Tests Statistiques

Rejeter, ne pas rejeter... Se risquer ?

Magalie Fromont

Année universitaire 2015-2016

Table des matières

Chapitre 1

Introduction

On considère un phénomène aléatoire modélisé par une variable aléatoire X, dont la loi de probabilité P_θ est connue à un paramètre $\theta \in \Theta$ (inconnu) prés. Comme dans le cadre d'un problème d'estimation, on dispose d'une observation x de cette variable X (souvent, X est un n-échantillon $(X_1, \dots, X_n)$ d'une variable aléatoire parente et $x = (x_1, \dots, x_n)$ est l'observation de cet échantillon, ou bien $X = (Y, Z)$, où $Y = (Y_1, \dots, Y_{n_1})$ et $Z = (Z_1, \dots, Z_{n_2})$ sont deux échantillons indépendants et $x = (y, z)$ est l'observation de ces échantillons).
Dans le cadre d'un problème de test qui nous intéresse ici, on dispose en outre d'informations laissant penser a priori que le paramètre θ appartient à un sous-ensemble Θ_0 de Θ, et on cherche à valider ou invalider cette hypothèse sur la base de l'observation x.

1.1 Problèmes de tests

1.1.1 Modèles statistiques : rappels

Définition 1 (Modèle statistique)**.** *Le* modèle statistique *correspondant à la variable X est défini par le triplet $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$, où $\mathcal{X}$ est l'ensemble des valeurs possibles de l'observation x de X, $\mathcal{A}$ est une tribu sur $\mathcal{X}$, P_θ est la loi de X, dépendant d'un paramètre θ, inconnu, supposé appartenir à un ensemble Θ fixé.*

Définition 2 (Modèle identifiable)**.** *Le modèle statistique $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ est dit* identifiable *si l'application $\theta \mapsto P_\theta$ est injective.*

Définition 3 (Modèle paramétrique/non paramétrique)**.** *Le modèle statistique $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ est dit* paramétrique *si $\Theta \subset \mathbb{R}^d$ pour $d \in \mathbb{N}^*$,* non paramétrique *sinon.*

1.1.2 Hypothèse nulle pas si nulle, hypothèse alternative

Définition 4 (Hypothèse statistique)**.** *Faire une hypothèse statistique sur le paramètre θ consiste à se donner un sous-ensemble Θ_δ de Θ, et à énoncer que θ appartient à Θ_δ. L'hypothèse s'écrit alors $(H_\delta) \ : \ \theta \in \Theta_\delta$.*

Définition 5 (Problème de test, hypothèse nulle/alternative)**.** *Se poser un* problème de test *consiste à :*

1. *Considérer deux hypothèses (H_0) et (H_1), appelées respectivement* hypothèse nulle *et* hypothèse alternative, *correspondant à des sous-ensembles disjoints Θ_0 et Θ_1 de Θ.*

2. *Chercher à décider, sur la base de l'observation x, laquelle des deux hypothèses (H_0) et (H_1) est vraie.*

On dit alors qu'on teste $(H_0) : \theta \in \Theta_0$ *contre* $(H_1) : \theta \in \Theta_1$.
Accepter (H_0) *revient à décider que* (H_0) *est vraie, rejeter* (H_0) *au profit de* (H_1) *revient à décider que* (H_1) *est vraie.*

Exemple courant et historique : $X = (X_1, \dots, X_n)$ est un échantillon d'une loi d'espérance θ, modélisant la diffèrence entre une quantité mesurée avant un événement, et la même quantité mesurée après cet événement. On teste l'hypothèse nulle $(H_0) : \theta \in \Theta_0 = \{0\}$ (absence d'effet moyen de l'événement) contre l'hypothèse alternative $(H_1) : \theta \in \Theta_1 = \Theta \backslash \{0\}$, ou encore $(H_0) : \theta = 0$ contre $(H_1) : \theta \neq 0$.

Cet exemple est à l'origine de la dénomination de l'hypothèse nulle... qui n'est en fait pas si "nulle", par le rôle privilégié qui lui est attribué.

Dissymétrie des hypothèses

En théorie des tests classique issue du principe de Neyman et Pearson, les hypothèses nulle et alternative ne jouent pas des rôles symétriques. (H_0) est l'hypothèse que l'on privilégie, dans le sens où elle est présumée vraie tant que l'observation x ne conduit pas à la rejeter au profit de (H_1). L'analogie entre un problème de test d'hypothèses et un procès d'assises peut aider à comprendre. Dans un procès d'assises, tout suspect est présumé innocent tant qu'on n'apporte pas la preuve de sa culpabilité, preuve qui doit de plus s'appuyer sur des éléments matériels. De la même façon, dans un problème de test de (H_0) contre (H_1), (H_0) est présumée vraie tant qu'on n'apporte pas de preuve (x constitue les éléments matériels ici) contre elle au profit de (H_1). Donc, accepter (H_0), c'est seulement dire qu'on n'a pas pu, au vu de l'observation x, la rejeter au profit de (H_1). Accepter (H_0), c'est "acquitter faute de preuve". On préférera donc souvent dire dans ce cas que l'on ne rejette pas (H_0) au profit de (H_1).

Différentes formes d'hypothèses

— Hypothèses simples : $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, i.e. $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta = \theta_1$. Exemples dans des cadres paramétriques et non paramétriques.
— Hypothèse simple contre hypothèse composite : $\Theta_0 = \{\theta_0\}$ et Θ_1 contient au moins deux éléments. Par exemple, dans le cas où $\Theta = \mathbb{R}$,

1. $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \mathbb{R}\backslash\{\theta_0\}$, i.e. $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta \neq \theta_0$. Le problème de test est dit *bilatère*.
2. $\Theta_0 = \{\theta_0\}$ et $\Theta_1 =]\theta_0, +\infty[$, i.e. $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta > \theta_0$, ou $\Theta_0 = \{\theta_0\}$ et $\Theta_1 =]-\infty, \theta_0[$, i.e. $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta < \theta_0$. Le problème de test est dit *unilatère*.

Exemples des tests non paramétriques d'adéquation à une loi.
— Hypothèses composites : Θ_0 et Θ_1 contiennent au moins deux éléments. Par exemple, dans le cas où $\Theta = \mathbb{R}$,

1. $\Theta_0 =]-\infty, \theta_0]$ et $\Theta_1 =]\theta_0, +\infty[$, i.e. $(H_0) : \theta \leq \theta_0$ contre $(H_1) : \theta > \theta_0$, ou $\Theta_0 = [\theta_0, +\infty[$ et $\Theta_1 =]-\infty, \theta_0[$, i.e. $(H_0) : \theta \geq \theta_0$ contre $(H_1) : \theta < \theta_0$. Le problème de test est dit *unilatère*.

2. $\Theta_0 = [\theta_1, \theta_2]$ et $\Theta_1 =]-\infty, \theta_1[\cup]\theta_2, +\infty[$. Le problème de test est dit *bilatère*.

Exemples des tests paramétriques de comparaison.
Exemples des tests non paramétriques d'appartenance à une famille de lois.
Exemples des tests non paramétriques d'indépendance.

1.2 Exemple à boire sans modération

Le ministère de la santé étudie régulièrement la nécessité de prendre des mesures contre la consommation d'alcool, et l'efficacité de telles mesures. L'INSEE fournit à cet effet notamment des données annuelles sur la consommation moyenne d'alcool par personne et par jour. En janvier 1991, la loi Evin limite la publicité pour les boissons alcoolisées, et une campagne de publicité "Tu t'es vu quand t'as bu ?" est lancée en parallèle. La consommation moyenne d'alcool pur* par personne de plus de 15 ans et par jour (en g) est supposée suivre une loi gaussienne $\mathcal{N}(m, \sigma^2)$ avec $\sigma = 2$. En 1990, avant la loi Evin, m est supposée égale à 35. Le ministère juge qu'un objectif immédiat à atteindre est que m passe de 35 à 33. Sur l'observation des consommations moyennes d'alcool pour les années 1991 à 1994 comprises, le ministère se fixe la règle de décision suivante : si la moyenne des consommations d'alcool par personne et par jour sur ces 4 années est supérieure au seuil de 34.2, les mesures prises sont jugées inefficaces.

* 10 grammes d'alcool pur correspondent à un verre de boisson alcoolisée servi dans un café soit à peu prés 10 cl de vin à 12,5 degrés, 25 cl de bière à 5 degrés, 6 cl de porto à 20 degrés et 3 cl de whisky ou autre spiritueux à 40 degrés.

Questions.
- Quels sont les risques associés à cette règle de décision arbitraire ?
- Peut-on se donner des critères objectifs pour fixer un autre seuil que 34.2 ?

1.2.1 Problème de test envisagé

Soit $X = (X_1, X_2, X_3, X_4)$ la variable aléatoire modélisant la consommation moyenne d'alcool pur par personne de plus de 15 ans et par jour au cours des années 1991 à 1994, et $x = (x_1, x_2, x_3, x_4)$ l'observation de cette variable.
On suppose (même si cela peut être critiquable) que la loi de X est de la forme $P_\theta = \mathcal{N}(\theta, 4)^{\otimes 4}$ avec θ paramètre inconnu ($\theta = m \in \Theta =]0, +\infty[$).
On souhaite, au vu de l'observation x, trancher entre les deux hypothèses $\theta = 33$ ou $\theta = 35$. On privilégie ici l'hypothèse $\theta = 33$ dans le sens où elle est présumée vraie tant qu'on n'a pas d'éléments pour lui préférer l'hypothèse $\theta = 33$, i.e. on ne souhaite pas la rejeter à tort. On veut donc tester $(H_0) \ : \ \theta = 33$ contre $(H_1) \ : \ \theta = 35$.

1.2.2 Règle de décision et risques associés

La règle de décision retenue par le ministère se traduit de la façon suivante :
- Si $\frac{1}{4}\sum_{i=1}^{4} x_i \geq s$, on rejette (H_0) au profit de (H_1).
- Si $\frac{1}{4}\sum_{i=1}^{4} x_i < s$, on ne rejette pas (H_0) au profit de (H_1).

Le seuil s est ici fixé arbitrairement à 34.2.

Chacune des décisions possibles a pour conséquence une erreur éventuelle :
- Décider que les mesures n'ont pas été efficaces (rejeter (H_0) au profit de (H_1)) alors qu'elles l'ont été (alors que (H_0) est vraie), et prendre de nouvelles mesures.
- Décider que les mesures ont été efficaces (accepter ou ne pas rejeter (H_0) au profit de (H_1)) alors qu'elles ne l'ont pas été (alors que (H_1) est vraie), et ne rien faire.

