DELEGUE Hugo - TANIEL Rémi GIS2A3

Compte rendu TP : Examen

Exercice 1 - Etude d'un élément chimique

Pour cette partie, nous utiliserons le jeu de donnée "échantillon.csv" représentant la durée de vie en année de noyaux radioactifs de Césium.
Nous savons que ces données sont distribués selon une loi exponentielle de paramètre λ.

Question 1

On commence par importer le jeu de données, voici les 10 premières lignes de ce fichier :

```
> head(echantillon)
  X   valeurs
1 1 18.596943
2 2 45.683665
3 3 32.598273
4 4 71.980621
5 5 89.726689
6 6  8.281879
```

Puis nous générons un échantillon $(X_1, X_2, ..., X_n)$ de taille $n = 50$ à partir de ces données, dans la suite nous utiliserons ce même échantillon pour toutes les estimations.

Question 2

Nous proposons un premier estimateur $\widehat{\lambda}_0$ calculé à partir de la variance empirique :

$\widehat{\lambda}_0 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0.03147776$

Avec $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Question 3

Nous calculons un second estimateur $\widehat{\lambda}_1$ de λ en utilisant cette fois-ci la méthode des moments. Nous utilisons le plus petit moment d'ordre convenable c'est à dire le moment d'ordre 1.

Soit $m_1(\lambda) = E(X_i) = \frac{1}{\lambda}$
Donc m_1^{-1} l'inverse de la fonction précédente est défini par :
$m_1^{-1}(\lambda) = \frac{1}{\lambda}$

L'estimateur $\widehat{\lambda}_1$ est donc obtenu grâce à la formule :
$\widehat{\lambda}_1 = m_1^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_i) = 1/(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{\bar{X}} = 0.0270138$

Question 4

Le but de cette question est le même que les 2 questions précédentes, c'est à dire proposer un estimateur de λ mais pour celle-ci nous allons proposer un estimateur $\widehat{\lambda}_2$ de λ en utilisant la méthode du maximum de vraisemblance.

Pour rappel, la fonction de densité d'une loi exponentielle est :
$f(x) = \lambda e^{-\lambda x}$ si $x \geq 0$ et 0 sinon

On pose $L(\lambda) = \prod_{i=1}^{n} f(X_i) = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$
Pour rappel $\widehat{\lambda}_2 = argmax\,(L(\lambda))$ qui est le même que $argmax\,(Log(L(\lambda))$
Soit $l(\lambda) = log(L(\lambda)) = n\,log(\lambda) - \lambda \sum_{i=1}^{n} X_i$
En dérivant cette fonction, on obtient :
$l'(\lambda) = n\frac{1}{\lambda} - \sum_{i=1}^{n} X_i$
Elle s'annule au point :
$\lambda = n / \sum_{i=1}^{n} X_i$
Donc
$\widehat{\lambda}_2 = argmax\,(L(\lambda)) = n / \sum_{i=1}^{n} X_i = 0.02700195$

Question 5

Pour rappel, voici les valeurs que nous obtenons pour les 3 estimateurs $\widehat{\lambda}_0$, $\widehat{\lambda}_1$ et $\widehat{\lambda}_2$:

$\widehat{\lambda}_0$	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$
0.03147776	0.0270138	0.02700195

On remarque que les 2 estimateurs calculés grâce à la méthode des moments et du maximum de vraisemblance sont très proche et beaucoup plus précis que celui calculé à partir de la variance empirique.

Afin de comparer les estimateurs plus précisément, on pourrait se décider de calculer pour chacun, son biais b et son risque quadratique moyen R.
Pour rappel, le biais b d'un estimateur $\widehat{\lambda}$ de λ est défini comme :
$b_{\widehat{\lambda}} = E(\widehat{\lambda}) - \lambda$
Et le risque quadratique R par :
$R_{\widehat{\lambda}} = E((\widehat{\lambda} - \lambda)^2) = V(\widehat{\lambda}) - b_{\lambda}^2$

Question 6

Afin d'améliorer le calcul de nos estimateurs, nous pourrions filtrer préalablement nos données afin d'écarter les valeurs les plus aberrantes, augmenter la taille de notre échantillon.

