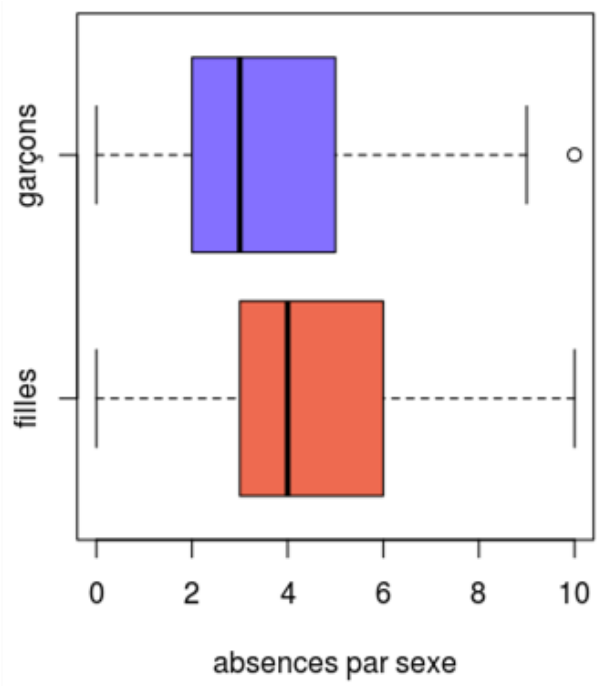
Question 1 :

Le jeu de données a été chargé grâce à la fonction load. Il est disponible sous le nom :

projet_binome_8_groupe_4.RData

Question 2 :

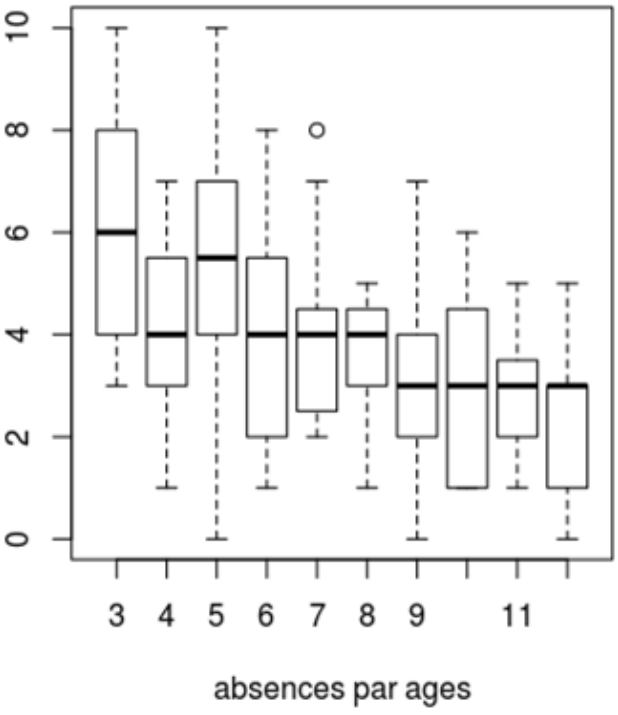
On trace les boites à moustaches du nombre de jours d'absence pour les filles et les garçons sur un même graphique. On obtient le graphique suivant :



On constate que les filles semblent avoir une légère tendance à être plus souvent absentes que les garçons. Cependant les distributions similaires en termes d'amplitude et de rapport aux quartiles, et la médiane est à peine plus élevée chez les filles.

Question 3 :