Le modèle statistique posé rend possible le calcul des probabilités ou risques associés à ces deux erreurs.
Soit α la probabilité de rejeter (H_0) au profit de (H_1) alors que (H_0) est vraie.

$$\alpha = P_{33}\left(\left\{x = (x_1, x_2, x_3, x_4), \frac{1}{4}\sum_{i=1}^{4} x_i \geq 34.2\right\}\right),$$

que l'on notera $P_{33}\left(\frac{1}{4}\sum_{i=1}^{4} X_i \geq 34.2\right)$, posant que X suit ici la loi $P_{33} = \mathcal{N}(33,4)^{\bigotimes 4}$.
Lorsque X suit la loi P_{33}, $\frac{1}{4}\sum_{i=1}^{4} X_i = \bar{X}$ suit la loi $\mathcal{N}(33,1)$, et $(\bar{X} - 33) \sim \mathcal{N}(0,1)$. On a donc

$$\alpha = P(N \geq 34.2 - 33),$$

où $N \sim \mathcal{N}(0,1)$, et finalement, si F désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$,

$$\alpha = 1 - F(1.2) = 0.1151.$$

De la même façon, on peut calculer la probabilité β de ne pas rejeter (H_0) au profit de (H_1) alors que (H_1) est vraie :

$$\beta = P_{35}\left(\bar{X} < 34.2\right) = F(34.2 - 35) = 0.2119.$$

Ces probabilités ne sont pas équivalentes, la deuxième étant supérieure à la première.
On a donc ici contrôlé en priorité le risque de décider que les mesures n'ont pas été efficaces (rejeter (H_0) au profit de (H_1)) alors qu'elles l'ont été (alors que (H_0) est vraie).
Si l'on souhaite maintenant que ce risque ne dépasse pas une valeur bien précise, par exemple 0.05, un seuil s différent de 34.2 doit être choisi.
Sachant que $P_{33}\left(\bar{X} \geq s\right) = 1 - F(s - 33)$, et que le 0.95-quantile de la loi $\mathcal{N}(0,1)$ vaut 1.645, $s = 34.645$ convient. On aura ainsi $P_{33}\left(\bar{X} \geq s\right) = 0.05$.
Il est à noter que le deuxième risque vaut alors $\beta = P_{35}\left(\bar{X} < 34.645\right) = F(34.645 - 35) = 0.3613$.

Cet exemple va nous permettre de définir la notion de test statistique (non randomisé), ainsi que les notions qui s'y rattachent.

Chapitre 2

Tests statistiques : premières pierres et construction

2.1 Tests statistiques (non randomisés)

Définition 6 (Test statistique (non randomisé), région de rejet ou critique)**.** *Un test statistique (non randomisé) consiste en une partition de $\mathcal{X}$ en deux ensembles : l'ensemble des valeurs possibles de x conduisant au rejet de* (H_0) *au profit de* (H_1)*, noté $\mathcal{R}_{(H_0)}$ et appelé* région de rejet *ou* région critique *du test, et son complémentaire dans $\mathcal{X}$ qui correspond à l'ensemble des valeurs possibles de x ne conduisant pas au rejet de* (H_0) *au profit de* (H_1)*.*

Exemple de la Section 1.2 : le test considéré par le ministère correspond à $\mathcal{R}_{(H_0)} = \{x, \bar{x} \geq 34.2\}$, celui considéré ensuite à $\mathcal{R}_{(H_0)} = \{x, \bar{x} \geq 34.645\}$.

Un test statistique (non randomisé) est donc caractérisé par sa région critique, ou encore par la fonction indicatrice de sa région critique.

Définition 7 (Fonction de test (non randomisé))**.** *On appelle* fonction du test (non randomisé) *de région critique $\mathcal{R}_{(H_0)}$ la statistique ϕ de $\mathcal{X}$ dans $\{0, 1\}$, telle que $\phi(x) = 1$ si et seulement si $x \in \mathcal{R}_{(H_0)}$. Autrement dit, si $\phi(x) = 1$, on rejette* (H_0) *au profit de* (H_1)*, et si $\phi(x) = 0$, on ne rejette pas* (H_0) *au profit de* (H_1)*.*

Remarques.

— On peut voir $\phi(x)$ comme la probabilité de rejeter (H_0) au profit de (H_1) avec le test ϕ à partir de l'observation x.

— Un test statistique étant entièrement caractérisé par sa fonction de test, on confondra souvent les deux dénominations.

Définition 8 (Statistique de test, valeurs critiques)**.** *La région critique d'un test s'écrit en général sous l'une des formes suivantes : $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s\}$, $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s\}$, $\mathcal{R}_{(H_0)} = \{x, |T(x)| \geq s\}$, $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s_1\} \cup \{x, T(x) \geq s_2\}$, ou $\mathcal{R}_{(H_0)} = \{x, T(x) \in [s_1, s_2]\}$ $(s_1 < s_2)$, où T est une statistique sur $\mathcal{X}$ appelée* statistique de test*, et où s, s_1 et s_2 sont appelées* les valeurs critiques *du test.*

Exemple de la Section 1.2 : $T(x) = \bar{x}$ est la statistique des tests considérés, $s = 34.2$ puis $s = 34.645$ sont les valeurs critiques des tests considérés.

2.2 Erreurs et risques de décision

Prendre la décision de rejeter ou non une hypothèse nulle au profit d'une hypothèse alternative, c'est prendre le risque de commettre une erreur. Deux types d'erreurs sont possibles. Si l'on rejette l'hypothèse nulle (H_0) au profit de (H_1) alors que (H_0) est vraie, on parle d'*erreur de première espèce* et, si l'on ne rejette pas (H_0) au profit de (H_1) alors que (H_1) est vraie, on parle d'*erreur de deuxième espèce*.

		Réalité	
		(H_0)	(H_1)
Décision	(H_0)	décision correcte	erreur de 2ème espèce
	(H_1)	erreur de 1ère espèce	décision correcte

2.2.1 Risque de première espèce, niveau et taille, probabilité critique

A chaque type d'erreur correspond une probabilité ou un risque de commettre cette erreur.

Définition 9 (Risque de première espèce)**.** *Le* risque de première espèce *d'un test ϕ est l'application, notée α, qui à chaque $\theta_0 \in \Theta_0$ donne la probabilité P_{θ_0} de rejeter (H_0) avec ϕ :*

$$\begin{aligned} \alpha \ : \Theta_0 &\to [0,1] \\ \theta_0 &\mapsto \alpha(\theta_0) = P_{\theta_0}(\mathcal{R}_{(H_0)}) = P_{\theta_0}(\{x, \phi(x) = 1\}). \end{aligned}$$

Pour simplifier les notations, on pourra écrire $P_{\theta_0}(\phi(X) = 1)$ pour désigner la probabilité $P_{\theta_0}(\{x, \phi(x) = 1\})$.

On remarque ici que $\alpha(\theta_0) = \mathbb{E}_{\theta_0}[\phi(X)]$, où $\mathbb{E}_{\theta_0}[\phi(X)]$ désigne l'espérance de $\phi(X)$ lorsque X est de loi P_{θ_0}.

Définition 10 (Risque de première espèce maximal)**.** *Le* risque de première espèce maximal *d'un test ϕ est donné par* $\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0)$.

Définition 11 (Test de niveau/taille α_0)**.** *Soit $\alpha_0 \in [0,1]$.*
Un test ϕ est de niveau α_0 *si son risque de première espèce maximal est inférieur ou égal à α_0 i.e. :*

$$\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0) \leq \alpha_0 \quad (\textit{inéquation du niveau}).$$

Un test ϕ est de taille α_0 *si son risque de première espèce maximal est égal à α_0 i.e. :*

$$\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0) = \alpha_0 \quad (\textit{équation de la taille}).$$

Exemple de la Section 1.2 : le test $\phi(x) = \mathbb{1}_{\bar{x} \geq 34.645}$ est de taille 0.05, donc également de niveau 0.05.

Définition 12 (Probabilité critique ou p-valeur)**.**

— *Si $\mathcal{R}_{(H_0)}$ est de la forme $\{x, T(x) \leq s\}$, où T est une statistique de test, s une valeur critique, on définit la* probabilité critique *ou p-*valeur *(p-*value *en anglais) du test de région critique $\mathcal{R}_{(H_0)}$ comme*

$$p(x) = \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T(X) \leq T(x)).$$

— *Si $\mathcal{R}_{(H_0)}$ est de la forme $\{x, T(x) \geq s\}$, on définit la* probabilité critique *ou p-*valeur *du test de région critique $\mathcal{R}_{(H_0)}$ comme*

$$p(x) = \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T(X) \geq T(x)).$$

Proposition 1. *Dans le cas où $\Theta_0 = \{\theta_0\}$ et où la loi de $T(X)$ lorsque $X \sim P_{\theta_0}$ est continue, si ϕ est un test de taille α_0, dont la région critique $\mathcal{R}_{(H_0)}$ est de l'une des formes ci-dessus, $\phi(X) = 1 \Leftrightarrow p(X) \leq \alpha_0$ P_{θ_0}-p.s. La p-valeur du test correspond donc à la valeur maximale du niveau pour laquelle on ne rejette pas l'hypothèse (H_0) au profit de (H_1) sur la base de l'observation x.*

Exemple de la Section 1.2 : les tests considérés sont de la forme $\phi(x) = \mathbb{1}_{\bar{x} \geq s}$, et puisque $(x_1, x_2, x_3, x_4) = (34.7, 34.4, 33.7, 33.3)$, leur p-valeur est $p(x) = P_{33}(\bar{X} \geq 34.025) = 0.1527$. Puisque $p(x) \geq 0.05$, on ne rejette pas (H_0) au profit de (H_1) pour un niveau 0.05. En effet, on a bien $\bar{x} < 34.645$.

2.2.2 Risque de deuxième espèce, fonction puissance

Définition 13 (Risque de deuxième espèce)**.** *Le* risque de deuxième espèce *d'un test ϕ est l'application, notée β, qui à chaque $\theta_1 \in \Theta_1$ associe la probabilité P_{θ_1} de ne pas rejeter (H_0) au profit de (H_1) avec ϕ :*

$$\begin{aligned} \beta : \Theta_1 &\to [0,1] \\ \theta_1 &\mapsto \beta(\theta_1) = P_{\theta_1}(\mathcal{X} \backslash \mathcal{R}_{(H_0)}) = P_{\theta_1}(\{x, \phi(x) = 0\}). \end{aligned}$$

Remarque. $\beta(\theta_1) = 1 - \mathbb{E}_{\theta_1}[\phi(X)]$.

Définition 14 (Fonction puissance)**.** *La* fonction puissance *du test ϕ est l'application, notée γ, qui à chaque $\theta_1 \in \Theta_1$ associe la probabilité P_{θ_1} de rejeter (H_0) au profit de (H_1) avec ϕ :*

$$\begin{aligned} \gamma : \Theta_1 &\to [0,1] \\ \theta_1 &\mapsto 1 - \beta(\theta_1) = P_{\theta_1}(\mathcal{R}_{(H_0)}) = P_{\theta_1}(\{x, \phi(x) = 1\}). \end{aligned}$$

Remarques. $\gamma(\theta_1) = \mathbb{E}_{\theta_1}[\phi(X)]$. Par ailleurs, lorsque $\Theta_1 = \{\theta_1\}$, on parle de puissance du test. La fonction puissance est le prolongement de la fonction α sur Θ_1.

