Projet Statistique

Kai HUANG ,Yixuan ZHANG

May 2016

1 Introduction

Nous étudions ici le nombre de jours d'absence dans une école maternelle et primaire. Le jeu de donnée comprend 150 élèves et 3 variables (le nombre de jour d'absence, le sexe de l'élève et son âge). Le but de ce projet est de réaliser une petite étude en utilisant ce qui a été vu durant le module de modélisation statistique. Le projet comprend plusieurs parties: statistiques descriptives, inferrence, intervalle de confiance, tests et prédiction.

2 Partie 1 [statistiques descriptives]

Dans cette partie, nous dessinons les graphes avec R, du coup, les graphes sont suivants.

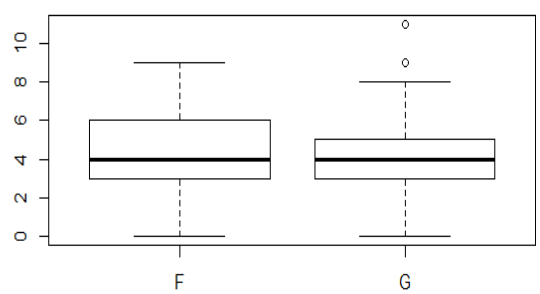

Figure 1: Ex1.2 boites à moustaches du abs pour G et F

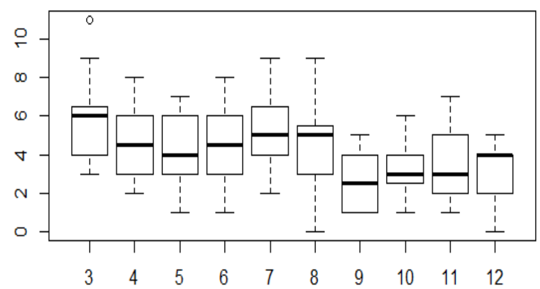

Figure 2: Ex1.3 boites à moustaches du abs pour chaque Age

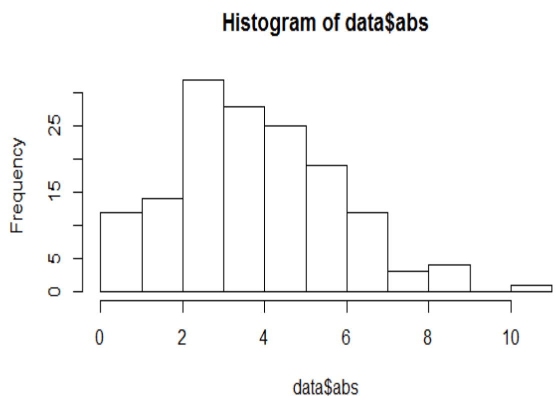

Figure 3: Ex1.4 l'histogramme du abs

3 Partie 2 [statistique inferrentielle]

On considère que le nombre de jour d'absence suit une loi de Poisson $P(\lambda)$. Dans cette partie nous cherchons à estimer λ

Ex2.1

Nous dessinons le graphe comme suivant (Ex2.1), dont le $\lambda = moyenne(abs)$

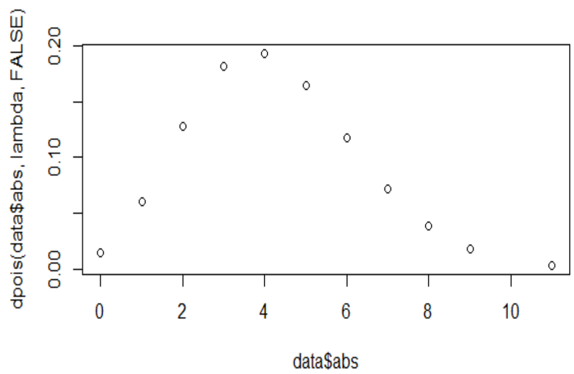

Figure 4: Ex2.1 l'histogramme du abs

Ex2.2

loi de Poisson: $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$
La vraisemblance:

$$\begin{aligned} L(X_1,\cdots+X_n;\lambda) &= \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{X_i}}{X_i!} \\ &= e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{X_i}}{X_i!} \end{aligned}$$

Ex2.3

$$\begin{aligned} L(X_1,\cdots+X_n;\lambda) &= -n\lambda + ln\prod_{i=1}^{n}\frac{\lambda^{X_i}}{X_i!} \\ &= -n\lambda + \sum_{i=1}^{n} X_i ln\lambda - \sum_{i=1}^{n} ln(X_i!) \end{aligned}$$

$\frac{\partial l}{\partial\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0$
Soit $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$
La dérivée seconde s'écrit :$\frac{\partial^2 L}{\partial\lambda^2} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} X_i$
Elle est toujours négative,
Donc, $\widehat{\lambda_n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ est le maximum du vraisemblance du paramètre λ

