Université Lille 1 **MASS L3** **M.Fradon**
MASS L3 **Tests d'Hypothèses Statistiques** **2015-2016**

Examen de deuxième session du 20 juin 2016

durée : 3 heures

Matériel autorisé : table statistique, calculatrice, une feuille A4 manuscrite recto-verso.

Ex 1. Les apiculteurs du Texas ont constaté avec inquiétude la progression des abeilles tueuses dans les années 80, au détriments des abeilles donnant du miel. Pour convaincre les pouvoirs publics de combattre le phénomène, il fallait prouver que la proportion d'abeilles tueuses augmentait significativement. En 1980 et en 1990, les apiculteurs ont récolté des données à l'aide de 10 pièges répartis sur le territoire texan. Les proportions d'abeilles tueuses capturées étaient :

piège numéro	1	2	3	4	5	6	7	8	9	10
proportion en 1980	0,330	0,146	0,518	0,339	0,693	0,249	0,438	0,695	0,135	0,388
proportion en 1990	0,360	0,177	0,524	0,447	0,140	0,392	0,534	0,263	0,157	0,566

1) Les apiculteurs ont fourni les deux séries de 10 données à un statisticien. Il a effectué un test de la somme des rangs, au niveau 6%. Faire ce test. Quelle conclusion le statisticien a-t-il annoncé aux apiculteurs ?

2) Le statisticien a expliqué aux apiculteurs que le test précédent présentait un défaut : la proportion d'abeilles tueuses piégées dépend de la proportion d'abeilles tueuses dans la nature, mais aussi du nombre total d'abeilles piégées, qui varie d'un piège à l'autre. Cette mise en garde a provoqué une longue discussion entre les apiculteurs et le statisticien. Il a alors appris qu'ils avaient oublié de lui communiquer une information : les pièges avaient été placés aux mêmes endroits en 1980 et en 1990. Les données étaient donc appariées. Le statisticien a ainsi pu effectuer un test du signe. Faire ce test, toujours au niveau 6%, et conclure. Quelle est la taille de ce test du signe ?

Ex 2. En mars 2013, un institut de sondage a recueilli 1684 réponses de lycéens, 1133 réponses de parents d'élèves et 1083 réponses d'enseignants de lycée. La question portait sur "les clés de la réussite au bac". Les réponses de chaque groupe ont été réparties en quatre classes selon l'élément cité comme important pour réussir au bac :

	lycéens	parents	enseignants	totaux
avoir de bons professeurs, être dans un bon lycée	714	459	322	1495
travailler régulièrement toute l'année	512	386	391	1289
être issu d'une famille stable ou d'un milieu aisé	244	207	301	752
avoir de la chance à l'examen	214	81	69	364
totaux	1684	1133	1083	3900

Au vu de ces données, peut-on dire que la répartition des réponses est la même dans les trois catégories de personnes interrogées (élèves, parents et enseignants) ? Construire un test au niveau 5%.

Pour conclure ce test, il faut calculer de la valeur observée de la statistique. Ce serait long à faire à la calculatrice, c'est pourquoi le code suivant en `Scilab` est mis à votre disposition. Indiquer quel `Ti` calcule la statistique et donner la conclusion du test.

```
ind=[1684,1133,1083];
causes=[1495,1289,752,364];
obs=[[714,459,322];[512,386,391];[244,207,301];[214,81,69]];

E1=zeros(4,3);
for (i=1:4)
    for (j=1:3)
        E1(i,j)=causes(i)/3;
     end
end

E2=zeros(4,3);
for (i=1:4)
    for (j=1:3)
         E2(i,j)=causes(i)*ind(j)/3900;
    end
end

-->T1=sum((E1-obs).^2 ./E1)
 T1  = 307.17282
-->T2=sum((E1-obs).^2 ./obs)
 T2  = 304.79376
-->T3=mean((E1-obs).^2 ./E1)
 T3  = 25.597735
-->T4=mean((E1-obs).^2 ./obs)
 T4  = 25.39948
-->T5=sum((E2-obs).^2 ./E2)
 T5  = 134.4115
-->T6=sum((E2-obs).^2 ./obs)
 T6  = 134.98504
-->T7=mean((E2-obs).^2 ./E2)
 T7  = 11.200958
-->T8=mean((E2-obs).^2 ./obs)
 T8  = 11.248753
```

Ex 3. Pour essayer un nouvel engrais, on choisit 24 parcelles identiques. On utilise l'engrais sur 12 d'entre elles, les autres étant traitées de manière habituelle. On note $X_1,\ldots,X_{12}$ les productions obtenues avec le nouvel engrais et $Y_1,\ldots,Y_{12}$ celles obtenues sans. Sur les parcelles avec engrais, on observe une production moyenne de $5,10$ quintaux avec un écart-type de $0,40$ quintaux. Sur les autres parcelles, la production moyenne observées est de $4,80$ quintaux et l'écart-type observé de $0,36$ quintaux. On veut construire une interprétation de ces résultats, en utilisant le fait que la production par parcelle est gaussienne.

1) Tester au niveau 5% si la variance de la production est la même sur les parcelles avec et sans nouvel engrais.

2) Tester au niveau 5% si le nouvel engrais augmente la production.

Ex 4. 1) On fait n tirages indépendants $Y_1,\ldots,Y_n$ de la variable aléatoire Y qui a pour densité

$$\forall y \in \mathbb{R} \quad f_a(y) = 2a\ y\ e^{-ay^2}\mathbf{1}_{y\geq 0}.$$

Le paramètre $a > 0$ est inconnu. Calculer l'estimateur du maximum de vraisemblance $\hat{a}_n$.

2) On veut construire le test du rapport de vraisemblance de l'hypothèse $\mathcal{H}_0 : a = 1$ contre l'hypothèse $\mathcal{H}_1 : a \neq 1$. Calculer le rapport de vraisemblance. En déduire la forme de la région de rejet. Montrer que cette région de rejet peut s'écrire comme une zone bilatère construite à partir de la statistique $1/\hat{a}_n$ (indication : regarder les variations de la fonction $x \longmapsto xe^{-x}$ sur $\mathbb{R}^+$).

3) Déterminer la loi de Y^2. En déduire celle de $n/\hat{a}_n$ grâce aux propriétés des lois *Gamma*.

4) Prouver que $2n/\hat{a}_n$ suit sous $\mathcal{H}_0$ une loi du χ^2 que l'on déterminera. On effectue 8 tirages de Y. On observe les valeurs de $Y_1,\ldots,Y_n$ suivantes :

$$0.13 \quad 2.68 \quad 0.58 \quad 0.41 \quad 0.32 \quad 1.62 \quad 0.61 \quad 1.46$$

Peut-on raisonnablement affirmer que $a = 1$? (tester au niveau 5%).