Exemple de la Section 1.2 : le test $\phi(x) = \mathbb{1}_{\bar{x} \geq 34.645}$ est, comme on l'a vu, de niveau 0.05. Son risque de deuxième espèce est égal à 0.3613 et sa puissance à $1 - 0.3613 = 0.6387$ (test peu puissant).

2.3 Construction de tests (non randomisés)

2.3.1 Le principe de Neyman et Pearson

Pour des hypothèses (H_0) et (H_1) fixées, un test de (H_0) contre (H_1) "idéal" serait un test de risques de première et deuxième espèces minimum. Cependant, on peut remarquer notamment qu'un test qui consiste à toujours accepter (H_0) a un risque de première espèce nul, mais un risque de deuxième espèce maximum. Réciproquement, un test qui consiste à toujours rejeter (H_0) au profit de (H_1) a un risque de deuxième espèce nul, mais un risque de première espèce maximum. A taille d'échantillon fixée, diminuer le risque de première espèce ne peut se faire en général qu'au détriment du risque de deuxième espèce et vice versa (principe des vases communicants). Il est donc impossible, dans la plupart des cas, de trouver un test minimisant à la fois les risques de première et deuxième espèces. Afin de sortir de cette impasse, Neyman et Pearson proposent, en 1933, de traiter les deux risques de façon non symétrique, définissant ainsi le principe de dissymétrie des hypothèses. Lors de la construction d'un test, ils choisissent de considérer comme prioritaire de contrôler l'un des deux risques, celui de première espèce, par une valeur α_0 fixée a priori (en pratique 0.01, 0.05 voire 0.1), et d'utiliser ensuite le risque de deuxième espèce ou la puissance pour évaluer la qualité du test.

2.3.2 La construction en pratique

La construction pratique d'un test peut se faire selon le schéma suivant :

1. Détermination d'un modèle statistique représentant le phénomène aléatoire étudié.
2. Détermination des hypothèses (H_0) et (H_1) à partir du problème posé, en respectant la dissymétrie des rôles des deux hypothèses.
3. Détermination d'une statistique de test $T : \mathcal{X} \to \mathbb{R}$ comme outil décisionnel, respectant les trois règles suivantes : la valeur de $T(x)$ doit être calculable quelle que soit la valeur de θ (elle ne doit donc pas dépendre de θ), la loi de $T(X)$ sous l'hypothèse (H_0) (ou à sa "frontière") doit être entièrement connue, libre du paramètre inconnu, tabulée si possible, et sa loi sous (H_1) doit différer de celle sous (H_0). Cette statistique de test est souvent construite à partir d'un estimateur de θ, ou d'un estimateur d'une distance choisie entre θ et Θ_0.
4. Détermination intuitive de la forme de la région critique et de la fonction de test, sur la base des seules connaissances en estimation généralement. Par exemple, $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s\}$, $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s\}$, $\mathcal{R}_{(H_0)} = \{x, |T(x)| \geq s\}$, $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s_2 \text{ ou } T(x) \leq s_1\}$, $\mathcal{R}_{(H_0)} = \{x, T(x) \in [s_1, s_2]\}$ $(s_1 < s_2)$...
5. Détermination précise des valeurs critiques s, s_1, s_2, de telle façon que le test soit bien de niveau α_0 fixé au préalable (c.f. inéquation du niveau), de taille la plus proche possible de α_0 (pour ne pas trop perdre en puissance).
6. Conclusion au vu de l'observation x.
7. Évaluation de la puissance du test, ou tracé de la fonction puissance éventuellement.

2.4 Intervalles de confiance et problèmes de test bilatères

Définition 15. *Étant donné* $\alpha_0 \in]0,1[$, *un* intervalle de confiance de niveau de confiance $1-\alpha_0$ *pour* $\tau(\theta)$ *est défini par deux statistiques A et B telles que pour tout* $\theta \in \Theta$,

$$P_\theta\left(\{x, \tau(\theta) \in [A(x), B(x)]\}\right) \geq 1-\alpha_0.$$

On considère le problème de test de l'hypothèse $(H_0) : \tau(\theta) = \tau_0$ contre $(H_1) : \tau(\theta) \neq \tau_0$.
Si A et B définissent un intervalle de confiance de niveau de confiance $1-\alpha_0$ pour $\tau(\theta)$, alors

$$\sup_{\theta, \tau(\theta)=\tau_0} P_\theta\left(\{x, \tau(\theta) \in [A(x), B(x)]\}\right) \geq 1-\alpha_0,$$

d'où

$$\sup_{\theta, \tau(\theta)=\tau_0} P_\theta\left(\{x, \tau_0 \notin [A(x), B(x)]\}\right) \leq \alpha_0.$$

Par conséquent, le test ϕ de région critique

$$\mathcal{R}_{(H_0)} = \{x, \tau_0 \notin [A(x), B(x)]\}$$

est un test de niveau α_0 de (H_0) contre (H_1). La conclusion de ce test s'exprime donc sous la forme simple suivante : si $\tau_0 \in [A(x), B(x)]$, alors on accepte l'hypothèse (H_0), sinon, on rejette (H_0) au profit de (H_1) pour un niveau α_0.

Fonction pivotale et statistique de test. La construction d'un intervalle de confiance est basée sur une fonction pivotale, c'est-à-dire une fonction $T : \mathcal{X} \times \tau(\Theta)$ telle que pour tout $\theta \in \Theta$, $T(X, \tau(\theta))$ suit une loi entièrement spécifiée. La statistique $T(x, \tau_0)$ peut alors être utilisée comme statistique de test pour le problème de test considéré.

Chapitre 3

Tests paramétriques de base pas si basiques

3.1 Tests en modèles gaussiens

3.1.1 Tests sur l'espérance

Soit $X = (X_1, \dots, X_n)$ un n-échantillon d'une loi gaussienne $\mathcal{N}(m, \sigma^2)$.

Variance connue

On suppose dans un premier temps que σ^2 est connue.

Si l'on souhaite tester $(H_0) : m = m_0$ contre $(H_1) : m = m_1$ avec $m_1 > m_0$, ou $(H_1) : m > m_0$, sur la base d'une observation x de X, on peut suivre le schéma de construction suivant.

- Modèle statistique : $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$, avec $\mathcal{X} = \mathbb{R}^n$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$, $\Theta \subset \mathbb{R}$ et pour $\theta = m \in \Theta$, $P_\theta = \mathcal{N}(\theta, \sigma^2)^{\otimes n}$.
- Statistique de test : $T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sigma}$.
- Fonction de test : $\phi(x) = \mathbb{1}_{T(x)\geq s}$.
- Calcul de s pour un test de niveau $\alpha_0 = 0.05$: s doit vérifier $P_{m_0}(\{x, T(x) \geq s\}) \leq 0.05$ c'est-à-dire $P_{m_0}\left(\left\{x, \sqrt{n}\frac{\bar{x}-m_0}{\sigma} \geq s\right\}\right) = P(N \geq s) \leq 0.05$, où $N \sim \mathcal{N}(0, 1)$. La valeur $s = 1.645$ convient donc.

Si l'on veut tester $(H_0) : m \leq m_0$ contre $(H_1) : m > m_0$, le schéma reste le même puisque par croissance et continuité de la fonction $m \mapsto P_m(\{x, T(x) \geq s\})$ sur $]-\infty, m_0]$,

$$\sup_{m\leq m_0} P_m(\{x, T(x) \geq s\}) = P_{m_0}(\{x, T(x) \geq s\}).$$

Donc choisir s de telle façon que le test soit de niveau 0.05 revient à choisir s tel que $P_{m_0}(\{x, T(x) \geq s\}) \leq 0.05$.

Si l'on veut tester $(H_0) : m = m_0$ contre $(H_1) : m = m_1$ avec $m_1 < m_0$, ou $(H_1) : m > m_0$, ou encore $(H_0) : m \geq m_0$ contre $(H_1) : m < m_0$, le test se construit à partir de la même statistique de test, mais sa région critique diffère.
On peut suivre le schéma suivant :

- Statistique de test : $T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sigma}$.
- Fonction de test : $\phi(x) = \mathbb{1}_{T(x)\leq s}$.

— Calcul de s pour un test de niveau $\alpha_0 = 0.05$: s doit vérifier $\sup_{m \geq m_0} P_m(\{x, T(x) \leq s\}) = P_{m_0}(\{x, T(x) \leq s\}) \leq 0.05$ c'est-à-dire $P_{m_0}\left(\left\{x, \sqrt{n}\frac{\bar{x}-m_0}{\sigma} \leq s\right\}\right) = P(N \leq s) \leq 0.05$, où $N \sim \mathcal{N}(0,1)$. On prend donc $s = -1.645$.

Si l'on veut tester $(H_0) : m = m_0$ contre $(H_1) : m \neq m_0$, on peut suivre le schéma suivant :
— Statistique de test : $T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sigma}$.
— Fonction de test : $\phi(x) = \mathbb{1}_{|T(x)| \geq s}$.
— Calcul de s pour un test de niveau $\alpha_0 = 0.05$: s doit vérifier $P_{m_0}(\{x, |T(x)| \geq s\}) \leq 0.05$ c'est-à-dire $P(|N| \geq s) \leq 0.05$, où $N \sim \mathcal{N}(0,1)$. On prend donc $s = 1.96$.

Remarque.
Les trois tests précédents peuvent alors s'exprimer sous la forme $\phi(x) = \mathbb{1}_{T(x) \geq s}$, $\phi(x) = \mathbb{1}_{T(x) \leq s}$ ou $\phi(x) = \mathbb{1}_{|T(x)| \geq s}$ avec des constantes s différentes, mais avec la même statistique de test $T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sigma}$, où T a comme propriété fondamentale que la loi de $T(X)$ est entièrement connue lorsque $X \sim P_{m_0}$ (elle ne dépend d'aucun paramètre inconnu).

Variance inconnue

On suppose maintenant que σ^2 est inconnue.