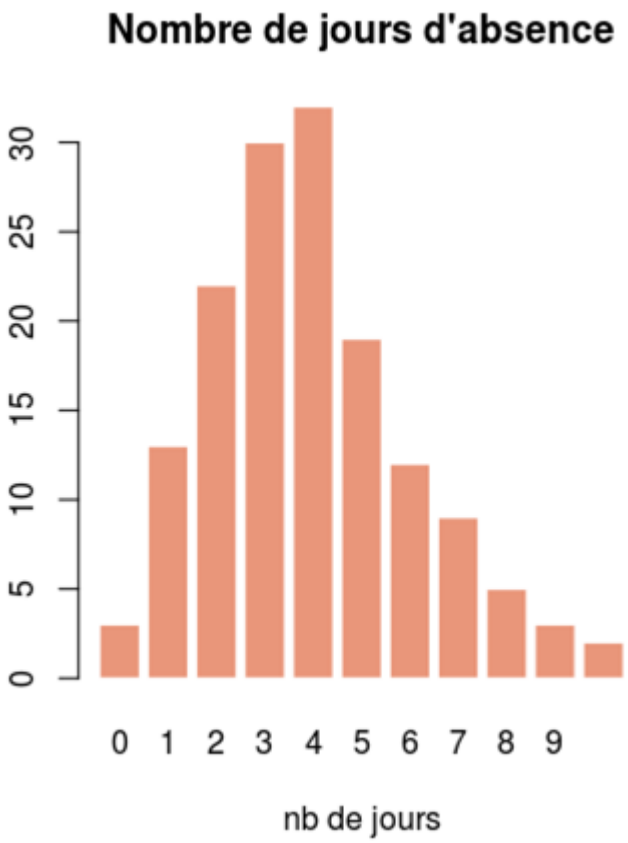
On trace les boîtes à moustache du nombre de jours d'absence pour chaque âge. On obtient le graphique suivant :



On constate que les boites sont peu homogènes entre elles.
La médiane oscille autour de 3 à 4 jours d'absences, avec des pics de 5.5 à 6 pour les enfants de 3 et 5 ans. Les répartitions des quartiles sont très différentes en fonctions des âges. Il semblerait qu'une tendance se dessine grossièrement : plus les enfants sont jeunes, plus ils ont des jours d'absence.

Question 4 :

On trace l'histogramme du nombre de jour d'absence. On obtient le graphique suivant :

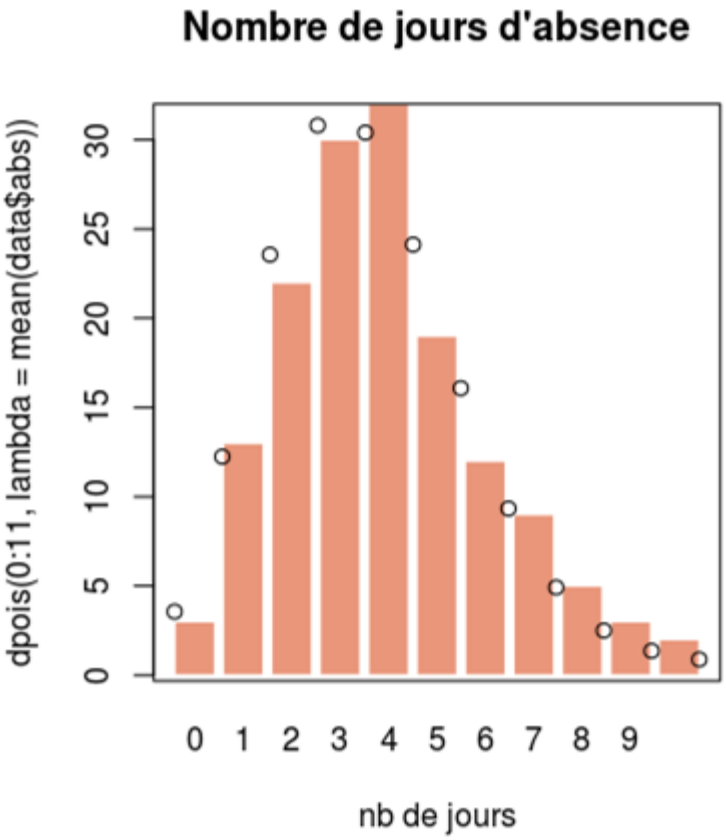


Il semblerait à première vue, étant donné l'allure de l'histogramme, que le nombre de jours d'absence suive une loi de poisson P(λ).
Le maximum de l'histogramme étant aux alentours de 4, on peut en dire autant du paramètre λ.

PARTIE II : statistiques inferrentielles

Question 1 :

On affiche la distribution de la loi de Poisson de paramètre $\hat{\lambda}$, avec $\hat{\lambda}$ la moyenne empirique de l'échantillon (qui est estimé grâce à R à **3.946667**) des nombres de jours d'absence, que l'on notera X_{exp} par la suite, sur le graphe précédent. On obtient le graphique suivant :



On constate que les observations semblent effectivement suivre une loi de Poisson puisque les observations correspondent aux estimations théoriques.