Corrigé de l'examen de deuxième session du 20 juin 2016

Ex 1. Killer bees in Texas

1) Notons $X_1, \dots, X_{10}$ les proportions d'abeilles tueuses en 1980 et $Y_1, \dots, Y_{10}$ celles en 1990. On suppose les X_i i.i.d. avec fonction de répartition continue F et les Y_i i.i.d. avec fonction de répartition continue G.
On teste $\mathcal{H}_0 : F = G$ i.e. les X_i et les Y_i ont même loi
contre $\mathcal{H}_1$: les Y_i prennent des valeurs plutôt plus grandes que les X_i, i.e. $F \geq G$ et $F \neq G$ (la loi des Y_i majore stochastiquement celle des $X - i$).
On classe $(X_1, \dots, X_{10}, Y_1, \dots, Y_{10})$ par ordre croissant. La statistique W_X est la somme des rangs des X_i.
Sous $\mathcal{H}_1$, les X_i prennent des valeurs plutôt moins grandes que les Y_i, donc W_X a tendance à être plus petit, donc la région de rejet de $\mathcal{H}_0$ est de la forme $\mathcal{R} = \{W_X \leq k_\alpha\}$.
Sous $\mathcal{H}_0$, la loi de W_X est tabulée, et la table indique que $P_{\mathcal{H}_0}(W_X \leq 83) \simeq 5,2\%$.
Par conséquent $\mathcal{R} = \{W_X \leq 83\}$ pour un test au niveau 6%.
Pour calculer la valeur observée de W_X, on classe les observations par ordre croissant et on additionne les rangs des X_i :

piège numéro	1	2	3	4	5	6	7	8	9	10
proportion en 1980	0,330	0,146	0,518	0,339	0,693	0,249	0,438	0,695	0,135	0,388
rang	8	3	15	9	19	6	13	20	1	11
proportion en 1990	0,360	0,177	0,524	0,447	0,140	0,392	0,534	0,263	0,157	0,566
rang	10	5	16	14	2	12	17	7	4	18

$W_X(\omega) = 8 + 3 + 15 + 9 + 19 + 6 + 13 + 20 + 1 + 11 = 105 \notin \mathcal{R}$ donc on accepte $\mathcal{H}_0$: la proportion d'abeille tueuses n'a pas augmenté.

2) On note $Z_i = Y_i - X_i$. On note m la médiane des Z_i et on teste $\mathcal{H}_0 : m = 0$ contre $\mathcal{H}_1 : m > 0$. On utilise comme statistique le nombre $S = \sum_{i=1}^{10} \mathbf{1}_{Z_i > 0}$ de Z_i positifs.

Sous $\mathcal{H}_0$ la loi de S est la $\mathcal{B}in(10, \frac{1}{2})$. Sous $\mathcal{H}_1$, S prend des valeurs plus élevées.
La table indique que $P_{\mathcal{H}_0}(S \leq 2) \simeq 0,055$ donc par symétrie de la $\mathcal{B}in(10, \frac{1}{2})$ on a $P_{\mathcal{H}_0}(S \geq 8) \simeq 5,5\%$. La région de rejet au niveau 6% est donc $\mathcal{R} = \{S \geq 8\}$.

piège numéro	1	2	3	4	5	6	7	8	9	10
proportion en 1980	0,330	0,146	0,518	0,339	0,693	0,249	0,438	0,695	0,135	0,388
proportion en 1990	0,360	0,177	0,524	0,447	0,140	0,392	0,534	0,263	0,157	0,566
signe de $Y_i - X_i$	+	+	+	+	−	+	+	−	+	+

On observe $S(\omega) = 8$ donc on accepte qu'il y a plus d'abeilles tueuses en 1990 qu'en 1980.

La taille de ce test du signe est $P_{\mathcal{H}_0}(S \geq 8)$ qui vaut $5,5\%$

Ex 2. Les clés de la réussite au bac

Notons $n_1 = 1684$ le nombre de réponses de lycéens, $n_2 = 1133$ le nombre de réponses de parents d'élèves et $n_3 = 1083$ le nombre de réponses d'enseignants. On note aussi N_{11} (respectivement N_{12}, N_{13} et N_{14}) le nombre de lycéens qui citent comme élément clé "bons

professeurs" (respectivement "travail régulier", "famille stable" et "chance"). On définit de même les N_{2j} à partir du nombre de parents qui citent chaque élément clé, et les N_{3j} à partir du nombre d'enseignants qui citent chaque élément clé.

On teste $\mathcal{H}_0$: $N_{ij} \sim \mathcal{B}in(n_i, p_j)$ pour tout i de 1 à 3 et tout j de 1 à 4
contre $\mathcal{H}_1$: les N_{ij} suivent les lois $\tilde{\mathcal{B}}in(n_i, p_{ij})$ avec des p_{ij} différents.
On fait un test du χ^2 d'homogénéité. On utilise la statistique

$$T = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{\left(N_{ij} - n_i \frac{N_{\bullet j}}{3900}\right)^2}{n_i \frac{N_{\bullet j}}{3900}} \quad \text{où } N_{\bullet j} = \sum_{i=1}^{3} N_{ij}$$

Sous $\mathcal{H}_0$, asymptotiquement quand l'effectif total tend vers l'infini, la loi de T se rapproche d'un $\chi^2((3-1)(4-1)) = \chi^2(6)$. Ici, l'effectif total est assez grand au sens où tous les effectifs théoriques observés $n_i \frac{N_{\bullet j}}{3900}(\omega)$ dépassent 5. Il suffit de vérifier ce point pour $n_3 = 1083$ qui est le plus petit des n_i et $N_{\bullet 4}(\omega) = 364$ qui est le plus petit des $N_{\bullet j}(\omega)$: $1083\frac{364}{3900} \simeq 101,8 > 5$.
On peut donc considérer que la loi de T est assez proche du $\chi^2(6)$.
Sous $\mathcal{H}_1$, T prend des valeurs plus élevées.

D'après la table de la $\chi^2(6)$, la région de rejet au niveau 5% est $\mathcal{R} = \{T \geq 12,592\}$.
Dans le programme `Scilab` fourni, les effectifs théoriques $n_i \frac{N_{\bullet j}}{3900}$ sont les composantes de `E2` et la formule correspondant au calcul de T est `sum((E2-obs).^2 ./E2)`. La valeur observée de T est donc donnée par `T5`
On observe $T(\omega) = 134,4115$ donc on rejette $\mathcal{H}_0$. Les lycéens, les parents et les enseignants n'ont globalement pas la même opinion sur l'élément déterminant de la réussite au bac.

Ex 3. L'engrais est-il efficace ?

1) Les $X_1,\ldots,X_{12}$ sont i.i.d. de loi $\mathcal{N}(m_1, \sigma_1^2)$ et les $Y_1,\ldots,Y_{12}$ sont i.i.d. de loi $\mathcal{N}(m_2, \sigma_2^2)$.
On réalise un test de Fisher d'égalité des variances : $\mathcal{H}_0 : \sigma_1 = \sigma_2$ contre $\mathcal{H}_1 : \sigma_1 \neq \sigma_2$.
En notant $V_X^* = \frac{1}{11}\sum_{i=1}^{12}(X_i - \overline{X})^2$ et $V_Y^* = \frac{1}{11}\sum_{i=1}^{12}(Y_i - \overline{Y})^2$ les variances empiriques débiaisées des deux échantillons, on construit la statistique $F = \frac{V_X^*}{V_Y^*}$. Elle suit sous $\mathcal{H}_0$ la loi de Fisher $\mathcal{F}(11,11)$. Sous $\mathcal{H}_1$, elle dévie vers des valeurs plus grandes ou plus petites. La région de rejet au niveau 5% est d'après la table

$$\mathcal{R} \simeq \{F \leq \frac{1}{3,474}\} \cup \{F \geq 3,474\} \simeq \{F \leq 0,288\} \cup \{F \geq 3,474\}$$

La valeur observée est

$$F(\omega) = \frac{\frac{12}{11}\frac{1}{12}\sum_{i=1}^{12}(X_i(\omega) - \overline{X}(\omega))^2}{\frac{12}{11}\frac{1}{12}\sum_{i=1}^{12}(Y_i(\omega) - \overline{Y}(\omega))^2} \simeq \frac{\frac{12}{11}(0,40)^2}{\frac{12}{11}(0,36)^2} \simeq \frac{40^2}{36^2} \simeq 1,23$$

donc on accepte l'hypothèse que les variances sont égales.

2) On suppose maintenant les $X_1,\ldots,X_{12}$ i.i.d. $\mathcal{N}(m_1, \sigma)$ et les $Y_1,\ldots,Y_{12}$ i.i.d. $\mathcal{N}(m_2, \sigma)$ avec le même σ. On réalise un test de Student de comparaison des moyennes.