On souhaite tester $(H_0) : m = m_0$ contre $(H_1) : m = m_1$ avec $m_1 \neq m_0$, sur la base d'une observation x de X, ou $(H_0) : m \leq m_0$ contre $(H_1) : m > m_0$, ou $(H_0) : m \geq m_0$ contre $(H_1) : m < m_0$, ou encore $(H_0) : m = m_0$ contre $(H_1) : m \neq m_0$.
On note $P_{(m,\sigma^2)}$ la loi $\mathcal{N}(m, \sigma^2)^{\otimes n}$.
Comme pour les tests précédents, on construit un test basé sur une statistique T telle que la loi de $T(X)$ est entièrement connue lorsque $X \sim P_{(m_0,\sigma^2)}$, c'est-à-dire dont la loi sous (H_0) ne dépend pas du paramètre inconnu σ^2.
Cette statistique est obtenue en remplaçant σ par son estimateur empirique classique $S(x)$, avec $S^2(x) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$: $T(x) = \sqrt{n}\frac{\bar{x}-m_0}{S(x)}$.
On sait alors que si $X \sim P_{(m_0,\sigma^2)}$, $T(X)$ suit une loi de Student de paramètre $(n-1)$, quelle que soit la valeur de σ^2, et on peut ainsi déterminer les valeurs critiques du test pour qu'il soit de niveau α_0 fixé a priori.

3.1.2 Tests sur la variance

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon d'une loi gaussienne $\mathcal{N}(m, \sigma^2)$.

On souhaite tester $(H_0) : \sigma^2 = \sigma_0^2$ contre $(H_1) : \sigma^2 = \sigma_1^2$ avec $\sigma_1^2 \neq \sigma_0^2$, sur la base d'une observation x de X, ou $(H_0) : \sigma^2 \leq \sigma_0^2$ contre $(H_1) : \sigma^2 > \sigma_0^2$, ou $(H_0) : \sigma^2 \geq \sigma_0^2$ contre $(H_1) : \sigma^2 < \sigma_0^2$, ou encore $(H_0) : \sigma^2 = \sigma_0^2$ contre $(H_1) : \sigma^2 \neq \sigma_0^2$.

Espérance connue

On suppose que m est connue. On note P_{σ^2} la loi $\mathcal{N}(m, \sigma^2)^{\otimes n}$.
On considère la statistique de test $T(x) = \frac{1}{\sigma_0^2}\sum_{i=1}^n (x_i - m)^2$. On sait en effet que si $X \sim P_{\sigma_0^2}$, $T(X) \sim \chi^2(n)$.

Espérance inconnue

On suppose maintenant que m est inconnue. On note $P_{(m,\sigma^2)}$ la loi $\mathcal{N}(m,\sigma^2)^{\bigotimes n}$.
On considère la statistique de test $T(x) = \frac{1}{\sigma_0^2}\sum_{i=1}^n (x_i - \bar{x})^2$. On sait en effet que si $X \sim P_{(m,\sigma_0^2)}$, $T(X) \sim \chi^2(n-1)$ quelle que soit la valeur de m.

3.1.3 Tests de comparaison d'espérances

Soit $Y = (Y_1, \dots, Y_{n_1})$ un n_1-échantillon de la loi $\mathcal{N}(m_1, \sigma_1^2)$ et $Z = (Z_1, \dots, Z_{n_2})$ un n_2-échantillon de la loi $\mathcal{N}(m_2, \sigma_2^2)$. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$.

On veut tester $(H_0) : m_1 = m_2$ contre $(H_1) : m_1 \neq m_2$, ou $m_1 < m_2$, ou $m_1 > m_2$.

Variances connues

On suppose que σ_1^2 et σ_2^2 sont connues, et on note $P_{(m_1,m_2)}$ la loi de X.
On considère la statistique de test $T(x) = \frac{\bar{y}-\bar{z}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$.
Lorsque $m_1 = m_2 = m$, i.e. lorsque $X \sim P_{(m,m)}$, $T(X) \sim \mathcal{N}(0,1)$.

Variances inconnues mais supposées égales

On suppose maintenant que σ_1^2 et σ_2^2 sont inconnues, mais que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. On note $P_{(m_1,m_2,\sigma^2)}$ la loi de X.
La statistique de test devient alors $T(x) = \frac{\bar{y}-\bar{z}}{S(y,z)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, où $S^2(y,z) = \frac{\sum_{i=1}^{n_1}(y_i-\bar{y})^2+\sum_{i=1}^{n_2}(z_i-\bar{z})^2}{n_1+n_2-2}$.
Lorsque $m_1 = m_2 = m$, i.e. lorsque $X \sim P_{(m,m,\sigma^2)}$, $T(X) \sim \mathcal{T}(n_1+n_2-2)$ quelle que soit la valeur de σ^2.

Variances inconnues (problème de Behrens-Fisher)

Si σ_1^2 et σ_2^2 sont inconnues, et non supposées égales, on peut choisir comme statistique de test $T(x) = \frac{\bar{y}-\bar{z}}{\sqrt{\frac{S^2(y)}{n_1}+\frac{S^2(z)}{n_2}}}$, avec $S^2(y) = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i-\bar{y})^2$ et $S^2(z) = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(z_i-\bar{z})^2$.
Lorsque $m_1 = m_2 = m$, $T(X)$ converge en loi vers une loi $\mathcal{N}(0,1)$ quelles que soient les valeurs de σ_1^2 et σ_2^2, donc on peut construire des tests dont le niveau est garanti asymptotiquement (ou tests asymptotiques), à condition que n_1 et n_2 soient suffisamment grands.
Si n_1 et n_2 ne sont pas suffisamment grands, on considère en général l'approximation de Welch-Satterthwaite-Aspin correspondant à une approximation de la loi sous (H_0) par une loi Student à ν degrés de liberté où ν est l'entier le plus proche de

$$\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{\sigma_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{\sigma_2^2}{n_2}\right)^2}.$$

Comme σ_1^2 et σ_2^2 sont inconnues, on les estime, donc on considère l'entier $\hat{\nu}$ le plus proche de :

$$\frac{\left(\frac{S_1^2(Y)}{n_1}+\frac{S_2^2(Z)}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2(Y)}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{S_2^2(Z)}{n_2}\right)^2}.$$

Attention :

1. Tous les tests vus dans cette section ne sont valables que si les deux échantillons sont indépendants. Si ce n'est pas le cas et si $n_1 = n_2 = n$, on parle de données appariées. On peut dans ce cas, considérer la différence $X_i = Y_i - Z_i$ et supposer, éventuellement, que l'échantillon $X = (X_1, \dots, X_n)$ est gaussien. On réalise alors un test de nullité de l'espérance.
2. Ces tests sont souvent peu robustes au changement de loi. On conseille, si l'on n'a aucune garantie que les lois peuvent être supposées gaussiennes, de considérer un test non-paramétrique de comparaison de lois.

Remarque : On peut aussi effectuer au préalable un test de comparaison de variances (c.f. section suivante). Si ce test conduit à supposer que les variances sont égales, on peut ensuite mettre en œuvre le test de comparaison d'espérances correspondant. Mais attention, le risque de première espèce global s'en trouve modifié. On a pour coutume de prendre pour les deux tests un niveau $\alpha/2$.

3.1.4 Tests de comparaison de variances

Soit $Y = (Y_1, \dots, Y_{n_1})$ un n_1-échantillon de la loi $\mathcal{N}(m_1, \sigma_1^2)$ et $Z = (Z_1, \dots, Z_{n_2})$ un n_2-échantillon de la loi $\mathcal{N}(m_2, \sigma_2^2)$. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$.

On veut tester $(H_0) : \sigma_1^2 = \sigma_2^2$ contre $(H_1) : \sigma_2^2 \neq \sigma_1^2$, ou $\sigma_1^2 > \sigma_2^2$, ou $\sigma_1^2 < \sigma_2^2$.

Espérances connues

On suppose que m_1 et m_2 sont connues, et on note $P_{(\sigma_1^2,\sigma_2^2)}$ la loi de X.

On considère la statistique de test $T(x) = \frac{\tilde{S}^2(y)}{\tilde{S}^2(z)}$, avec $\tilde{S}^2(y) = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - m_1)^2$ et $\tilde{S}^2(z) = \frac{1}{n_2}\sum_{i=1}^{n_2}(z_i - m_2)^2$. Lorsque $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e. lorsque $X \sim P_{(\sigma^2,\sigma^2)}$, $T(X) \sim \mathcal{F}(n_1, n_2)$.

Espérances inconnues

On suppose maintenant que m_1 et m_2 sont inconnues, et on note $P_{(m_1,\sigma_1^2,m_2,\sigma_2^2)}$ la loi de X.

On considère la statistique de test $T(x) = \frac{S^2(y)}{S^2(z)}$.
Lorsque $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e. lorsque $X \sim P_{(m_1,\sigma^2,m_2,\sigma^2)}$, $T(X) \sim \mathcal{F}(n_1 - 1, n_2 - 1)$ quelles que soient les valeurs de m_1 et m_2.

Remarque (utile pour la lecture de certaines tables statistiques). Si $F \sim \mathcal{F}(q_1, q_2)$, alors $1/F \sim \mathcal{F}(q_2, q_1)$.

3.2 Tests en modèles de Bernoulli

3.2.1 Tests sur une probabilité

Soit $X = (X_1, \dots, X_n)$ un n-échantillon d'une loi de Bernoulli $\mathcal{B}(p)$ de paramètre p (p est la probabilité de réalisation d'un événement). On note P_p la loi $\mathcal{B}(p)^{\otimes n}$.
On veut tester $(H_0) : p = p_0$ contre $(H_1) : p = p_1$ (avec $p_1 \neq p_0$ donné), ou $(H_0) : p \leq p_0$ contre $(H_1) : p > p_0$, ou $(H_0) : p \geq p_0$ contre $(H_1) : p < p_0$, ou encore $(H_0) : p = p_0$ contre $(H_1) : p \neq p_0$.
On considère la statistique de test $T(x) = \sum_{i=1}^{n} x_i$. Alors, si $X \sim P_{p_0}$, $T(X)$ suit une loi binomiale de paramètres (n, p_0).

Si n est suffisamment grand, il est courant d'utiliser une approximation de la loi binomiale $\mathcal{B}(n, p_0)$ par une loi de Poisson $\mathcal{P}(np_0)$ ou par une loi gaussienne $\mathcal{N}(np_0, np_0(1-p_0))$. Dans ce cas, le test ne sera pas nécessairement précisément de niveau α_0.
Les critères les plus souvent utilisés pour justifier l'approximation par une loi de Poisson sont $n > 30$ et $np_0 < 5$ ou $n > 50$ et $p_0 < 0.10$.
Celui le plus utilisé pour justifier l'approximation par une loi gaussienne est $n > 30$, $np_0 \geq 5$ et $n(1-p_0) \geq 5$.

3.2.2 Tests de comparaison de probabilités

Soit $Y = (Y_1, \dots, Y_{n_1})$ un n_1-échantillon d'une loi de Bernoulli $\mathcal{B}(p_1)$ et $Z = (Z_1, \dots, Z_{n_2})$ un n_2-échantillon d'une loi de Bernoulli $\mathcal{B}(p_2)$. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$. On note $P_{(p_1,p_2)}$ la loi de X.