Question 7

Dans cette question, nous souhaitons estimer le paramètre β d'une loi Gamma de paramètres $\alpha = 1$ et β par la méthode du maximum de vraisemblance.

On rappelle la densité d'une loi Gamma $\Gamma(\alpha, \beta)$:
$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
En remplaçant $\alpha = 1$, on obtient :
$f(x) = \beta e^{-\beta x}$

On suit la même méthode qu'à la question 4 et en appelant $\widehat{\beta}$ notre estimateur, on obtient que :
$\widehat{\beta} = n / \sum_{i=1}^{n} X_i = \widehat{\lambda}_2$
Nos 2 estimateurs sont donc égaux.

Question 8

On décide de garder la valeur obtenu grâce à l'estimateur $\widehat{\lambda}_2$, la durée de vie moyenne m est obtenu à partir de l'espérance de la loi exponentielle, c'est à dire :
$m = E(X) = \frac{1}{\lambda} = \frac{1}{0.02700195} = 37.034$

Exercice 2 - Sécurité routière

Dans cette partie, nous nous intéressons à la vitesse moyenne des automobilistes sur une route, les données ont été relevés par un contrôle routier.
Nous savons également que la vitesse des véhicules suit une loi normale $N(\mu, \sigma^2)$

Partie 1 : Vitesse moyenne de l'ensemble des automobilistes

Question 1

Nous importons le jeu de données "police.csv", voici les 10 premières lignes de ce fichier :

Voici les informations relatives aux différentes variables :

```
> summary(police)
       X             Vitesse       Distance_Reaction Distance_freinage
 Min.   :  1.00   Min.   :  5.00   Min.   : 1.389    Min.   :  0.123
 1st Qu.: 53.75   1st Qu.: 47.53   1st Qu.:13.205    1st Qu.: 11.123
 Median :106.50   Median : 49.85   Median :13.845    Median : 12.225
 Mean   :106.50   Mean   : 52.36   Mean   :14.546    Mean   : 16.124
 3rd Qu.:159.25   3rd Qu.: 52.45   3rd Qu.:14.570    3rd Qu.: 13.535
 Max.   :212.00   Max.   :251.00   Max.   :69.722    Max.   :310.044
 Exces_Vitesse
 Non:113
 Oui: 99

   Conducteur  Nombre_passagers              Heure_controle
 Femme  : 96   0      :99       Entre 11h30 et 13h30:109
 Homme  :113   2      :43       Entre 9h00 et 9h30  : 99
 Inconnu:  3   4      :67       Entre 9h30 et 10h00 :  4
               Inconnu: 3

> head(police)
  X Vitesse Distance_Reaction Distance_freinage Conducteur
1 1  250.00           69.4444          307.5787    Inconnu
2 2  251.00           69.7222          310.0443      Femme
3 3    5.00            1.3889            0.1230    Inconnu
4 4  224.00           62.2222          246.9291    Inconnu
5 5   51.93           14.4200           13.2700      Homme
6 6   51.05           14.1800           12.8300      Homme
```

Nombre_passagers	Heure_controle	Exces_Vitesse
Inconnu	Entre 9h30 et 10h00	Non
2	Entre 9h30 et 10h00	Non
Inconnu	Entre 9h30 et 10h00	Non
Inconnu	Entre 9h30 et 10h00	Non
0	Entre 9h00 et 9h30	Oui
0	Entre 9h00 et 9h30	Oui

Question 2

Le but de cette question est de savoir si nous pouvons affirmer avec un risque $\alpha = 5\%$ que la vitesse moyenne des automobilistes est différente de 50 km/h. Pour cela, nous allons établir les hypothèses nulle et alternative, construire une zone de rejet (ou d'acceptation)