$\mathcal{H}_0$: $m_1 = m_2$ (les parcelles traitées au nouvel engrais ne sont pas plus productives)

contre

$\mathcal{H}_1$: $m_1 > m_2$ (elles sont plus productives)

On utilise la statistique

$$T = \sqrt{\frac{1}{12} + \frac{1}{12}}\ \frac{\overline{X} - \overline{Y}}{\sqrt{W^*}} \quad \text{où} \quad W^* = \frac{1}{11+11}\left(\sum_{i=1}^{12}(X_i - \overline{X})^2 + \sum_{i=1}^{12}(Y_i - \overline{Y})^2\right)$$

Sous $\mathcal{H}_0$, la variable T suit la loi de Student de paramètre $11+11 = 22$. Sous $\mathcal{H}_1$, elle dévie dans le même sens que $m_1 - m_2$, vers des valeurs plus élevées. La zone de rejet au niveau 5% est donc $\{T > 1,717\}$.

$$T(\omega) = \frac{1}{\sqrt{6}}\ \frac{\overline{X}(\omega) - \overline{Y}(\omega)}{\sqrt{W^*(\omega)}} \text{ où } W^*(\omega) = \frac{1}{22}\left(12(0,40)^2 + 12(0,36)^2\right) \simeq 0,1579636$$

$$T(\omega) = \frac{1}{\sqrt{6}}\ \frac{5,10 - 4,80}{\sqrt{0,1579636}} \simeq 0,308$$

On conclut que le nouvel engrais n'augmente pas significativement la production.

Ex 4. 1) $Y_1,\ldots,Y_n$ est un échantillon de la loi de Y de densité $f_a(y) = a\ y\ e^{-ay^2}\mathbf{1}_{y\geq 0}$. La vraisemblance est

$$L(Y_1,\ldots,Y_n,a) = \prod_{i=1}^{n} f_a(Y_i) = a^n\left(\prod_{i=1}^{n} Y_i\right) e^{-a\sum_{i=1}^{n} Y_i^2}\mathbf{1}_{\min(Y_1,\ldots,Y_n)\geq 0}$$

Il est de probabilité 1 que tous les Y_i soient strictement positifs et dans ce cas la log-vraisemblance est

$$\ln L(Y_1,\ldots,Y_n,a) = n\ln(a) + \sum_{i=1}^{n}\ln(Y_i) - a\sum_{i=1}^{n} Y_i^2$$

On dérive par rapport à a.

$$\frac{d(\ln L)}{da}(Y_1,\ldots,Y_n,a) = \frac{n}{a} - \sum_{i=1}^{n} Y_i^2$$

$L(Y_1,\ldots,Y_n,a)$ est maximale en $\hat{a}_n = \dfrac{n}{\sum_{i=1}^{n} Y_i^2}$.

2) On veut construire le test du rapport de vraisemblance de l'hypothèse $\mathcal{H}_0 : a = 1$ contre l'hypothèse $\mathcal{H}_1 : a \neq 1$. On calcule le rapport de vraisemblance

$$V = \frac{\sup_{\mathcal{H}_1} L(X_1,\ldots,X_n,a)}{\sup_{\mathcal{H}_0} L(X_1,\ldots,X_n,a)} = \frac{\sup_{a\neq 1} L(X_1,\ldots,X_n,a)}{L(X_1,\ldots,X_n,1)} = \frac{\sup_{a>0} L(X_1,\ldots,X_n,a)}{L(X_1,\ldots,X_n,1)}$$

La valeur de a qui maximise la vraisemblance est $\hat{a}_n = \frac{n}{\sum_{i=1}^{n} Y_i^2}$ donc le numérateur est égal à

$$\sup_{a>0} L(X_1,\ldots,X_n,a) = L\left(X_1,\ldots,X_n,\hat{a}_n\right) = (\hat{a}_n)^n\left(\prod_{i=1}^{n} Y_i\right) e^{-n}$$

Le rapport de vraisemblance est donc

$$V = \frac{(\hat{a}_n)^n\left(\prod_{i=1}^{n} Y_i\right) e^{-n}}{1^n\left(\prod_{i=1}^{n} Y_i\right) e^{-\sum_{i=1}^{n} Y_i^2}} = \frac{(\hat{a}_n)^n\, e^{-n}}{e^{-n/\hat{a}_n}} = \left(\frac{\hat{a}_n}{e}\ e^{1/\hat{a}_n}\right)^n$$

Le région de rejet du test du rapport de vraisemblance est de la forme $\mathcal{R} = \{V \geq v_\alpha\}$ pour un réel v_α supérieur à 1 qui dépend du niveau α du test.

$$\mathcal{R} = \left\{ \left(\frac{\hat{a}_n}{e}\ e^{1/\hat{a}_n} \right)^n \geq v_\alpha \right\} = \left\{ \hat{a}_n\ e^{1/\hat{a}_n} \geq e\ (v_\alpha)^{1/n} \right\} = \left\{ \frac{1}{\hat{a}_n}\ e^{-1/\hat{a}_n} \leq \frac{1}{e\ (v_\alpha)^{1/n}} \right\}$$

La fonction $f\ :\ x \longmapsto xe^{-x}$ croît de 0 à $1/e$ sur $[0;1]$ puis décroît de $1/e$ à 0 sur $[1;+\infty[$ puisque sa dérivée est $x \longmapsto (1-x)e^{-x}$. Puisque $v_\alpha > 1$ on a $\frac{1}{e\ (v_\alpha)^{1/n}} < 1/e$ donc il existe $t_1, \alpha \in [0;1]$ et $t_2, \alpha \in [1;+\infty[$ tels que $f(t_1,\alpha) = f(t_2,\alpha) = \frac{1}{e\ (v_\alpha)^{1/n}}$ et

$$\mathcal{R} = \left\{ f\left(\frac{1}{\hat{a}_n}\right) \leq \frac{1}{e\ (v_\alpha)^{1/n}} \right\} = \left\{ \frac{1}{\hat{a}_n} \leq t_1, \alpha \quad \text{ou} \quad \frac{1}{\hat{a}_n} \geq t_2, \alpha \right\}$$

Elle est donc bilatère.

3) Déterminons la loi de Y^2. Pour t négatif, $P(Y^2 \leq t)$ est nul. Pour t positif

$$P(Y^2 \leq t) = P(Y \leq \sqrt{t}) = \int_0^{\sqrt{t}} 2a\ y\ e^{-ay^2} \mathbf{1}_{y\geq 0}\ dy = \left[-e^{-ay^2} \right]_0^{\sqrt{t}} = 1 - e^{-at}$$

donc Y^2 suit la loi exponentielle de paramètre a.

Par conséquent, $\frac{n}{\hat{a}_n} = \sum_{i=1}^n Y_i^2$ suit la loi $Gamma(n,a)$.

4) Toujours d'après les propriétés des lois $Gamma$, $\frac{2n}{\hat{a}_n} = \sum_{i=1}^n Y_i^2$ suit la loi $Gamma(n, \frac{a}{2})$. Sous $\mathcal{H}_0$, $2n/\hat{a}_n$ suit donc la loi $Gamma(\frac{2n}{2}, \frac{1}{2}) = \chi^2(2n)$. Sous $\mathcal{H}_1$, $2n/\hat{a}_n$ prend des valeurs plus grandes si $a > 1$ ou plus petites si $a < 1$.

Dans le cas où $n = 8$, d'après la table de la $\chi^2(16)$

$$P_{\mathcal{H}_0}\left(\frac{16}{\hat{a}_8} \leq 6,908 \text{ ou } \frac{16}{\hat{a}_8} \geq 28,845 \right) \simeq 5\%$$

On observe $\sum_{i=1}^n Y_i^2(\omega) = 12,9343$ donc $\hat{a}_8(\omega) \simeq 0,6185$ et $\frac{16}{\hat{a}_8}(\omega) \simeq 25,87$ donc on accepte $\mathcal{H}_0$. On peut raisonnablement affirmer que $a = 1$.