On veut tester $(H_0) : p_1 = p_2$ contre $(H_1) : p_1 \neq p_2$, ou $(H_1) : p_1 > p_2$, ou $(H_1) : p_1 < p_2$.
On considère la statistique de test

$$T(x) = \frac{\bar{y} - \bar{z}}{\sqrt{\frac{n_1\bar{y}+n_2\bar{z}}{n_1+n_2}\left(1 - \frac{n_1\bar{y}+n_2\bar{z}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Si $p_1 = p_2 = p$, i.e. si $X \sim P_{(p,p)}$, $T(X)$ converge en loi vers une loi $\mathcal{N}(0,1)$, donc on peut construire des tests asymptotiques à condition que n_1 et n_2 soient suffisamment grands.

Remarque. Si l'on souhaite construire un test non asymptotique, on peut utiliser un test non paramétrique comme par exemple le test exact de Fisher ou le test de Barnard.

3.3 Test du rapport de vraisemblance maximale

L'emploi de la vraisemblance très présent en estimation paramétrique permet de construire un test, dit du rapport de vraisemblance maximale, de façon intuitive.
Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique dominé par une mesure μ, de vraisemblance notée $L(x, \theta)$. Soit $\Theta_0 \subset \Theta$ et $\Theta_1 \subset \Theta$, avec $\Theta_0 \cap \Theta_1 = \emptyset$. On considère le problème de test de $(H_0) : \theta \in \Theta_0$ contre $(H_1) : \theta \in \Theta_1$. On introduit alors le rapport

$$\frac{\sup_{\theta\in\Theta_0} L(x, \theta)}{\sup_{\theta\in\Theta_1} L(x, \theta)},$$

et on décide de rejeter (H_0) lorsque ce rapport est petit.

En fait on montre que ce test est équivalent au test du rapport de vraisemblance maximale défini de la façon suivante.

Définition 16. *Soit*

$$\lambda(x) = \frac{\sup_{\theta\in\Theta_0} L(x,\theta)}{\sup_{\theta\in\Theta} L(x,\theta)}.$$

Un test du rapport de vraisemblance maximale *est de la forme* $\phi(x) = \mathbb{1}_{\lambda(x)\leq s}$.

Remarques.

— La statistique du rapport de vraisemblance maximale λ est liée à l'estimation par le maximum de vraisemblance puisque si $\hat{\theta}$ est un estimateur du maximum de vraisemblance de θ, et si $\hat{\theta}_0$ est un estimateur du maximum de vraisemblance de θ pour l'espace des paramètres réduit à Θ_0 (on parle d'EMV restreint), $\lambda(x) = L(x,\hat{\theta}_0)/L(x,\hat{\theta})$.

— Même si dans les cas simples, le test du rapport de vraisemblance maximale est équivalent aux tests optimaux vus dans le chapitre suivant, ce test n'est dans le cas général pas forcément optimal !

Le calcul de λ se trouve souvent simplifié lorsque l'on dispose d'une statistique exhaustive T pour le modèle considéré.

Thèorème 1. *Soit* $(\mathcal{X},\mathcal{A},(P_\theta)_{\theta\in\Theta})$ *un modèle statistique dominé par une mesure* μ*, de vraisemblance notée* $L(x,\theta)$*. Si* T *est une statistique exhaustive pour* θ*, alors, pour toute partie* Θ_0 *de* Θ*, la statistique du rapport de vraisemblance maximale* λ *factorise à travers* T*, i.e. il existe une fonction* $\tilde{\lambda}$ *telle que pour tout* x *de* $\mathcal{X}$*,* $\lambda(x) = \tilde{\lambda}(T(x))$*.*

Preuve. Le théorème de factorisation assure l'existence de fonctions g et h telles que

$$\forall x \in \mathcal{X},\ \forall\theta\in\Theta,\ L(x,\theta) = g(T(x),\theta)h(x).$$

On a donc pour tout x de $\mathcal{X}$,

$$\lambda(x) = \frac{\sup_{\theta\in\Theta_0} L(x,\theta)}{\sup_{\theta\in\Theta} L(x,\theta)} = \frac{\sup_{\theta\in\Theta_0} g(T(x),\theta)h(x)}{\sup_{\theta\in\Theta} g(T(x),\theta)h(x)}.$$

Remarque importante : On a le résultat asymptotique suivant. Si $\Theta = \mathbb{R}^d$ et Θ_0 est un sous-espace vectoriel de dimension k, sous certaines conditions sur la loi de l'estimateur du maximum de vraisemblance de θ, alors $-2\ln\lambda(X)$ converge en loi sous (H_0) vers une loi $\chi^2(d-k)$. On trouvera les détails de la preuve et des conditions sur la loi dans l'ouvrage de Wilks, *Mathematical statistics*, ou son article de 1938 aux Annals of Mathematical Statistics.

3.4 Exercices

Exercice 1 : Test sur l'espérance d'une loi gaussienne

Un négociant en vin s'intéresse à la contenance des bouteilles d'un producteur soupçonné par certains clients de frauder. Il souhaite s'assurer que cette contenance respecte bien en moyenne la limite légale de 75 cl. À cet effet, il mesure le contenu de 10 bouteilles prises au hasard et obtient les valeurs suivantes (en cl) :

$$73.2,\ 72.6,\ 74.5,\ 75,\ 75.5,\ 73.7,\ 74.1,\ 75.8,\ 74.8,\ 75.$$

1. On suppose que la contenance des bouteilles (en cl) suit une loi gaussienne d'espérance θ inconnue, d'écart-type connu égal à 1.
a) Écrire le modèle statistique considéré.
b) Le négociant décide de tester l'hypothèse nulle $(H_0)\ :\ \theta = 75$ contre l'alternative $(H_1)\ :\ \theta < 75$. Quel point de vue le négociant adopte-t-il en choisissant ces hypothèses ? Justifier précisément la réponse.
c) Construire, à l'aide d'une règle de décision intuitive basée sur la moyenne empirique, un test de niveau 1% de (H_0) contre (H_1). Quelle est la conclusion de ce test ?
d) Tracer la courbe de puissance du test.
e) Le négociant veut pouvoir détecter, avec une probabilité élevée (99%), une contenance moyenne de 74.8 cl tout en gardant un test de niveau 1%. Que doit-il faire ?
f) Quel test le négociant peut-il considérer s'il souhaite en fait tester l'hypothèse nulle multiple $(H_0)\ :\ \theta \geq 75$ contre l'alternative $(H_1)\ :\ \theta < 75$?
2. On suppose maintenant que la contenance des bouteilles suit une loi gaussienne d'écart-type inconnu.
a) Écrire le modèle statistique considéré.
b) Le négociant adopte toujours le même point de vue. Construire un nouveau test de niveau 5% lui permettant de s'assurer que le producteur ne fraude pas.
c) Peut-on tracer la courbe de puissance du test ?

Exercice 2 : Test sur la variance d'une loi gaussienne

Sur le marché des Lices à Rennes, un inspecteur des poids et mesures vérifie la précision de la balance d'un vendeur de fruits et légumes. Il effectue pour cela 10 pesées d'un poids étalonné à 100g et note les indications de la balance : $x_1, \ldots, x_{10}$. On admet que le résultat d'une pesée est une variable aléatoire de loi gaussienne d'espérance 100, de variance σ^2 avec $\sigma = 5$ si la balance est juste, et $\sigma > 5$ si elle est déréglée. Les mesures ont donné $\sum_{i=1}^{10}(x_i - 100)^2 = 512$.
1. Écrire le modèle statistique considéré et les hypothèses à tester. Justifier ce choix.
2. Construire un test basé sur un estimateur de la variance permettant de vérifier la précision de la balance.
3. Quelle sera la conclusion de l'inspecteur avec ce test pour un niveau 5% ? Pour un niveau 1% ?
4. Calculer la p valeur du test.
5. Peut-on construire un test similaire lorsque l'espérance de la loi gaussienne n'est plus connue ?
6. Et lorsque la loi elle-même est inconnue ?

Exercice 3 : Tests sur une probabilité

Avant le second tour d'une élection présidentielle, un candidat commande un sondage à une société spécialisée pour savoir s'il a une chance d'être élu.
1. Si p est la proportion d'électeurs qui lui est favorable dans la population, il souhaite tout d'abord résoudre le problème de test :

$$(H_0) \ : \quad p = 0.48 \quad \text{contre} \quad (H_1) \ : \quad p = 0.52.$$

a) Écrire le modèle statistique considéré.
b) Quelle est la signification du choix de $p = 0.48$ comme hypothèse nulle ?
c) Quelle statistique de test pourra-t-on considérer ?
d) Construire un test de niveau exact 10%, puis de niveau asymptotique 10% lorsque le sondage est effectué auprès de $n = 100$ électeurs.
e) Combien d'électeurs devra-t-on interroger si l'on souhaite avoir un niveau asymptotique α et un risque de deuxième espèce asymptotique inférieur à β, avec α et β donnés ?
2. Le candidat souhaite maintenant tester les hypothèses composites :

$$(H_0) \ : \quad p \leq 0.5 \quad \text{contre} \quad (H_1) \ : \quad p > 0.5.$$

Que peut-on conclure ?

Exercice 4 : Test sur le paramètre d'une loi de Poisson

On admet, s'inspirant de l'argument historique de Siméon Denis Poisson (1837), que le nombre mensuel d'erreurs judiciaires commises dans une juridiction administrée par un juge soupçonné de ne pas avoir toute sa tête, peut être modélisé par une variable aléatoire suivant une loi de Poisson de paramètre $\lambda = 2$. après avoir écarté ce juge, on s'attend à ce que le nombre mensuel moyen d'erreurs judiciaires commises dans la juridiction diminue.
1. Construire un test, basé sur le nombre total d'erreurs judiciaires sur six mois suivant l'éviction du juge, permettant de le vérifier.
2. Que décide-t-on pour un niveau 10% si le nombre total d'erreurs judiciaires sur six mois suivant l'éviction du juge est de 10 ?
3. Tracer la courbe de puissance du test.

Exercice 5 : Test sur le paramètre d'une loi uniforme

Un programme informatique de simulation de la loi uniforme sur $[0, \theta]$ destiné à un nouveau logiciel statistique a généré les nombres suivants : 95, 24, 83, 52, 68.
1. Donner l'estimateur du maximum de vraisemblance de θ et déterminer sa loi.
2. En déduire un test de niveau 5% de l'hypothèse $(H_0) : \theta = 100$ contre $(H_1) : \theta > 100$.
3. Que peut-on conclure ?

Exercice 6 : Test sur le paramètre d'une loi exponentielle

On admet que la durée de vie de la batterie d'un smartphone (d'une célèbre marque symbolisée par un fruit) est modélisée par une variable aléatoire X de loi exponentielle de paramètre $1/\theta$, avec $\theta > 0$ inconnu. On considère un n-échantillon $(X_1, \ldots, X_n)$ de la loi de X, dont on a une observation $(x_1, \ldots, x_n)$.