Ex2.4

$\widehat{\lambda_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

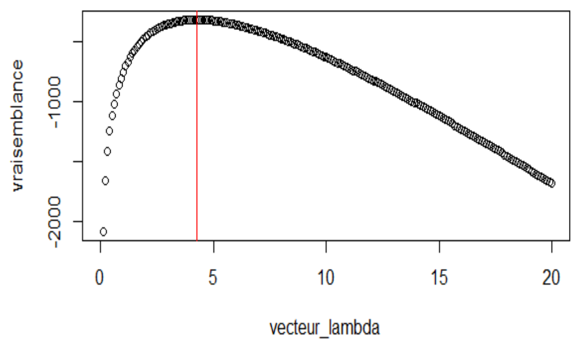

Figure 5: Ex2.3 et Ex2.4 log-vraisemblance de l'échantillon en fonction de λ

4 Partie 3 [intervalle de confiance]

Ex3.1

$\widehat{\lambda_n}$ est l'estimateur du maximum du vraisemblance de λ, nous savons que $\widehat{\lambda}$ est asymptotiquement normal,
i.e. :
$\sqrt{n}(\widehat{\lambda_n} - \lambda) \longrightarrow_{n\to\infty}^{loi} N(0, \frac{1}{I_1(\lambda)})$
Autrement dit: $\sqrt{nI_1(\lambda)}(\widehat{\lambda_n} - \lambda) \longrightarrow_{n\to\infty}^{loi} N(0,1)$

on obtient: $P(-U_{\frac{\alpha}{2}} \leq \sqrt{nI_1(\lambda)(\widehat{\lambda_n} - \lambda) \leq U_{\frac{\alpha}{2}}}) \longrightarrow_{n\to\infty}^{loi} N(0,1)$

$\widehat{\lambda_n} \longrightarrow_{n\to\infty}^{proba} \lambda$, donc,$I_1(\widehat{\lambda_n}) \longrightarrow_{n\to\infty}^{proba} I_1(\lambda)$

d'où l'intervalle de confiance asymptotique: $[\widehat{\lambda_n} - \frac{U_{\frac{\alpha}{2}}}{\sqrt{nI_1(\widehat{\lambda})}} \quad , \widehat{\lambda_n} + \frac{U_{\frac{\alpha}{2}}}{\sqrt{nI_1(\widehat{\lambda})}}]$

$\widehat{\lambda_n} = \bar{X} \qquad U_{\frac{\alpha}{2}} \longrightarrow quantile$

$f(X;\lambda) = e^{-\lambda}\frac{\lambda^X}{X!}$

$m(f(X;\lambda)) = -\lambda + Xln\lambda - ln(X!)$

$\frac{\partial^2 m(f(X;\lambda))}{\partial\lambda^2} = -\frac{X}{\lambda^2}$

$I_1(\lambda) = -[E_\lambda[\frac{\partial^2}{\partial\lambda^2}lnf(X;\lambda)]] = \frac{1}{\lambda}$

$I_2(\widehat{\lambda_n}) = \frac{1}{\widehat{\lambda_n}}$

Donc, l'intervalle de confiance asymptotique est :

$[\widehat{\lambda_n} - \frac{U_{\frac{\alpha}{2}}}{\sqrt{n\frac{1}{\widehat{\lambda_n}}}} \quad , \widehat{\lambda_n} + \frac{U_{\frac{\alpha}{2}}}{\sqrt{n\frac{1}{\widehat{\lambda_n}}}}]$

Ex3.2

Du coup, avec l'aide de R, nous calculons les bornes de l'intervalle de confiance asymptotique est [3.9296949,4.5903051].

5 Partie 4 [tests]

En dessinant les courbes des deux échantillons, on pose qu'ils suivent la loi normale Gaussien.
Pour $X_i (i = 1, \cdots, n)$, il suit une loi $f_X(X; m_0, \sigma_0) = \frac{1}{\sqrt{2\pi}\sigma_0} exp(-\frac{(X-m_0)^2}{2\sigma_0^2})$

Pour $Y_i (i = 1, \cdots, n)$, il suit une loi $f_Y(Y; m_1, \sigma_1) = \frac{1}{\sqrt{2\pi}\sigma_1} exp(-\frac{(X-m_1)^2}{2\sigma_1^2})$

Pour simplifier la question, on suppose que $\sigma_0 = \sigma_1 = \sigma$
On désire tester l'hypothèse $H_0 : \quad m_0 \geq m_1$ contre l'alternative $H_1 : \quad m_0 < m_1$

La vraisemblance de les échantillons $(X_1, \cdots, X_n)$

$$\begin{aligned} L_0 &= \prod_{i=1}^{n} f_X(X_i; m_0, \sigma) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} exp(-\frac{(X_i - m_0)^2}{2\sigma^2}) \end{aligned}$$

La vraisemblance de les échantillons $(Y_1, \cdots, Y_n)$

$$\begin{aligned} L_1 &= \prod_{i=1}^{n} f_Y(Y_i; m_1, \sigma) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} exp(-\frac{(Y_i - m_1)}{2\sigma^2}) \end{aligned}$$

$$\begin{aligned} R(L_0, L_1) &= \frac{L_1}{L_0} \\ &= exp(-\frac{(Y_i - m_1)^2 - (X_i - m_0)^2}{2\sigma^2}) \end{aligned}$$

Nous cherchons à comparer la moyenne des jours d'absence entre les filles et les garçons.

Item pour la question 4.2.

6 Partie 5 [prédiction]

Ex5.1

Le modèle est
$M = \{f(X;\lambda) \mid \lambda \in R^+\}$

Ex5.2

Nous utilisons le code suivant pour déterminer les coefficients α et β.

model ← glm(data*abs data*Age,family = poisson(link = "log"))
summary(model)

Au final, nous calculons α=1.89977 et β=-0.06370 .

Ex5.3

Avec les résultats de questions précédentes, nous avons une formule comme suivante.

$\frac{nb\_garcon}{nb\_total}(\lambda\alpha + age * \beta)$

nb_garcon = 81;   nb_total = 150;
λ = 4.26; α = 1.89;   age = 10; β = -0.063

Du coup, nous savons que la réponse est 4.01