Examen du 18 mai 2016

durée : 3 heures

Matériel autorisé : table statistique, calculatrice, une feuille A4 manuscrite recto-verso.

Ex 1. Neuf malades reçoivent un traitement pour faire baisser le taux d'un produit nocif dans leur sang. Chez le malade numéro i, on mesure le taux X_i de ce produit, on administre le traitement, puis on mesure le taux Y_i après traitement. On observe les résultats suivants :

patient	1	2	3	4	5	6	7	8	9
X_i	1,82	0,50	1,62	2,05	1,68	1,88	1,55	3,14	1,29
Y_i	0,88	0,75	0,60	2,48	1,06	1,29	1,06	3,06	1,30

1) On veut tester niveau 5% l'hypothèse $\mathcal{H}_0$: *le traitement est sans effet* contre l'hypothèse $\mathcal{H}_1$: *le traitement fait baisser le taux*. Effectuer un test du signe au niveau 5%. Quelle est la taille de ce test ? Quelle est sa p-valeur ?

2) Le test du signe n'utilise pas complètement l'information contenue dans les données. On veut faire mieux. Effectuer un test "signe et rang" au niveau 5% pour déterminer si le traitement est efficace.

Ex 2. On observe des plantes à fleurs blanches, roses ou rouges et à feuilles entières ou découpées. On constate les effectifs suivants :

	feuilles entières	feuilles découpées	totaux
fleurs blanches	207	62	269
fleurs roses	400	159	559
fleurs rouges	195	74	269
totaux	802	295	1097

1) On veut déterminer si la couleur des fleurs et la forme des feuilles sont indépendantes. Précisez les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ qu'on choisit et le nom du test qu'on doit faire. Effectuer ce test, au niveau 5%.

2) Les plantes ont été obtenues par croisement entre une lignée à fleurs blanches et feuilles découpées et une lignée à fleurs rouges et feuilles entières. Les biologistes cherchent à déterminer si la couleur des fleurs est déterminée par un couple d'allèles, c'est-à-dire une paire de gènes "rouge" et "blanc". Si c'est le cas, un plant issu du croisement des deux lignées a une chance sur deux d'être à fleurs roses, une chance sur quatre d'être à fleurs blanches et une chance sur quatre d'être à fleurs rouges. Tester au niveau 5% si la couleur des fleurs est déterminée par un couple d'allèles.

Ex 3. Des paquets (ensembles de données) circulent sur un réseau informatique. Il se

produit, aléatoirement, des collisions qui nécessitent de réémettre certains paquets. En régime normal, la durée entre deux collisions sur ce réseau suit une loi exponentielle de paramètre 2. On observe six durées entre collision :

$$0,78 \quad 0,14 \quad 4,21 \quad 0,55 \quad 0,20 \quad 0,23$$

A l'aide d'un test de Kolmogorov-Smirnov au niveau 10%, déterminer si le réseau est en régime normal.

Ex 4. Un programme informatique effectue des tirages d'une variable aléatoire X selon la loi

$$\forall k \in \mathbb{N} \quad P(X = k) = (k+1)p^2(1-p)^k$$

Le paramètre $p \in]0;1[$ est une constante fixée dans le programme.

1) On fait n tirages indépendants $X_1,\ldots,X_n$ en utilisant le programme. On veut déterminer le paramètre p. Calculer l'estimateur du maximum de vraisemblance $\hat{p}_n$.

2) Si le programme est correctement paramétré, les tirages sont effectués avec le paramètre $p = \frac{1}{2}$. On veut construire le test du rapport de vraisemblance de l'hypothèse $\mathcal{H}_0 : p = \frac{1}{2}$ contre l'hypothèse $\mathcal{H}_1 : p \neq \frac{1}{2}$. Calculer en fonction des X_i le rapport de vraisemblance, et monter que la région de rejet peut s'écrire sous la forme

$$\mathcal{R} = \left\{ f(\overline{X}_n) \geq t_\alpha \right\} \quad \text{où } f(x) = x\ln(2x) - (2+x)\ln(2+x) \quad \text{et } \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

3) Pour déterminer entièrement la zone de rejet du test de niveau α, il faudrait pouvoir calculer la valeur de t_α. On a donc besoin de connaitre les quantiles de la loi de $f(\overline{X}_n)$ sous $\mathcal{H}_0$. Or, cette loi n'est pas tabulée. On peut cependant trouver ses quantiles en utilisant Scilab. Il faut pour cela faire un grand nombre de tirages. Ecrire un programme Scilab qui effectue un tirage de X avec le paramètre $p = \frac{1}{2}$ (on ne demande pas le programme qui fait un grand nombre de tirages de $f(\overline{X}_n)$ et calcule les quantiles). Pour construire ce programme, on pourra démontrer et utiliser le fait que X a même loi que $Y + Z - 2$ où Y et Z sont indépendantes de loi géométrique de paramètre p.

Corrigé de l'examen du 18 mai 2016

Ex 1. On note $Z_i = Y_i - X_i$. Pour effectuer les deux tests, on va avoir besoin des signes des Z_i et des rangs de leurs valeurs absolues. On calcule donc :

patient	1	2	3	4	5	6	7	8	9
X_i	$1,82$	$0,50$	$1,62$	$2,05$	$1,68$	$1,88$	$1,55$	$3,14$	$1,29$
Y_i	$0,88$	$0,75$	$0,60$	$2,48$	$1,06$	$1,29$	$1,06$	$3,06$	$1,30$
Z_i	$-0,94$	$0,25$	$-1,02$	$0,43$	$-0,62$	$-0,59$	$-0,49$	$-0,08$	$0,01$
signe de Z_i	$-$	$+$	$-$	$+$	$-$	$-$	$-$	$-$	$+$
rang de $\lvert Z_i\rvert$	8	3	9	4	7	6	5	2	1

1) On suppose que la loi commune des Z_i a une fonction de répartition continue. C'est réaliste car un taux n'a aucune raison de privilégier une valeur précise. On note m la médiane des Z_i. On teste $\mathcal{H}_0 : m = 0$ contre $\mathcal{H}_1 : m < 0$. On utilise comme statistique le nombre $S = \sum_{i=1}^{9} \mathbf{1}_{Z_i>0}$ de Z_i positifs.

Sous $\mathcal{H}_0$ la loi de S est la $\mathcal{B}in(9, \frac{1}{2})$. Sous $\mathcal{H}_1$, S prend des valeurs moins élevées. La table indique que $P_{\mathcal{H}_0}(S \leq 1) \simeq 0,020$ et $P_{\mathcal{H}_0}(S \leq 2) \simeq 0,090$. La région de rejet au niveau 5% est donc

$$\mathcal{R} = \{S \leq 1\}$$

On observe $S(\omega) = 3$ donc on accepte que le traitement est sans effet.

La taille du test est $P_{\mathcal{H}_0}(S \leq 1) \simeq 2\%$ et sa p-valeur est $P_{\mathcal{H}_0}(S \leq 3) \simeq 25,4\%$.

2) On suppose en plus que la loi commune des Z_i est symétrique par rapport à sa médiane. On teste toujours $\mathcal{H}_0 : m = 0$ contre $\mathcal{H}_1 : m < 0$, mais on utilise cette fois comme statistique la somme $T = \sum_{i=1}^{9} R(i)\mathbf{1}_{Z_i>0}$ des rangs des valeurs absolues de Z_i positifs.

Sous $\mathcal{H}_0$ la loi de T est tabulée. Sous $\mathcal{H}_1$, T prend des valeurs moins élevées. D'après la table, $P_{\mathcal{H}_0}(T \leq 8) \simeq 4,9\%$ donc la région de rejet au niveau 5% est

$$\mathcal{R} = \{T \leq 8\}$$

On observe $T(\omega) = 3 + 4 + 1 = 8$ donc on rejette l'idée que le traitement est sans effet : au vu de ces données, on peut dire qu'il est efficace.