1. Déterminer l'estimateur $\hat{\theta}_n$ du maximum de vraisemblance du paramètre θ.
2. On rappelle que la densité d'une loi du Khi-Deux à $2k$ degrés de liberté est donnée par

$$g_k(x) = \frac{1}{2^k(k-1)!} x^{(k-1)} e^{-x/2} \mathbb{1}_{x>0}.$$

Montrer que la variable $\frac{2X}{\theta}$ suit une loi du Khi-Deux à 2 degrés de liberté. En déduire que $\frac{2}{\theta}\sum_{i=1}^{n} X_i$ suit une loi du Khi-Deux à $2n$ degrés de liberté.
3. Des études passées avaient permis d'attribuer au paramètre θ la valeur θ_0. L'évolution des méthodes de fabrication pouvant avoir entraîné une augmentation de θ, on considère le problème de test de l'hypothèse

$$(H_0) \ : \ \theta = \theta_0 \quad \text{contre} \quad (H_1) \ : \ \theta > \theta_0.$$

Construire un test de niveau α de ces hypothèses.
4. On a relevé les durées de vie suivantes en années :
0.60, 0.19, 6.44, 1.74, 0.02, 2.34, 4.08, 0.17, 1.16, 1.23, 1.74, 0.55, 0.25, 3.43, 1.64, 5.32, 3.80, 1.27, 0.68, 2.22, 2.77, 2.85, 2.85, 3.67, 0.26, 2.48, 4.08, 1.23, 0.35, 5.05, 3.22.
Quelle est la conclusion du test pour un niveau 5% et $\theta_0 = 2$ (durée de la garantie) ?
5. On suppose maintenant que l'on n'observe pas les durées de vie X_i directement mais les variables $Y_i = \mathbb{1}_{X_i \geq 2}$ pour $i = 1 \dots n$. Proposer un nouveau test de (H_0) contre (H_1).

Exercice 7 : Comparaison de deux lois gaussiennes

On admet que les notes des étudiants de Licence MASS en Tests statistiques à l'Université Rennes 2 dans un contexte "normal" suit une loi gaussienne. On a relevé 15 notes de ces étudiants en 2005.
Après le mouvement social s'opposant au Contrat première Embauche en 2006, on a également relevé 15 notes en Tests statistiques.

Notes en 2005	12.8	15	8.5	12.7	10.4	15.5	9.6	10.3
	8.5	8.1	7.8	14	12.5	8.6	7	
Notes en 2006	10.1	8.9	6.1	4.8	9.1	11.9	14.2	13.5
	16	12.9	11.1	11	8.8	10	9.2	

On souhaite savoir si le mouvement social a eu un effet sur les résultats des étudiants.

1. Quel modèle statistique peut-on poser ?
2. Quels tests statistiques peut-on faire dans ce modèle ?
3. Quelles seront les conclusions des tests ?

Exercice 8 : Comparaison de deux probabilités

En juillet 2010, un sondage Ifop sur un échantillon dit représentatif de 958 personnes donnait le résultat suivant : 8 sympathisants PS sur 100 contre 6 sympathisants UMP sur 100 sont tatoués. Imaginant que sur les 958 personnes interrogées, 249 se sont déclarées sympathisantes PS, dont 20 tatouées, et 297 sympathisantes UMP, dont 18 tatouées, que peut-on penser de la déclaration faite dans les journaux : "les sympathisants PS sont plus tatoués que les sympathisants UMP" ?

Exercice 9

Sur la base d'un échantillon de taille n de la densité $f : x \mapsto \frac{1}{\sigma} e^{-(x-\theta)/\sigma} 1_{[\theta,+\infty[}(x)$, où $\theta \in \mathbb{R}$ et $\sigma > 0$ sont inconnus, on désire tester l'hypothèse $(H_0) : \ \theta \leq \theta_0$ contre l'alternative $(H_1) : \ \theta > \theta_0$. Déterminer la forme d'un test de rapport de vraisemblance maximale.

Exercice 10

Le revenu annuel des individus d'une population est distribué selon une loi de Pareto (généralisant la loi connue sous le nom de loi des 80/20, établie par Vilfredo Pareto, qui avait noté que 80% des revenus d'une population sont détenus par seulement 20% de la population), de densité :

$$f : x \mapsto \frac{ak^a}{x^{a+1}} 1_{[k,+\infty[}(x).$$

Les paramètres k (revenu minimum) et a (paramètre de forme ou constante de Pareto) sont inconnus.
1. Estimer les paramètres par la méthode du maximum de vraisemblance.
2. On désire tester l'hypothèse $(H_0) : \ a = 1$ contre l'alternative $(H_1) : \ a \neq 1$. Déterminer la forme d'un test de rapport de vraisemblance maximale.

Exercice 11

Démontrer la Proposition 1.

3.5 Problèmes corrigés

3.5.1 Faire le ménage ou l'amour, faut-il choisir ?

Problème

Les chiffres donnés ici sont fictifs, bien qu'inspirés d'une étude réelle !

Selon une dépêche de l'AFP du 30 janvier 2013, "Plus un homme marié accorde de temps aux tâches ménagères, moins il a de relations sexuelles"... Les titres de certains journaux ne se sont pas fait attendre, du "Faire le ménage ou l'amour, il faut choisir" du figaro.fr au "Insolite : un homme qui passe l'aspirateur, c'est moins excitant ?" de lexpress.fr.
L'AFP dit s'appuyer sur une étude publiée dans *American Sociological Review 78 (1)* par Sabino Kornrich, Julie Brines et Katrina Leupp. En réalité, les auteurs de cette étude se penchent sur l'idée souvent avancée par les médias que les hommes mariés participant plus au travaux domestiques ont une fréquence de rapports sexuels plus élevée : les femmes échangeraient des rapports sexuels contre la participation des hommes aux travaux domestiques. Les auteurs cherchent à démontrer que, dans le mariage, la fréquence des relations sexuelles comme l'accomplissement des tâches domestiques, loin d'être régis selon des principes d'échange de services, obéissent en fait à des schémas liés au genre. Ils étudient ainsi la fréquence des relations sexuelles dans le mariage en fonction de la participation des hommes aux travaux domestiques, en différenciant les travaux dits traditionnellement féminins (courses, ménage, cuisine...) des travaux dits traditionnellement masculins (entretien de la voiture, jardinage, paiement des factures...).

Nous décidons de nous intéresser à la même question et de mener notre propre étude, selon deux angles de vue différents.

Partie I : un premier angle de vue

Pour les hommes mariés participant à moins de 30% aux travaux domestiques traditionnellement féminins, la probabilité d'avoir au moins 4 rapports sexuels par mois s'élève à 0.6.
Sur 900 hommes mariés interrogés participant aux travaux domestiques traditionnellement féminins à plus de 30%, 513 déclarent avoir au moins 4 rapports sexuels par mois.
On souhaite savoir si la probabilité θ pour un homme accomplissant plus de 30% des travaux domestiques traditionnellement féminins d'avoir une fréquence sexuelle mensuelle au moins égale à 4 est inférieure à 0.6 ou non.

1. Décrire la variable aléatoire X modélisant les données et le modèle statistique considéré.

.....
.....
.....

2. Précisez votre choix d'hypothèses nulle (H_0) et alternative (H_1), en le justifiant.

.....
.....

3. Déterminer une statistique de test $T(X)$, et donner sa loi lorsque $\theta = 0.6$.

. . . . . .
. . . . . .

4. Choisir la forme de la région critique du test :

- ☐ $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, |T(x)| \geq s\}$

5. On veut construire un test de niveau $\alpha = 5\%$. Déterminer précisément la valeur de s à l'aide des tables statistiques fournies.

. . . . . .
. . . . . .
. . . . . .
. . . . . .

6. Quelle est la conclusion du test ?

. . . . . .

Partie II : un deuxième angle de vue

On admet que le pourcentage de participation aux travaux domestiques traditionnellement féminins d'un homme ayant une fréquence sexuelle mensuelle inférieure à 4 suit une loi gaussienne d'espérance 30.
On relève les pourcentages de participation aux travaux domestiques traditionnellement féminins de 100 hommes déclarant avoir au moins 4 rapports sexuels par mois, et on note $(x_i)_{i=1...100}$ ces pourcentages. On calcule la moyenne et la variance empirique des données, et on obtient : $\bar{x} = \sum_{i=1}^{100} x_i/100 = 25.99$ et $\sum_{i=1}^{100}(x_i - \bar{x})^2 = 41435.35$.
On suppose que le pourcentage de participation aux travaux domestiques traditionnellement féminins d'un homme ayant au moins 4 rapports sexuels par mois suit une loi gaussienne d'espérance m et de variance σ^2 inconnues.

1. Décrire la variable aléatoire X modélisant les données et le modèle statistique considéré.

. . . . . .
. . . . . .
. . . . . .

2. On souhaite tester l'hypothèse nulle $(H_0) : m = 30$ contre l'alternative $(H_1) : m < 30$.
De quel risque se prémunit-on en priorité en considérant ces hypothèses ?

. . . . . .
. . . . . .

3. Déterminer une statistique de test $T(X)$, et donner sa loi sous (H_0).

. . . . . .
. . . . . .

4. Choisir la forme de la région critique du test :

- ☐ $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, |T(x)| \geq s\}$

5. On veut construire un test de niveau $\alpha = 5\%$. Représenter sur un graphe commenté la valeur de s qu'il faut prendre pour que le test soit bien du niveau choisi.

Déterminer précisément la valeur de s à l'aide des tables statistiques fournies.

. . . . . .
. . . . . .

6. Quelle est la conclusion du test ?

. . . . . .

7. Donner un encadrement de la p-valeur du test. Que peut-on en déduire ?

. . . . . .
. . . . . .

Partie III

Après avoir mené la présente étude, quelle(s) critique(s) pouvez-vous faire à propos de la conclusion avancée dans la dépêche de l'AFP : "Plus un homme marié accorde de temps aux tâches domestiques, moins il a de relations sexuelles", et des titres de journaux cités ?
En particulier, pensez-vous qu'un test statistique peut permettre d'établir une relation de cause à effet ? Expliquez.

Extraits des tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi gaussienne : on donne, pour différentes valeurs de q, la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0,1)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(N \leq q)$	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(N \leq q)$	0.90	0.92	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	0.99

Table de la loi de Student : on donne, pour différentes valeurs de $\alpha \in [0,1]$, t_α tel que $P(T \leq t_\alpha) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique), avec $n = 99$, 100 ou 101.