L'hypothèse nulle H_0 est donc $H_0 : \mu = \mu_0 = 50$
Et l'hypothèse alternative H_1 est $H_1 : \mu \neq \mu_0 = 50$
C'est donc un test bilatéral

On pose $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
Donc sous H_0 on a $\overline{X} \sim N(\mu_0, \frac{\sigma^2}{n})$

Dans notre cas, σ^2 est inconnu donc on l'estime par la variance empirique :
$\overline{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_I - \overline{X})^2$
Et nous utilisons la statistique de test suivante :
$T = \sqrt{n}\ (\overline{X} - \mu_0) / \overline{S}\ \sim T_{n-1}$

Pour un risque α donné, on a donc :
$P(-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2}) = 1 - \alpha$

On obtient une zone de rejet égale à l'intervalle suivant :
$]-\infty; \mu_0 - \overline{S}\frac{1}{\sqrt{n}} t_{1-\alpha/2}[\cup]\mu_0 + \overline{S}\frac{1}{\sqrt{n}} t_{1-\alpha/2}; +\infty[$

La p-value du test est obtenu grâce à la formule :

α est le niveau de signification du test, c'est selon cette valeur que nous allons rejeter ou conserver les hypothèses, cette valeur est toujours fixé par l'utilisateur et ne dépends jamais du problème, c'est la probabilité de rejeter H_0 alors que H_0 est vrai, de rejeter H_0 à tord.

La p-value d'un test statistique est le risque limite avec lequel on passe de conserver H_0 à rejeter H_0.

On obtient l'intervalle de confiance suivant:
$[46.86309; 53.13691]$

Sachant que la moyenne empirique est égale à: 52.3641
On ne peut donc pas rejeter H_0, par conséquent on le conserve et on ne peut pas affirmer au risque de 5 % que la vitesse moyenne des automobilistes est différente de 50km/h.

Question 3

Partie a

Cette fois-ci, nous voulons affirmer que la vitesse moyenne des automobilistes est supérieure à 45 km/h, les hypothèses changent, elles deviennent :
$H_0 : \mu = \mu_0 = 45$
$H_1 : \mu > \mu_0 = 45$
Le test devient unilatéral, mais la statistique de test ne change pas.

Pour un risque α donné, on a maintenant :
$P(T \leq t_{1-\alpha/2}) = 1 - \alpha$

Donc la zone de rejet du test devient :
$[\mu_0 + \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}; +\infty[$

On obtient la zone de rejet suivante:
$[48.137; +\infty[$

Sachant que la moyenne empirique est égale à: 52.3641
Etant donné que la moyenne appartient à la zone de rejet on peut donc rejeter H_0 et affirmer que la vitesse moyenne des automobiliste est supérieur à 45km/h.

Partie b

Dans le cas où nous voulons affirmer que la vitesse moyenne des automobilistes est inférieure à 65 km/h, nos hypothèses deviennent :
$H_0 : \mu = \mu_0 = 65$
$H_1 : \mu < \mu_0 = 65$
Le test reste unilatéral, et la statistique de test ne change pas.

Pour un risque α donné, on a maintenant :
$P(T \leq t_{1-\alpha/2}) = 1 - \alpha$

Donc la zone de rejet du test devient :
$]-\infty; \mu_0 - \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}]$

On obtient la zone de rejet suivante:
$]-\infty; 61.863]$

Sachant que la moyenne empirique est égale à: 52.3641
Etant donné que la moyenne appartient à la zone de rejet on peut donc rejeter H_0 et affirmer que la vitesse moyenne des automobiliste est inférieur à 65 km/h.

On remarque que dans les 2 cas on rejette les tests, donc la vitesse moyenne des automobilistes est à la fois supérieur à 45 et inférieure à 45 km/h. On peut donc encadrer la vitesse moyenne des automobilistes par ces deux valeurs.