Question 2 :

On cherche maintenant l'estimateur du maximum de vraisemblance de λ. On calcule donc dans un premier temps la vraisemblance $L(X_{exp}, \lambda)$. On a un échantillon de 150 élèves, donc n=150.

$$L(X_{exp}, \lambda) = \prod_{i=1}^{150} e^{-\lambda} \frac{\lambda^{X_{exp_i}}}{X_{exp_i}!}$$

Soit encore :

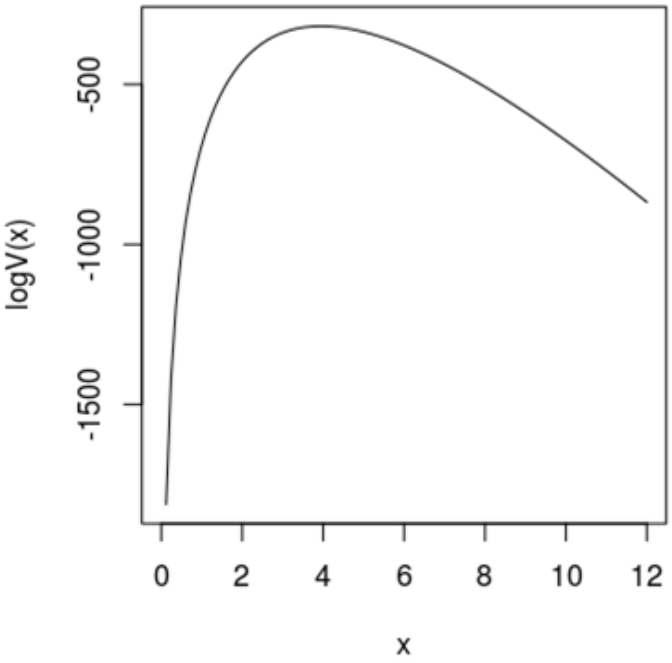
$$L(X_{exp}, \lambda) = e^{-150\lambda} \frac{\lambda \sum_{i=1}^{150} X_{exp_i}}{\prod_{i=1}^{150} X_{exp_i}!}$$

On obtient ensuite la log-vraisemblance :

$$\mathcal{L}(X_{exp}, \lambda) = -150\lambda + ln(\lambda) \sum_{i=1}^{150} X_{exp_i} - \sum_{i=1}^{150} ln(X_{exp_i}!)$$

Question 3 :

On crée une fonction log-vraisemblance et on affiche sa courbe. On obtient le graphique suivant :



On constate qu'il semble que l'EMV se situe aux alentours de 4.

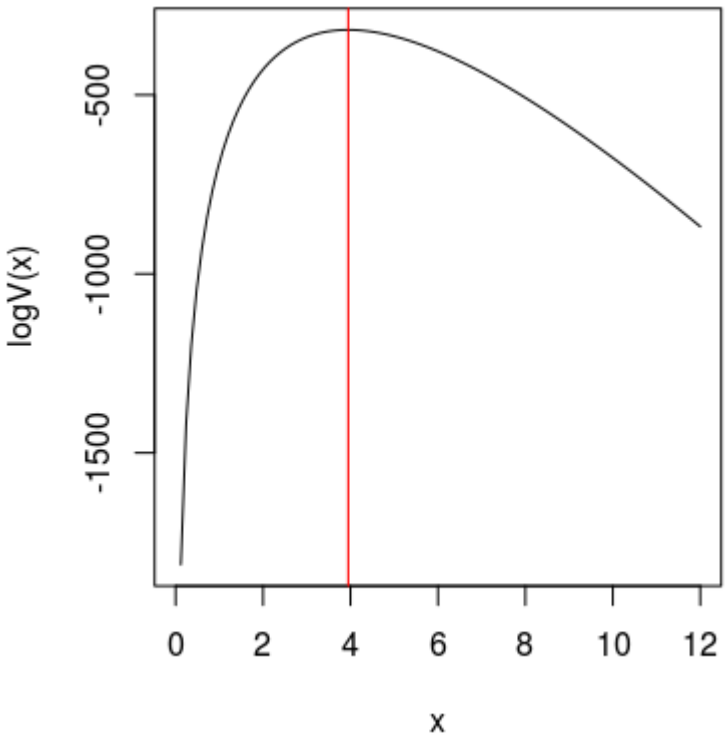
On va maintenant déterminer l'EMV :

$$\frac{\partial \mathcal{L}(X_{exp}, \lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{150} X_{exp_i} - 150$$