α	0.9	0.95	0.975
t_α	1.290	1.66	1.984

Table de la loi de Student : on donne, pour différentes valeurs de t, la valeur de $P(T \leq t)$ lorsque $T \sim \mathcal{T}(n)$, avec $n = 99$, 100 ou 101.

t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(T \leq t)$	0.54	0.579	0.618	0.655	0.691	0.725	0.757	0.787	0.815	0.84	0.863	0.884
t	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(T \leq t)$	0.902	0.918	0.932	0.944	0.954	0.963	0.97	0.976	0.981	0.985	0.988	0.991

Table de la loi binomiale : on donne, pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi binomiale de paramètres 900 et 0.6.

k	514	515	516	517	...	562	563	564	565
$P(X \leq k)$	0.042	0.048	0.055	0.063	...	0.938	0.946	0.953	0.959

Correction

Partie I : un premier angle de vue

1. Soit $X = (X_1, \dots, X_{900})$ un 900-échantillon de la loi de Bernoulli de paramètre θ, modélisant les réponses des 900 hommes mariés interrogés, tel que $X_i = 1$ si le i-éme homme interrogé répond avoir au moins 4 rapports sexuels par mois, 0 sinon. Soit $x = (x_1, \dots, x_{900})$ l'observation de cet échantillon, $\mathcal{X} = \{0,1\}^{900}$, $\mathcal{A} = \mathcal{P}(\{0,1\}^{900})$. Pour $\theta \in \Theta =]0,1[$, on note $P_\theta = \mathcal{B}(\theta)^{\otimes 900}$. Le modèle statistique considéré est alors représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.

2. On peut tester $(H_0) : \theta \geq 0.6$ contre $(H_1) : \theta < 0.6$. On souhaite alors se prémunir en priorité du risque de déclarer que la probabilité pour un homme marié d'avoir au moins 4 rapports sexuels par mois est plus faible s'il participe pour plus de 30% aux travaux domestiques traditionnellement féminins, alors que ce n'est pas vrai.

3. On considère la statistique de test $T(X) = \sum_{i=1}^{900} X_i$ correspondant au nombre d'hommes déclarant avoir au moins 4 rapports sexuels par mois parmi les 900 hommes interrogés. Lorsque $\theta = 0.6$, $T(X)$ suit une loi binomiale de paramètres $(900, 0.6)$.

4. On sait que $T(X)/900$ est un estimateur sans biais, asymptotiquement normal de θ, donc de petites valeurs de $T(X)$ favorisent plutôt le choix de (H_1). Par conséquent, la région critique du test de (H_0) contre (H_1) est de la forme $\mathcal{R}_{(H_0)} = \{x,\ T(x) \leq s\}$, avec $s \in \mathbb{N}$.

5. La valeur critique s doit vérifier $\sup_{\theta\geq 0.6} P_\theta(\{x,\ T(x) \leq s\}) \leq 0.05$ ou encore $P_{0.6}(\{x,\ T(x) \leq s\}) \leq 0.05$, avec $s \in \mathbb{N}$ la plus grande possible. On a vu que lorsque $X \sim P_{0.6}$, $T(X) \sim \mathcal{B}(900, 0.6)$, donc on choisit $s = 515$.

6. Ici, $T(x) = 513$, donc on rejette (H_0) au profit de (H_1) pour un niveau 5%.

Partie II : un deuxième angle de vue

1. Soit $X = (X_1, \dots, X_{100})$ un 100 échantillon de la loi $\mathcal{N}(m, \sigma^2)$ modélisant le pourcentage de participation aux travaux domestiques traditionnellement féminins pour 100 hommes ayant au moins 4 rapports sexuels par mois. Soit $x = (x_1, \dots, x_{100})$ l'observation de cet échantillon. Le modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, avec $\mathcal{X} = \mathbb{R}^{100}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{100})$, $\Theta = [0, 100] \times \mathbb{R}_+^*$, et pour $\theta = (m, \sigma^2) \in \Theta$, $P_\theta = \mathcal{N}(m, \sigma^2)^{\otimes 100}$, c'est-à-dire la loi d'un 100-échantillon de la loi gaussienne d'espérance m et de variance σ^2.

2. On se prémunit en priorité du risque de déclarer qu'un homme marié ayant au moins 4 rapports sexuels par mois participe moins en moyenne aux travaux domestiques traditionnellement féminins qu'un homme dont la fréquence sexuelle mensuelle est inférieure à 4, à tort.

3. On peut prendre comme statistique de test $T(X) = \sqrt{100}\frac{\bar{X}-30}{S(X)}$, avec $S^2(X) = \frac{1}{99}\sum_{i=1}^{100}(X_i - \bar{X})^2$. Lorsque X suit la loi $P_{(30,\sigma^2)}$, $T(X)$ suit une loi de Student à 99 degrés de liberté.

4. $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s\}$.

5. L'équation du niveau s'écrit $\sup_{\sigma^2\in\mathbb{R}_+^*} P_{(30,\sigma^2)}(\mathcal{R}_{(H_0)}) \leq 0.05$. Si $X \sim P_{(30,\sigma^2)}$, on a vu que $T(X) \sim \mathcal{T}(99)$, donc on prend $s = -1.66$.

6. Ici, $T(x) = -1.96$, donc on rejette (H_0) au profit de (H_1) pour un niveau 5%.

7. La p-valeur du test est donnée par $p(x) = P_{(30,\sigma^2)}(\{y, T(y) \leq -1.96\}) \in [0.024, 0.03]$. On a $p(x) < 0.05$, ce qui confirme bien qu'on rejette (H_0) au profit de (H_1) pour un niveau 5%.

3.5.2 Élections présidentielles 2012

Problème 1

Début 2011, les sondages concernant le premier tour des élections présidentielles de 2012 se multiplient. Dans le scénario des candidatures de Dominique Strauss-Kahn, Marine Le Pen et Nicolas Sarkozy, les pourcentages d'intention de vote pour chaque éventuel candidat obtenus par les sondages de Harris Interactive, Ifop et CSA sont les suivants :

Sondage	Nombre de personnes sondées	D. Strauss-Kahn	M. Le Pen	N. Sarkozy
Harris Interactive	1347	23%	24%	20%
Ifop	1046	29%	21%	23%
CSA	853	30%	21%	19%

Les méthodes utilisées pour le recueil des données sont différentes pour chaque sondage.

On s'intéresse ici aux pourcentages d'intention de vote pour D. Strauss-Kahn.
On note p_1, p_2 et p_3 les probabilités pour un électeur de voter pour D. Strauss-Kahn selon les sondages Harris Interactive, Ifop et CSA respectivement.
On souhaite construire un test multiple de l'hypothèse $(H_0) : p_1 = p_2 = p_3$ contre l'alternative (H_1) : il existe (i, j), $i \neq j$ tel que $p_i \neq p_j$.

1. Construire un test de niveau asymptotique 5% de l'hypothèse $(H_0') : p_1 = p_2$ contre l'alternative $(H_1') : p_1 \neq p_2$. On note $\mathcal{R}_{1,2}$ la région critique correspondante.
2. Peut-on affirmer que les résultats sur les intentions de vote pour D. Strauss-Kahn fournis par les sondages Harris Interactive et Ifop sont fiables ?
3. Tester de la même façon à l'aide d'un test de niveau asymptotique 5% l'hypothèse : $p_1 = p_3$ contre l'alternative : $p_1 \neq p_3$, puis l'hypothèse : $p_2 = p_3$ contre l'alternative : $p_2 \neq p_3$. On note $\mathcal{R}_{1,3}$ et $\mathcal{R}_{2,3}$ les régions critiques correspondantes.
4. On décide de rejeter l'hypothèse (H_0) du test multiple si l'on rejette l'hypothèse nulle pour au moins l'un des trois tests précédents, autrement dit la région critique du test multiple est définie par

$$\mathcal{R}_{(H_0)} = \mathcal{R}_{1,2} \cup \mathcal{R}_{1,3} \cup \mathcal{R}_{2,3}.$$

En utilisant l'inégalité de Bonferroni $P(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} P(A_i)$, donner un niveau asymptotique pour le test multiple. Quelle est la conclusion du test pour ce niveau asymptotique ? Pour justifier les résultats obtenus, le journal *Libération* invoque "la volatilité de l'opinion", qu'en pensez-vous ?
5. On souhaite maintenant tester l'hypothèse que $p_3 \leq 30\%$ contre $p_3 > 30\%$. Expliquer la démarche à adopter dans ce cas.