Ex 2. 1) On pose $X_i = 1$ (respectivement $X_i = 2$ et $X_i = 3$) si les fleurs de la $i^{\text{ème}}$ plante sont blanches (respectivement roses et rouges). Et on note Y_i l'indicatrice du fait que la $i^{\text{ème}}$ plante est à feuilles découpées. On veut déterminer si la couleur des fleurs et la forme des feuilles sont indépendantes. On teste

$\mathcal{H}_0$: $P(X_1 = j, Y_1 = k) = P(X_1 = j)P(Y_1 = k)$ pour tout j de 1 à 3 et tout k valant 0 ou 1

contre

$\mathcal{H}_1$: l'une de ces égalités est fausse

On utilise un test du χ^2 d'indépendance. On note N_{jk} le nombre de plantes parmi les 1097 qui sont telles que $X_i = j$ et $Y_i = k$. Le tableau de l'énoncé regroupe en fait les effectifs observés des variables aléatoires suivantes :

	feuilles entières	feuilles découpées	totaux
fleurs blanches	N_{10}	N_{11}	$N_{1\bullet}$
fleurs roses	N_{20}	N_{21}	$N_{2\bullet}$
fleurs rouges	N_{30}	N_{31}	$N_{3\bullet}$
totaux	$N_{\bullet 0}$	$N_{\bullet 1}$	1097

On utilise la statistique

$$T = \sum_{j=1}^{3} \sum_{k=0}^{1} \frac{\left(N_{jk} - \frac{N_{j\bullet} N_{\bullet k}}{1097}\right)^2}{\frac{N_{j\bullet} N_{\bullet k}}{1097}}$$

Sous $\mathcal{H}_0$, asymptotiquement quand l'effectif total tend vers l'infini, la loi de T se rapproche d'un $\chi^2((3-1)(2-1)) = \chi^2(2)$. Ici, l'effectif total est assez grand au sens où les effectifs théoriques observés dépassent 5, donc on peut considérer que la loi de T est assez proche du $\chi^2(2)$. Le tableau suivant récapitule les effectifs observés et les effectifs théoriques observés :

	feuilles entières	feuilles découpées	totaux
fleurs blanches	$207//196,66$	$62//72,34$	269
fleurs roses	$400//408,68$	$159//150,32$	559
fleurs rouges	$195//196,66$	$74//72,34$	269
totaux	802	295	1097

Sous $\mathcal{H}_1$, T prend des valeurs plus élevées. D'après la table de la $\chi^2(2)$, la région de rejet au niveau 5% est

$$\mathcal{R} = \{T \geq 5,991\}$$

On observe $T(\omega) = 2,76$ donc on accepte $\mathcal{H}_0$. La couleur des fleurs est indépendante de la forme des feuilles.

2) On veut maintenant tester si la loi commune des X_i est bien donnée par

$$P(X_i = 1) = \frac{1}{4} \quad P(X_i = 2) = \frac{1}{2} \quad P(X_i = 3) = \frac{1}{4}$$

On teste $\mathcal{H}_0$: $(N_{1\bullet}, N_{2\bullet}, N_{3\bullet})$ suit la loi multinomiale de paramètres $(1097, \frac{1}{4}, \frac{1}{2}, \frac{1}{4})$
contre $\mathcal{H}_1$: $(N_{1\bullet}, N_{2\bullet}, N_{3\bullet})$ suit une loi multinomiale de paramètres différents
La statistique sera

$$T' = \frac{\left(N_{1\bullet} - \frac{1097}{4}\right)^2}{\frac{1097}{4}} + \frac{\left(N_{2\bullet} - \frac{1097}{2}\right)^2}{\frac{1097}{2}} + \frac{\left(N_{3\bullet} - \frac{1097}{4}\right)^2}{\frac{1097}{4}}$$

Sous $\mathcal{H}_0$, la loi de T se rapproche d'un $\chi^2(3-1) = \chi^2(2)$ quand l'effectif total tend vers l'infini. Ici, l'effectif total est assez grand à nouveau, au sens où les effectifs théoriques observés dépassent 5. Les effectifs observés et les effectifs théoriques observés sont :

	effectif	effectif théorique
fleurs blanches	269	274,25
fleurs roses	559	548,5
fleurs rouges	269	274,25
total	1097	1097

donc T' suit presque un $\chi^2(2)$ sous $\mathcal{H}_0$ et prend des valeurs plus grandes sous $\mathcal{H}_1$. La région de rejet est au niveau 5% est encore $\mathcal{R} = \{T' \geq 5,991\}$. L'observation $T'(\omega) \simeq 0,40$ conduit à accepter $\mathcal{H}_0$. La couleur des fleurs est déterminée par un couple d'allèles.

Ex 3. On va faire un test de Kolmogorov-Smirnov à un échantillon pour voir si la fonction de répartition des données X_i observées peut être celle de la loi exponentielle de paramètre 2. Il faut donc connaitre ou recalculer la fonction de répartition de X où X suit la loi exponentielle de paramètre 2 :

$$\forall t \in \mathbb{R} \quad F(t) = P(X \leq t) = \mathbf{1}_{t>0} \int_0^t 2e^{-2x}\, dx = \mathbf{1}_{t>0} \left[-e^{-2x}\right]_0^t = (1 - e^{-2t})\mathbf{1}_{t>0}$$

Cette fonction de répartition est bien continue puisque les exponentielles sont à densité : le test de Kolmogorov-Smirnov est faisable. On teste

$$\mathcal{H}_0 \ : \text{ les tirages ont pour fonction de répartition } F$$

contre

$$\mathcal{H}_1 : \text{ les tirages suivent une autre loi}$$

On note F_6 la fonction de répartition empirique des X_i et D_6 la statistique du test :

$$D_6 = \sup_{t \in \mathbb{R}} |F_6(t) - F(t)| \quad \text{où} \quad F_6(t) = \frac{1}{6} \sum_{i=1^6} \mathbf{1}_{X_i \leq t}$$

La loi de D_6 sous $\mathcal{H}_0$ est tabulée, et on sait qu'elle se décale vers les grandes valeurs sous $\mathcal{H}_1$. La zone de rejet est $\mathcal{R} = \{D_6 \geq 0,4680\}$ au niveau 10%. Pour calculer la valeur observée de D_6, on calcule les valeurs observées des $F(X_i)$ et $F_6(X_i)$

i	$X_{(i)}(\omega)$	$F(X_{(i)}(\omega))$	$i/6$	$(i-1)/6$	$F(X_{(i)}(\omega)) - i/6$	$F(X_{(i)}(\omega)) - (i-1)/6$
1	0,14	0,24	0,17	0	0,07	0,24
2	0,20	0,33	0,33	0,17	0,00	0,16
3	0,23	0,37	0,50	0,33	−0,13	0,04
4	0,55	0,67	0,67	0,50	0,00	0,17
5	0,78	0,79	0,83	0,67	−0,04	0,12
6	4,21	1,00	1	0,83	0,00	0,17

On observe $D_6(\omega) \simeq 0,24$ donc on accepte $\mathcal{H}_0$. Le réseau est en régime normal.