Nous pouvons même être plus précis, on peut encadrer la vitesse moyenne par les bornes non infinies de nos deux test précédents ce qui nous donne l'intervalle suivant :
$[48.1; 61.9]$

Question 4

En utilisant la fonction t.test de R, nous obtenons des résultats similaire, donc cela confirme ceux obtenus précédemment.

Voici les deux intervalles de confiance obtenus:

Vitesse > 45 km/h	Vitesse < 65 km/h
$[49.7;+\infty[$	$]-\infty; 55]$

On rejette également dans les deux cas.

Question 5

On a vu précédemment que notre statistique de test était :
$T = \sqrt{n}\ (\overline{X} - \mu_0) / \overline{S}\ \sim T_{n-1}$

Donc :
$P(-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2}) = 1 - \alpha$

En isolant μ_0 , on obtient :
$P(\overline{X} - \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2} \leq \mu_0 \leq \overline{X} + \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}) = 1 - \alpha$

Et donc, l'intervalle de confiance suivant :
$[\overline{X} - \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}; \overline{X} + \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}]$

Intervalle de confiance de niveau 95% ($\alpha = 0.05$) :
$[49.2; 55.5]$

Partie 2 : Influence d'automobilistes atypiques sur la vitesse moyenne

Question 1

Cette fois-ci, on importe les données du fichier "police2.csv", ce fichier ne comporte pas les 4 premières lignes du fichier "police.csv" qui sont aberrantes par rapport aux restes des lignes.

Question 2

On décide calculer un nouvel intervalle de confiance de la vitesse moyenne des automobilistes à partir de ce nouveau jeu de données.

Rappel de la formule de l'intervalle de confiance :
$[\overline{X} - \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}; \overline{X} + \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}]$

Intervalle de confiance de niveau 95% ($\alpha = 0.05$) :
$[49.5; 50.3]$

Cette vient confirmer tous nos précédents résultats et est même plus précis.

Question 3

Grâce à l'intervalle de confiance obtenu dans la question précédente, nous pouvons dire que la vitesse réelle autorisée sur cette route est de 50 km/h.
Cela corrobore avec les valeurs dans la table :

```
> summary(police2)
       X             Vitesse      Distance_Reaction Distance_freinage
 Min.   :  1.00   Min.   :45.02   Min.   :12.51     Min.   : 9.97
 1st Qu.: 52.75   1st Qu.:47.53   1st Qu.:13.21     1st Qu.:11.12
 Median :104.50   Median :49.81   Median :13.84     Median :12.21
 Mean   :104.50   Mean   :49.86   Mean   :13.85     Mean   :12.28
 3rd Qu.:156.25   3rd Qu.:52.31   3rd Qu.:14.53     3rd Qu.:13.47
 Max.   :208.00   Max.   :54.75   Max.   :15.21     Max.   :14.75
 Exces_Vitesse
 Non:109
 Oui: 99

 Conducteur  Nombre_passagers              Heure_controle
 Femme: 95   Min.   :0.000    Entre 11h30 et 13h30:109
 Homme:113   1st Qu.:0.000    Entre 9h00 et 9h30  : 99
             Median :2.000
             Mean   :1.692
             3rd Qu.:4.000
             Max.   :4.000
```

Question 4

Les commentaires des policiers laissent penser une erreur de calibrage du radar pour les 4 premiers automobilistes étant donné que peu d'informations ont été remplis et que les vitesses sont soient énorme soit nulle, on peut soit penser à une erreur de calibrage soit des chauffards que les policiers n'ont pas pris en chasse.
Dans les 2 cas, il n'est pas nécessaire de les inclure dans le calcul des intervalles de confiance

Partie 3 : Quelques résultats

Question 1

Nous souhaitons obtenir des intervalles de confiance de la vitesse moyenne et de l'écart-type des mesures de vitesses en différenciant le sexe du conducteur.