La dérivée de la log-vraisemblance s'annule donc pour

$\hat{\lambda}_{EMV} = \frac{1}{150} \sum_{i=1}^{150} X_{exp_i} =$ **3.946667** (ce qui correspond à la moyenne empirique)

Question 4 :



On trace la valeur obtenu théoriquement sur le graphe précédent.

On constate que la valeur obtenue avec l'EMV coïncide bien avec le graphique.

PARTIE III : intervalle de confiance

Question 1 :

Maintenant que nous avons une estimation de λ, on cherche à déterminer un intervalle de confiance asymptotique pour l'échantillon. On note n la taille de l'échantillon.
D'après le théorème central limite, dans le cas d'une loi de Poisson :

$$\sqrt{n}\frac{\hat{\lambda}_n - \lambda}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1)$$

On a donc, pour un intervalle de seuil de confiance α :

$$P\left(-q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\hat{\lambda}_n - \lambda}{\sqrt{\lambda}} \leq q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Soit encore :

$$P\left(\hat{\lambda}_n - q_{1-\frac{\alpha}{2}}\frac{\sqrt{\lambda}}{\sqrt{n}} \leq \lambda \leq \hat{\lambda}_n + q_{1-\frac{\alpha}{2}}\frac{\sqrt{\lambda}}{\sqrt{n}}\right) = 1 - \alpha$$

On a donc pour λ un intervalle de confiance asymptotique de la forme :

$$\left[\hat{\lambda}_n - q_{1-\frac{\alpha}{2}}\frac{\sqrt{\lambda}}{\sqrt{n}}\ , \hat{\lambda}_n + q_{1-\frac{\alpha}{2}}\frac{\sqrt{\lambda}}{\sqrt{n}}\right]$$

avec $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On peut donc calculer empiriquement les deux bornes.

Question 2 :

En prenant $\alpha = 0.05$ (donc un intervalle de confiance à 95%), on obtient l'intervalle :

[3.610479 , 4.282854]

PARTIE IV : tests

Question 1 :

On implémente un test du χ^2 pour vérifier que la répartition du nombre de journées d'absence ne dépend pas du sexe.
On obtient les résultats suivants sous R ($\alpha = 0.05$):

```
Pearson's Chi-squared test
X-squared = 11.305, df = 10, p-value = 0.3342
```

On constate que la p-value de ce test est **supérieure à 0.05**. On accepte donc l'hypothèse selon laquelle le sexe n'a pas d'influence sur le nombre de journées d'absence.

Question 2 :

On implémente un test du χ^2 pour vérifier que la répartition du nombre de journées d'absences ne dépend pas du fait d'avoir plus ou moins de 9 ans.
On obtient les résultats suivants sous R ($\alpha = 0.05$):

```
Pearson's Chi-squared test
X-squared = 24.733, df = 10, p-value = 0.005876
```

On constate que la p-value de ce test est **inférieure à 0.05**. On rejette donc l'hypothèse selon laquelle le fait d'avoir plus ou moins de 9 ans n'a pas d'influence sur le nombre de journées d'absence.

PARTIE V : prédiction

Question 1 :

On veut prédire le nombre de jours d'absences d'un élève selon son âge puisque nous avons vu dans la partie précédente que ce paramètre avait une influence.
On considère que le nombre de jours d'absences suit une loi de Poisson, et que ce modèle s'écrit dans notre cas, en notant Y_{exp} l'âge des élèves et toujours X_{exp} le nombre de jours d'absence :

$$ln(X_{exp}) = aY_{exp} + b$$

Question 2 :

On utilise la fonction glm pour effectuer une régression et calculer les coefficients a et b. On obtient :

a = -0.09015
b = 1.98498

Question 3 :

On modifie le modèle précédent en ajoutant le sexe de l'élève même si celui-ci n'influence pas. On prévoit ainsi avec ce modèle qu'un garçon de 10 ans sera absent **2.79435 jours** par an.