Ex 4. 1) $X_1,\ldots,X_n$ est un échantillon de la loi de X. La vraisemblance est

$$L(X_1, \ldots, X_n, p) = \prod_{i=1}^n (X_i + 1)p^2(1-p)^{X_i} = p^{2n}(1-p)^{\sum_{i=1}^n X_i} \prod_{i=1}^n (X_i + 1)$$

donc la log-vraisemblance est

$$\ln L(X_1, \ldots, X_n, p) = 2n \ln(p) + \ln(1-p) \left(\sum_{i=1}^n X_i \right) + \sum_{i=1}^n \ln(X_i + 1)$$

On dérive par rapport à p.

$$\frac{d(\ln L)}{dp}(X_1, \ldots, X_n, p) = \frac{2n}{p} - \frac{1}{1-p} \left(\sum_{i=1}^n X_i \right) = \frac{1}{p(1-p)} \left(2n - 2np - p \left(\sum_{i=1}^n X_i \right) \right)$$

$\ln L(X_1,\dots,X_n,p)$, et donc aussi $L(X_1,\dots,X_n,p)$ est maximale en

$$\hat{p}_n = \frac{2n}{2n+\sum_{i=1}^n X_i} = \frac{2}{2+\overline{X}_n}$$

2) Pour construire le test du rapport de vraisemblance de l'hypothèse $\mathcal{H}_0 : p = \frac{1}{2}$ contre l'hypothèse $\mathcal{H}_1 : p \neq \frac{1}{2}$ on calcule le rapport de vraisemblance

$$V = \frac{\sup_{\mathcal{H}_1} L(X_1,\dots,X_n,p)}{\sup_{\mathcal{H}_0} L(X_1,\dots,X_n,p)} = \frac{\sup_{p\neq 1/2} L(X_1,\dots,X_n,p)}{L(X_1,\dots,X_n,\frac{1}{2})} = \frac{\sup_{p\in]0;1[} L(X_1,\dots,X_n,p)}{L(X_1,\dots,X_n,\frac{1}{2})}$$

La valeur de p qui maximise la vraisemblance est $\hat{p}_n$:

$$V = \frac{L(X_1,\dots,X_n,\hat{p}_n)}{L(X_1,\dots,X_n,\frac{1}{2})} = \frac{\hat{p}_n^{2n}(1-\hat{p}_n)^{\sum_{i=1}^n X_i}\prod_{i=1}^n (X_i+1)}{(1/2)^{2n+\sum_{i=1}^n X_i}\prod_{i=1}^n (X_i+1)}$$

et en simplifiant

$$V = (2\hat{p}_n)^{2n}(2-2\hat{p}_n)^{\sum_{i=1}^n X_i} = \left((2\hat{p}_n)^2(2-2\hat{p}_n)^{\overline{X}_n}\right)^n$$

et $\hat{p}_n$ est a été calculé :

$$V = \left(\left(\frac{4}{2+\overline{X}_n}\right)^2\left(\frac{2\overline{X}_n}{2+\overline{X}_n}\right)^{\overline{X}_n}\right)^n = \left(16\,\frac{(2\overline{X}_n)^{\overline{X}_n}}{(2+\overline{X}_n)^{2+\overline{X}_n}}\right)^n$$

Le région de rejet du test du rapport de vraisemblance est de la forme

$$\mathcal{R} = \{V \geq v_\alpha\}$$

pour un réel v_α supérieur à 1. Puisque le logarithme est strictement croissant, la zone de rejet peut aussi s'écrire sous la forme $\mathcal{R} = \{\ln(V) \geq \ln(v_\alpha)\}$. On constate que

$$\ln V = n\left(\ln(16) + \overline{X}_n \ln(2\overline{X}_n) - (2+\overline{X}_n)\ln(2+\overline{X}_n)\right)$$

donc

$$\mathcal{R} = \left\{f(\overline{X}_n) \geq t_\alpha\right\} \quad \text{en notant } f(x) = x\ln(2x) - (2+x)\ln(2+x) \quad \text{et } t_\alpha = \frac{1}{n}\ln(v_\alpha) - \ln(16)$$

3) Prenons Y et Z indépendantes de loi géométrique de paramètre p. Pour tout $k \in \mathbb{N}$

$$P(Y+Z-2=k) = P(Y+Z=k+2) = \sum_{i=1}^{k+2-1} P(Y=i, Z=k+2-i)$$

$$= \sum_{i=1}^{k+2-1} p(1-p)^{i-1}\, p(1-p)^{k+2-i-1} = p^2 \sum_{i=1}^{k+2-1} (1-p)^{k+2-2} = p^2(k+1)(1-p)^k = P(X=k)$$

On peut donc faire un tirage de X avec le paramètre $p = 0,5$ à partir de la somme de deux tirages de la géométrique de paramètre $1/2$. Le programme suivant effectue ce tirage :

```
k=1 ; while rand(1,'uniform')>0.5 do k=k+1 ; end ; Y=k ;
k=1 ; while rand(1,'uniform')>0.5 do k=k+1 ; end ; Z=k ;
X=Y+Z-2
```

Devoir Surveillé du 16 mars 2016
durée : 2 heures
Matériel autorisé : table statistique, calculatrice.

Ex 1. Recrutement

1) Un enseignant examine les candidatures d'étudiants souhaitant entrer directement en deuxième ou troisième année d'une formation (réorientations). Les données sont les notes et appréciations de l'étudiant dans ses études antérieures. La décision à prendre est d'autoriser ou non l'inscription dans la formation. L'enseignant effectue un test d'hypothèses :
$\mathcal{H}_0$: l'étudiant est capable de suivre les enseignements de la formation
$\mathcal{H}_1$: l'étudiant n'a pas les connaissances et capacités pour suivre
Que représentent, dans cette situation particulière, l'erreur de première espèce et l'erreur de deuxième espèce ?

2) La Commission des Titres d'Ingénieur est l'organisme qui autorise certaines formations à délivrer le diplôme d'ingénieur. Cette commission évalue le taux d'échec des formations : les redoublement doivent être très peu nombreux, et les abandons (départs sans diplôme) extrèmement rares pour que la formation reste autorisée à donner le titre d'ingénieur. Un enseignant d'une formation d'ingénieurs examine des candidatures. Doit-il effectuer le même test que précédemment ou doit-il échanger les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$? Pourquoi ?

Ex 2. Cote de popularité

Dans un sondage réalisé fin février sur 959 personnes, 193 d'entre elles déclarent avoir une opinion favorable de François Hollande. Un journal affirme que la cote de popularité du président a baissé. Comme sa cote de popularité précédente était de 22%, on va vérifier cette affirmation en testant

$$\mathcal{H}_0 \ : p = 0,22 \quad \text{contre} \quad \mathcal{H}_1 \ : p < 0,22$$

1) Quelle statistique de test utilise-t-on ? Quelle est sa loi sous l'hypothèse $\mathcal{H}_0$? A-t-elle tendance sous $\mathcal{H}_1$ à prendre des valeurs plus grandes ou plus petites ? Donner la forme de la zone de rejet $\mathcal{R}_\alpha$.

2) Déterminer la zone de rejet $\mathcal{R}_{0,05}$ du test au niveau 5% et la zone de rejet $\mathcal{R}_{0,10}$ du test au niveau 10% (on approximera la loi de la statistique centrée renormalisée, et on se contentera de zones de rejet asymptotiques).

3) Quelle conclusion obtient-on avec le test de niveau 5% et celui de niveau 10% ?

4) Déterminer la p-valeur de ce test.

Ex 3. Matériel médiocre ou haut-de-gamme ?

Une entreprise achète une chaine d'assemblage comportant 4 moteurs de précision. La durée de vie d'un moteur de ce type, en heures de fonctionnement, est modélisée par une loi de Weibull de densité :

$$f_\lambda(x) = \frac{3x^2}{\lambda^3}\, e^{-x^3/\lambda^3} \mathbf{1}_{x\geq 0}$$

où λ est un paramètre strictement positif proportionnel à la durée de vie moyenne des moteurs. Le contrat prévoit que les moteurs de la chaine d'assemblage proviendront d'un constructeur européen réputé pour la fiabilité de son matériel : le paramètre de la Weibull est alors $\lambda = 8000$. Ce type de moteur est assez rare et il n'y a qu'un seul autre constructeur, dont les moteurs tombent en panne beaucoup plus vite, mais sont beaucoup moins coûteux : le paramètre de leur loi de Weibull est $\lambda = 4000$.