Pour rappel, voici la formule de l'intervalle de confiance de la vitesse moyenne :
$[\overline{X} - \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}; \overline{X} + \overline{S}\frac{1}{\sqrt{n}}t_{1-\alpha/2}]$

Et voici celui de l'écart-type :
$[\overline{S}\sqrt{n-1} / \sqrt{u_{n-1,1-\alpha/2}}; \overline{S}\sqrt{n-1} / \sqrt{u_{n-1,\alpha/2}}]$

On obtient les intervalles suivants :

	Vitesse moyenne	Ecart-type
Homme	[49.06891 ; 50.43286]	[2.348141 ; 3.314393]
Femme	[49.15886 ; 50.82725]	[2.587771 ; 3.770061]

Question 2

La variance des relevés de vitesse effectué des conducteurs diffère légèrement de celle des conductrices cependant ce n'est pas assez significatifs.

Question 3

De la même manière que pour la variance, les moyennes des conducteurs et des conductrices sont pratiquement les mêmes avec un intervalle très légèrements plus étendue pour les femmes mais les deux intervalles restent entre 49 et 51 km/h ils sont donc extrêmement proche on ne peut rien en conclure.

Question 4

On estime la proportion d'automobilistes qui ne transportait aucun passager entre 9h00 et 9h30, par le nombre de ces automobilistes divisés par le nombre total d'automobilistes:

Cela nous donne la proportion suivante :
$p = \frac{99}{208} = 0.4759615$

Question 5

Pour considérer que les automobilistes qui sont en excès de vitesse ont une distance de freinage supérieur à 13 mètres 30 on effectue un test à un niveau de 95% avec pour hypothèse:

$H_0 : m = m_0$
$H_1 : m > m_0$
avec $m_0 = 12.3$

De la même manière que dans la question 3 de la partie 2 on construira une zone de rejet, si la moyenne appartient à cette zone on pourra rejeter H_0.

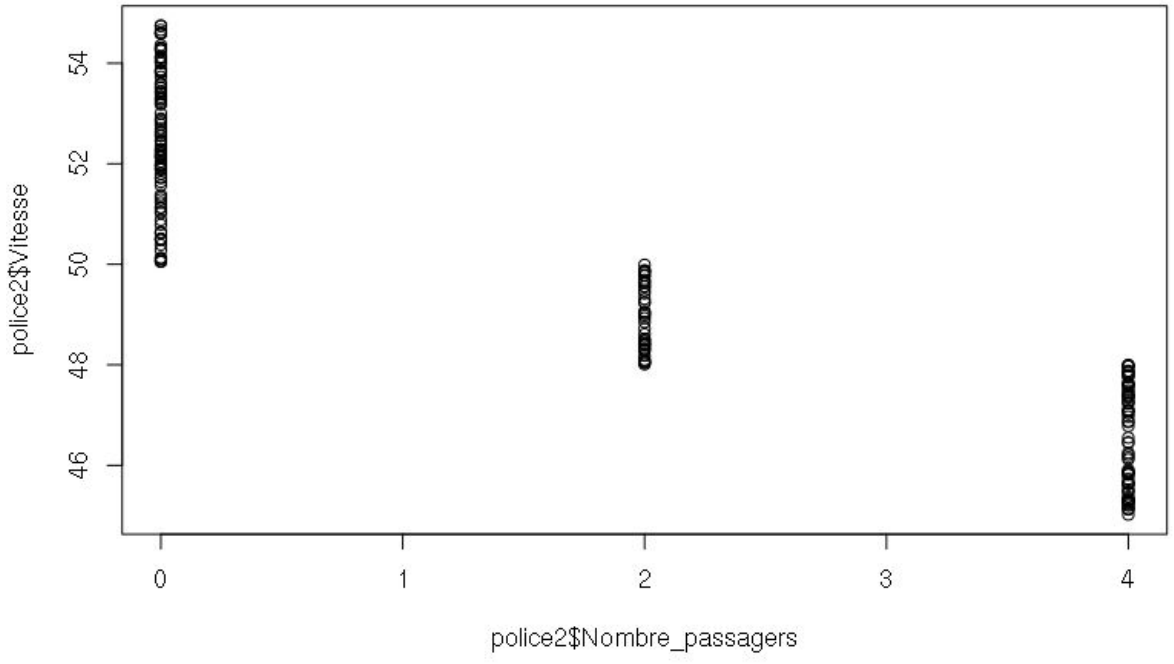
On obtient la zone de rejet suivante :
$[12.44291;+\infty[$

Sachant que la moyenne des distance de freinage des conducteurs en excès de vitesse est de : 13.58384

On peut donc rejeter H_0 et considérer que les automobilistes qui sont en excès de vitesse ont une distance de freinage supérieur à 12.30 mètres.

Question 6

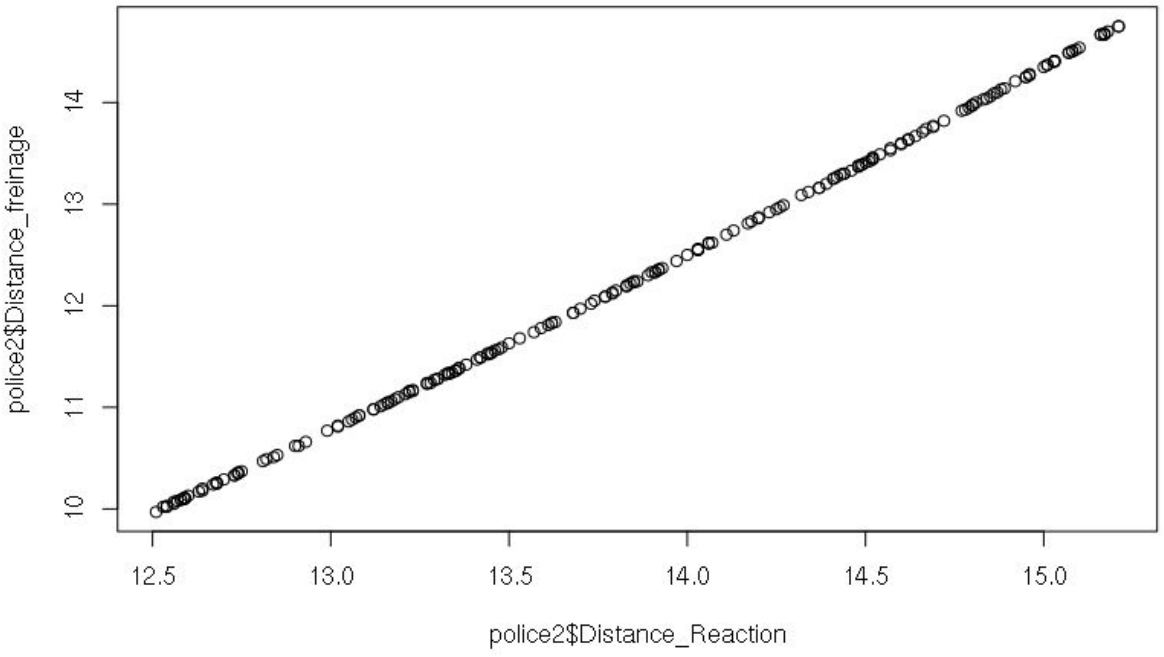
a. Lien entre le nombre de passagers et la vitesse



b. Lien entre le nombre de passagers et l'heure de controle

/

c. Lien entre la distance de réaction et la distance de freinage



Il y a bien un lien entre la distance de réaction et la distance de freinage, en effet plus la distance de réaction est grande plus la distance de freinage le sera aussi et inversement

Bonus

Les policiers ont effectué un contrôle en ville (vitesse limité à 50 km/h)
Les policiers ont effectué un contrôle proche d'une école,pratiquement pas de passagers entre 9h00 et 9h30, à l'inverse beaucoup de passagers entre 11h30 et 13h30 (heure de sortie et rentrée des écoles)