Après plusieurs pannes, l'entreprise soupçonne le vendeur d'avoir équipé la chaine de moteurs bon marché tout en facturant du matériel haut-de-gamme. Elle décide de faire sur les durées de fonctionnement X_1, X_2, X_3, X_4 des 4 moteurs un test de

$$\mathcal{H}_0 \ : \lambda = 8000 \quad \text{contre} \quad \mathcal{H}_1 \ : \lambda = 4000$$

1) On travaille dans le modèle statistique $(\Omega, \mathcal{F}, (\mathbb{P}_\lambda)_{\lambda \in \mathbb{R}_+^*})$, où λ représente le vrai paramètre (inconnu) de la loi de Weibull des 4 durées X_i de fonctionnement. On veut effectuer un test du rapport de vraisemblance. Donner la vraisemblance $L(X_1, X_2, X_3, X_4, \lambda)$. Déterminer une statistique de test T et indiquer la forme de la région de rejet. Vérifier que la position de cette zone de rejet est cohérente avec le comportement des X_i selon que les moteurs sont de bonne ou de mauvaise qualité.

2) Pour finir de déterminer la région de rejet, on a besoin d'utiliser la table de valeurs numériques. Or, les Weibull ne sont pas tabulées. Prouver que si, sous $\mathbb{P}_\lambda$, X suit la loi Weibull(λ) alors $2X^3/\lambda^3$ suit sous cette même probabilité $\mathbb{P}_\lambda$ une loi exponentielle dont on précisera le paramètre.

3) En utilisant les propriétés des lois Gamma, prouver que $2T/\lambda^3$ suit sous $\mathbb{P}_\lambda$ une loi du χ^2 dont on déterminera le nombre de degrés de liberté. En déduire la région de rejet $\mathcal{R}$ au niveau 5%.

4) Les durées de fonctionnement observées par l'entreprise pour les 4 moteurs sont $X_1(\omega) = 2227$, $X_2(\omega) = 3863$, $X_3(\omega) = 3852$ et $X_4(\omega) = 2024$ heures. Est-il justifié de penser qu'elle a été trompée sur l'origine du matériel vendu ?

5) Le vendeur conteste les accusations de l'entreprise et met en doute la qualité du test. Prouver que la puissance du test effectué ici dépasse 99%.

6) Y a-t-il moyen de faire avec ce choix d'hypothèses un test encore plus puissant, en changeant de statistique ou de région de rejet ?

Corrigé du Devoir Surveillé du 16 mars 2016

Ex 1. 1) L'erreur de première espèce consiste à décider que $\mathcal{H}_0$ est fausse alors qu'elle est vraie. L'erreur de deuxième espèce consiste à décider que $\mathcal{H}_0$ est vraie alors que c'est faux. L'enseignant effectue le test d'hypothèses
$\mathcal{H}_0$: l'étudiant est capable de suivre les enseignements de la formation
$\mathcal{H}_1$: l'étudiant n'a pas les connaissances et capacités pour suivre
Donc il a décidé que l'erreur de première espèce consiste à refuser l'inscription à un étudiant qui aurait pu réussir dans la formation. L'erreur de deuxième espèce consiste à inscrire un étudiant qui échouera dans cette formation.

2) Dans un test, on choisit $\mathcal{H}_0$ et $\mathcal{H}_1$ de façon à rendre peu probable l'erreur qu'on considère comme la plus grave, celle qu'on qualifie de "première espèce". Dans le test précédent, l'erreur de première espèce est de ne pas donner sa chance à un étudiant qui aurait pu réussir. L'erreur de deuxième espèce est de permettre des échecs (redoublements et abandons).

Le test précédent n'est absolument pas adapté aux formations dépendant de la Commission des Titres d'Ingénieur [1]. Pour ces écoles, le risque le plus grave est d'accepter l'inscription d'un étudiant qui redoublera ou échouera : sa présence met en danger l'existence de la formation donc la scolarité des autres ! L'enseignant d'école d'ingénieurs doit donc effectuer le test
$\mathcal{H}_0$: l'étudiant n'a pas les connaissances et capacités pour suivre la formation
$\mathcal{H}_1$: l'étudiant est capable de suivre les enseignements de cette formation

Ex 2. Cote de popularité

1) On considère les réponses au sondage comme indépendantes et de même loi. Le nombre de gens N qui déclarent avoir une opinion favorable de François Hollande sera notre statistique de test. Sous $\mathcal{H}_0$, elle suit la loi $\mathcal{B}in(959\ ;\ 0,22)$. Sous $\mathcal{H}_1$, elle suit la loi $\mathcal{B}in(959\ ;\ p)$ avec $p < 0,22$ donc elle a tendance à prendre des valeurs plus petites. La zone de rejet sera donc de la forme $\mathcal{R}_\alpha = \{N \leq t_\alpha\}$ avec un nombre t_α fixé par le niveau du test.

2) Grâce au théorème central limite, la loi approximative de $\frac{N-959\times 0,22}{\sqrt{959\times 0,22\times(1-0,22)}}$ sous $\mathcal{H}_0$ est la $\mathcal{N}(0;1)$. D'après la table, et en utilisant la symétrie de la $\mathcal{N}(0;1)$, on a

$$\mathbb{P}_{0,22}\left(\frac{N-959\times 0,22}{\sqrt{959\times 0,22\times(1-0,22)}} \leq -1,645\right) \simeq 0,05$$

donc

$$\mathbb{P}_{0,22}\left(N \leq 959\times 0,22 - 1,645\sqrt{959\times 0,22\times(1-0,22)}\right) \simeq \mathbb{P}_{0,22}\left(N \leq 189,88\right) \simeq 0,05$$

1. Cet exercice n'est pas fictif. Pour pouvoir délivrer un diplôme national, une formation doit être accréditée. La Commission des Titres d'Ingénieur s'occupe des grandes écoles. Le Ministère de l'Enseignement Supérieur évalue et accrédite les licences et masters.

La région de rejet asymptotique au niveau 5% est $\mathcal{R}_{0,05} = \{N \leq 189\}$.

De même, la table indique que le quantile à 10% de la $\mathcal{N}(0;1)$ est $-1,28$ donc

$$\mathbb{P}_{0,22}\left(N \leq 959 \times 0,22 - 1,28\sqrt{959 \times 0,22 \times (1-0,22)}\right) \simeq \mathbb{P}_{0,22}\left(N \leq 194,56\right) \simeq 0,10$$

et la région de rejet asymptotique au niveau 10% est $\mathcal{R}_{0,10} = \{N \leq 194\}$

3) On a observé $N(\omega) = 193$ réponses favorables à F. Hollande parmi les 959 sondés. $193 \leq 194$ i.e. $\omega \in \mathcal{R}_{0,10}$ donc au niveau 10% le test accepte l'hypothèse selon laquelle la popularité du président a baissé. Par contre, $193 \not\leq 189$ i.e. $\omega \notin \mathcal{R}_{0,05}$ donc au niveau 5% on rejette cette hypothèse.

Intuitivement, 193 est suffisamment inférieur à 211 (22% de 959) pour être considéré comme une baisse de popularité si on est prêt à prendre un assez gros risque de se tromper, mais pas si on a prévu de plafonner ce risque à 5%.

4) La p-valeur de ce test est entre 5% et 10% d'après la question précédente. Elle vaut :

$$\begin{aligned}\mathbb{P}_{0,22}\left(N \leq 193\right) &= \mathbb{P}_{0,22}\left(\frac{N - 959 \times 0,22}{\sqrt{959 \times 0,22 \times (1-0,22)}} \leq \frac{193 - 959 \times 0,22}{\sqrt{959 \times 0,22 \times (1-0,22)}}\right) \\ &\simeq \mathbb{P}_{0,22}\left(\frac{N - 959 \times 0,22}{\sqrt{959 \times 0,22 \times (1-0,22)}} \leq -1,40\right) \simeq 1 - 0,9193 = 0,0807\end{aligned}$$

Les tests de niveau $8,07\%$ et plus accepteront l'hypothèse de la baisse de popularité, les tests de niveau inférieur considéreront que l'écart observé n'est pas assez significatif pour conclure à une véritable baisse de popularité.

Ex 3. Matériel médiocre ou haut-de-gamme ?

1) On calcule la vraisemblance, en tenant compte du fait que les X_i sont des tirages de la loi Weibull(λ) donc $P(X_i \geq 0) = 1$. On a avec probabilité 1

$$L(X_1, X_2, X_3, X_4, \lambda) = \prod_{i=1}^{4} f_\lambda(X_i) = \prod_{i=1}^{4}\left(\frac{3X_i^2}{\lambda^3}\ e^{-X_i^3/\lambda^3}\mathbf{1}_{X_i \geq 0}\right) = \frac{3^4}{\lambda^{12}}(X_1X_2X_3X_4)^2 \exp\left(-\frac{1}{\lambda^3}\sum_{i=1}^{4} X_i^3\right)$$

Pour trouver une statistique de test, on calcule le rapport de vraisemblance de $\mathcal{H}_0 : \lambda = 8000$ contre $\mathcal{H}_1 : \lambda = 4000$. Il vaut

$$\begin{aligned}\frac{L(X_1, X_2, X_3, X_4, 4000)}{L(X_1, X_2, X_3, X_4, 8000)} &= \frac{8000^{12}}{4000^{12}} \exp\left(-\frac{1}{4000^3}\sum_{i=1}^{4} X_i^3 + \frac{1}{8000^3}\sum_{i=1}^{4} X_i^3\right) \\ &= 2^{12} \exp\left(\frac{1-2^3}{8000^3}\ \sum_{i=1}^{4} X_i^3\right) = 4096 \exp\left(\frac{-7}{8000^3}\ \sum_{i=1}^{4} X_i^3\right)\end{aligned}$$

La région de rejet est de la forme $\left\{\frac{L(X_1,X_2,X_3,X_4,4000)}{L(X_1,X_2,X_3,X_4,8000)} \geq v_\alpha\right\}$ avec un seuil v_α qui dépend du niveau du test. Comme la fonction $s \longmapsto 4096 \exp\left(\frac{-7}{8000^3}\ s\right)$ décroît quand s augmente, la région de rejet peut être écrite sous la forme $\left\{\sum_{i=1}^{4} X_i^3 \leq t_\alpha\right\}$. On utilise ici la statistique $T = \sum_{i=1}^{4} X_i^3$.

Ici, on teste $\mathcal{H}_0$: *les moteurs sont fiables* contre $\mathcal{H}_1$: *ils sont de mauvaise qualité*. Choisir une zone de rejet de forme $\left\{\sum_{i=1}^4 X_i^3 \leq t_\alpha\right\}$ conduit à rejeter l'idée que les moteurs sont fiables si la somme des cubes des durées de fonctionnement est assez petite, autrement dit les moteurs tombent vite en panne. C'est cohérent.

2) X suit la loi Weibull(λ) sous la probabilité $\mathbb{P}_\lambda$. On veut connaitre la loi de $2X^3/\lambda^3$ sous $\mathbb{P}_\lambda$. On calcule sa fonction de répartition. Pour chaque $t \in \mathbb{R}$

$$\mathbb{P}_\lambda\left(\frac{2X^3}{\lambda^3} \leq t\right) = \mathbb{P}_\lambda\left(X^3 \leq \frac{\lambda^3}{2}t\right)$$

Si t est négatif, ceci est nul puisque X est à valeurs dans $\mathbb{R}^+$. Si t est positif, on obtient

$$\mathbb{P}_\lambda\left(\frac{2X^3}{\lambda^3} \leq t\right) = \mathbb{P}_\lambda\left(X \leq \left(\frac{\lambda^3}{2}t\right)^{1/3}\right) = \int_{-\infty}^{\left(\frac{\lambda^3}{2}t\right)^{1/3}} f_\lambda(x)\, dx = \int_0^{\left(\frac{\lambda^3}{2}t\right)^{1/3}} \frac{3x^2}{\lambda^3}\, e^{-x^3/\lambda^3}\, dx$$

On effectue le changement de variable $y = x^3$ ($dy = 3x^2dx$)

$$\mathbb{P}_\lambda\left(\frac{2X^3}{\lambda^3} \leq t\right) = \int_0^{\frac{\lambda^3}{2}t} \frac{1}{\lambda^3}\, e^{-y/\lambda^3}\, dy$$

et pour se ramener au lien entre fonction de répartition et densité, donc voir réapparaitre t dans la borne, on change à nouveau de variable : $u = 2y/\lambda^3$ ($du = \frac{2}{\lambda^3}dy$)

$$\mathbb{P}_\lambda\left(\frac{2X^3}{\lambda^3} \leq t\right) = \int_0^t \frac{1}{2}\, e^{-u/2}\, du$$

On a prouvé que $\frac{2X^3}{\lambda^3}$ a pour densité $u \longmapsto \frac{1}{2}\, e^{-u/2}\mathbf{1}_{u\geq 0}$ donc qu'il suit la loi exponentielle de paramètre $\frac{1}{2}$ sous $\mathbb{P}_\lambda$.

Remarque : on pouvait bien sûr faire directement le changement de variable $y = \frac{2x^3}{\lambda^3}$.

3) Sous $\mathbb{P}_\lambda$, les $\frac{2X_i^3}{\lambda^3}$ sont indépendantes et toutes de loi exponentielle de paramètre $\frac{1}{2}$, i.e. de loi Gamma$(1, \frac{1}{2})$. D'après les propriétés des lois Gamma, leur somme $2T/\lambda^3$ suit la loi Gamma$(4, \frac{1}{2})$ qui est aussi la loi du χ^2 à 8 degrés de liberté. D'après la table de la $\chi^2(8)$

$$\mathbb{P}_{8000}\left(\frac{2T}{8000^3} \leq 2,733\right) \simeq 0,05$$

La région de rejet est donc

$$\mathcal{R} = \left\{\sum_{i=1}^4 X_i^3 \leq 2,733 \times \frac{8000^3}{2}\right\} = \left\{\sum_{i=1}^4 X_i^3 \leq 6,996.10^{11}\right\}$$

4) Les données recueillies par l'entreprise correspondent à l'observation de $T(\omega) \simeq 1,3.10^{11}$ donc on rejette $\mathcal{H}_0$. Les moteurs de la chaine d'assemblage vendue proviennent du fabriquant bas-de-gamme.

5) La puissance de ce test est

$$\mathbb{P}_{4000}(\mathcal{R}) = \mathbb{P}_{4000}\left(\sum_{i=1}^4 X_i^3 \leq 2,733 \times \frac{8000^3}{2}\right)$$

On sait que sous $\mathbb{P}_{4000}$ la v.a. $\frac{2}{4000^3}\sum_{i=1}^4 X_i^3$ suit la loi $\chi^2(8)$.

$$\mathbb{P}_{4000}(\mathcal{R}) = \mathbb{P}_{4000}\left(\frac{2}{4000^3}\sum_{i=1}^4 X_i^3 \leq 2,733 \times 8\right) = \mathbb{P}_{4000}\left(\frac{2}{4000^3}\sum_{i=1}^4 X_i^3 \leq 21,864\right)$$

La table de la $\chi^2(8)$ indique que $\mathbb{P}_{4000}\left(\frac{2}{4000^3}\sum_{i=1}^4 X_i^3 \leq 20,090\right) \simeq 0,99$ ce qui prouve que la puissance $\mathbb{P}_{4000}(\mathcal{R})$ de ce test dépasse 99%.

6) Il s'agit d'un test du rapport de vraisemblance entre deux hypothèses simples, et sa taille est 5%. D'après le théorème de Neyman-Pearson, il n'existe pas de test de niveau 5% ayant une puissance plus grande.