Correction

1. Modèle statistique : Soit $Y = (Y_1, \dots, Y_{n_1})$ un n_1-échantillon de la loi $\mathcal{B}(p_1)$, modélisant le comportement des personnes sondées par Harris Interactive (Y_i vaut 1 si la personne i i a l'intention de voter pour DSK, 0 sinon), et $Z = (Z_1, \dots, Z_{n_2})$ un n_2-échantillon de la loi $\mathcal{B}(p_2)$, modélisant le comportement des personnes sondées par Ifop (Z_i vaut 1 si la personne i a l'intention de voter pour DSK, 0 sinon), avec $(p_1, p_2) \in \Theta = [0,1]^2$ inconnu. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$. Soit $x = (y, z)$ avec $y = (y_1, \dots, y_{n_1})$, $z = (z_1, \dots, z_{n_2})$, l'observation de ces deux échantillons indépendants. Soit $\mathcal{X} = \{0,1\}^{n_1+n_2}$, et $\mathcal{A}$ l'ensemble des parties de $\mathcal{X}$. Pour $\theta = (p_1, p_2) \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{B}(p_1)^{\bigotimes n_1} \bigotimes \mathcal{B}(p_2)^{\bigotimes n_2}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.
Hypothèses : $(H_0') : p_1 = p_2$ contre $(H_1') : p_1 \neq p_2$.
Statistique de test : $T(x) = \dfrac{\bar{y} - \bar{z}}{\sqrt{\frac{n_1\bar{y}+n_2\bar{z}}{n_1+n_2}\left(1-\frac{n_1\bar{y}+n_2\bar{z}}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$.
Loi asymptotique de $T(X)$ sous (H_0') : sous (H_0'), $T(X)$ converge en loi vers une loi $\mathcal{N}(0,1)$.
Fonction de test : puisque $\bar{Y}$ et $\bar{Z}$ sont les estimateurs empiriques de p_1 et p_2, on rejette (H_0') pour de grandes valeurs de $|T(x)|$. La région critique s'écrit $\mathcal{R}_{1,2} = \{x, |T(x)| \geq s\}$. La fonction de test correspondante $\phi_{1,2}(x) = \mathbb{1}_{|T(x)|\geq s}$. Ce test est asymptotique puisque la loi de $T(X)$ n'est connue qu'asymptotiquement lorsque $p_1 = p_2$.
Calcul de s pour un niveau asymptotique 5% : si $p_1 = p_2 = p$, i.e. si $X \sim P_{(p,p)}$, $T(X)$ converge en loi vers une loi $\mathcal{N}(0,1)$, donc comme $n_1 = 1347$ et $n_2 = 1046$, on peut choisir $s = 1.96$. La région critique asymptotique s'écrit finalement $\mathcal{R}_{1,2} = \{x, |T(x)| \geq 1.96\}$.
2. Conclusion : on a ici $T(x) = -3.33$, donc on rejette l'hypothèse d'égalité des probabilités p_1 et p_2 pour un niveau asymptotique 5%. Les deux sondages, effectués à quelques jours d'intervalle, ne semblent pas fiables.
3. Les deux autres tests se construisent de la même façon, les valeurs des statistiques de test sont égales respectivement à -3.66 et -0.476. Pour des niveaux asymptotiques 5%, on rejette donc l'hypothèse d'égalité de p_1 et p_3, mais on accepte l'hypothèse d'égalité de p_2 et p_3.
4. Sous (H_0), la probabilité de $\mathcal{R}_{(H_0)}$ est majorée par 15% asymptotiquement. Pour un niveau asymptotique de 15%, les probabilités d'intention de vote pour DSK ne sont pas toutes égales suivant les sondages. Problème de "volatilité de l'opinion" ou de fiabilité des sondages ?
5. Modèle : Soit $X = (X_1, \dots, X_{853})$ un 853-échantillon d'une loi de Bernoulli de paramètre p_3, avec $p_3 \in \Theta = [0,1]$, modélisant l'intention de vote des personnes sondées par CSA ($X_i = 1$ si l'électeur i est favorable à l'éventuel candidat, 0 sinon), et $x = (x_1, \dots, x_{853})$ l'observation de cet échantillon. Soit $\mathcal{X} = \{0,1\}^{853}$, et $\mathcal{A}$ l'ensemble des parties de $\mathcal{X}$. Pour $p_3 \in \Theta$, on note P_{p_3} la loi de X : $P_{p_3} = \mathcal{B}(p_3)^{\bigotimes 853}$. Le modèle statistique est alors défini par $(\mathcal{X}, \mathcal{A}, (P_{p_3})_{p_3\in[0,1]})$.
Statistique de test : $T(x) = \sum_{i=1}^{853} x_i$.
Fonction de test : puisque $T(x)/853$ est l'estimateur empirique du paramètre p_3, on choisit de rejeter l'hypothèse nulle lorsque $T(x)$ prend de grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{T(x)\geq s}$.
Loi de $T(X)$ sous l'hypothèse $p_3 = 30\%$: si $X \sim P_{0.3}$, $T(X) \sim \mathcal{B}(853, 0.3)$. On peut ici utiliser l'approximation gaussienne. Dans ce cas, on sait que si $X \sim P_{0.3}$, $T(X)$ peut être approchée par la loi $\mathcal{N}(255.9, 179.13)$. Choisir s tel que $P_{0.3}(\{x, T(x) \geq s\}) \leq 0.05$ revient donc à choisir s tel que $P\left(N \geq \frac{s-255.9}{\sqrt{179.13}}\right) \leq 0.05$, où $N \sim \mathcal{N}(0,1)$. Alors $\frac{s-255.9}{\sqrt{179.13}} = 1.645$ convient, d'où $s = 277.92$.
On prend donc $s = 278$, i.e. $\phi(x) = \mathbb{1}_{T(x)\geq 278}$.
Ici, $T(x) = 256$ donc on ne rejette pas l'hypothèse : $p_3 \leq 0.3$ pour un niveau 5%.

Problème 2

On s'intéresse aux sondages Ifop du 15 mars 2012 portant sur l'élection présidentielle de 2012 en France. Ces sondages sont réalisés sur un échantillon représentatif de 928 personnes. On souhaite étudier, à partir de ces sondages, la probabilité qu'une personne prise au hasard parmi les électeurs vote pour un candidat donné. On note θ cette probabilité, et on souhaite réaliser un test de l'hypothèse nulle (H_0) : "$\theta \leq \theta_0$" contre l'aternative (H_1) : "$\theta > \theta_0$", où θ_0 est une valeur donnée dans $\Theta =]0, 1[$.

1. Décrire la variable aléatoire X modélisant les réponses des 928 personnes sondées à la question : "Avez-vous l'intention de voter pour le candidat en question ?", puis décrire le modèle statistique considéré.

. . . . . .
. . . . . .
. . . . . .
. . . . . .

2. Que signifie le choix des hypothèses (H_0) et (H_1) ici ?

. . . . . .
. . . . . .
. . . . . .

3. Donner la statistique de test à considérer, ainsi que sa loi lorsque $\theta = \theta_0$.

. . . . . .
. . . . . .
. . . . . .

Préciser sous quelles conditions sur θ_0 cette loi peut être approchée par une loi de Poisson et par une loi gaussienne.

. . . . . .
. . . . . .
. . . . . .

4. Donner la forme de la région critique du test de (H_0) contre (H_1), en la justifiant.

. . . . . .
. . . . . .
. . . . . .

5. On considère le premier tour des élections, dont les taux de remboursement de frais de campagne sont régis par la loi organique du 28 février 2012. Cette loi prévoit, pour un plafond de dépenses de 16,851 millions d'euros, un taux de remboursement s'élevant à 4.75% des dépenses pour un candidat obtenant 5% ou moins des voix au premier tour, et un taux s'élevant à 47.5% pour un candidat obtenant plus de 5% des voix. La barre des 5% est donc un enjeu d'importance pour chacun des candidats du premier tour de l'élection présidentielle.

a) Préciser la région critique du test de (H_0) contre (H_1) de niveau $\alpha = 0.05$ lorsque $\theta_0 = 0.05$, à l'aide de la table de la loi exacte de la statistique de test lorsque $\theta = 0.05$.

.....
.....
.....
.....

b) Préciser la région critique obtenue à l'aide d'une approximation de cette loi. Que constate-t-on ?

.....
.....
.....

6. Sur les 928 personnes interrogées, 92 ont répondu qu'elles avaient l'intention de voter pour Jean-Luc Mélenchon. Quel taux de remboursement des frais de campagne Jean-Luc Mélenchon doit-il s'attendre à obtenir d'après ce sondage ?

.....
.....

7. On s'intéresse maintenant au second tour des élections, et à la probabilité que François Hollande soit élu face à Nicolas Sarkozy.

a) Préciser la région critique du test de (H_0) contre (H_1) de niveau $\alpha = 0.05$ lorsque $\theta_0 = 0.5$, à l'aide de la table de la loi exacte de la statistique de test lorsque $\theta = 0.5$.

.....
.....
.....
.....

b) Préciser la région critique obtenue à l'aide d'une approximation de cette loi. Que constate-t-on ?

.....
.....
.....
.....

8. Sur les 928 personnes interrogées, 497 ont répondu qu'elles avaient l'intention de voter pour François Hollande (53.5% d'intention de vote) contre Nicolas Sarkozy au second tour. Que peut-on conclure (à la date du sondage) ?

. . . . . .
. . . . . .

Extraits des tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi binomiale : on donne pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi binomiale de paramètres 928 et 0.05.

k	34	35	36	37	...	56	57	58	59
$P(X \leq k)$	0.032	0.046	0.064	0.087	...	0.933	0.949	0.962	0.972

Table de la loi binomiale : on donne pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi binomiale de paramètres 928 et 0.5.

k	437	438	439	440	...	487	488	489	490
$P(X \leq k)$	0.041	0.047	0.054	0.061	...	0.939	0.946	0.953	0.959

Table de la loi de Poisson : on donne pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi de Poisson de paramètre 46.4.

k	34	35	36	37	...	56	57	58	59
$P(X \leq k)$	0.0355	0.0501	0.0688	0.0923	...	0.9274	0.9445	0.9582	0.9689

Table de la loi de Poisson : on donne pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi de Poisson de paramètre 464.

k	427	428	429	430	...	498	499	500	501
$P(X \leq k)$	0.044	0.048	0.053	0.059	...	0.944	0.949	0.954	0.958

Correction

1. Soit $X = (X_1, \dots, X_{928})$ un 928-échantillon de la loi de Bernoulli de paramètre θ, modélisant les réponses des 928 personnes sondées, tel que $X_i = 1$ si la i-ème personne sondée répond avoir l'intention de voter pour le candidat, 0 sinon. Soit $x = (x_1, \dots, x_{928})$ l'observation de cet échantillon, $\mathcal{X} = \{0,1\}^{928}$, $\mathcal{A} = \mathcal{P}(\{0,1\}^{928})$. Pour $\theta \in \Theta$, on note $P_\theta = \mathcal{B}(\theta)^{\otimes 928}$. Le modèle statistique considéré est alors représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.

2. Par ce choix d'hypothèses, on souhaite se prémunir en priorité du risque de déclarer que $\theta > \theta_0$ alors qu'en réalité $\theta \leq \theta_0$.

3. On considère la statistique de test $T(X) = \sum_{i=1}^{928} X_i$ correspondant au nombre de personnes ayant l'intention de voter pour le candidat donné parmi les 928 personnes sondées. Lorsque $\theta = \theta_0$, $T(X)$ suit une loi binomiale de paramètres $(928, \theta_0)$.
Si $\theta_0 < 0.1$ ou si $\theta_0 < 5/928$ c'est-à-dire si $\theta_0 < 0.0054$, la loi binomiale de paramètres $(928, \theta_0)$ peut être approchée par une loi de Poisson de paramètre $928\theta_0$. Si $\theta_0 \geq 5/928$ et $\theta_0 \leq 1 - 5/928$ donc si $\theta_0 \geq 0.0054$ et $\theta_0 \leq 0.9946$, elle peut être approchée par une loi gaussienne d'espérance $928\theta_0$ et de variance $928\theta_0(1-\theta_0)$.

4. On sait que $T(X)/928$ est un estimateur sans biais, asymptotiquement normal de θ, donc de grandes valeurs de $T(X)$ favorisent plutôt le choix de (H_1). Par conséquent, la région critique du test de (H_0) contre (H_1) est de la forme $\mathcal{R}_{(H_0)} = \{x,\ T(x) \geq s\}$.

5. a) La valeur critique s doit vérifier $P_{0.05}(\{x,\ T(x) \geq s\}) \leq 0.05$, avec $s \in \mathbb{N}$, ce qui est équivalent à $P_{0.05}(\{x,\ T(x) \leq s-1\}) \geq 0.95$. Or lorsque $X \sim P_{0.05}$, $T(X) \sim \mathcal{B}(928, 0.05)$, donc on choisit $s - 1 = 58$, c'est-à-dire $s = 59$.

5. b) En approchant la loi $\mathcal{B}(928, 0.05)$ par la loi de Poisson de paramètre 46.4, on obtient $s = 59$ également. On obtient la même région critique.

6. Ici, $T(x) \geq 59$, donc on rejette (H_0) au profit de (H_1) pour un niveau $\alpha = 0.05$.

7. a) La valeur critique s doit vérifier ici : $P_{0.5}(\{x,\ T(x) \leq s-1\}) \geq 0.95$, avec $s \in \mathbb{N}$. Or lorsque $X \sim P_{0.5}$, $T(X) \sim \mathcal{B}(928, 0.5)$, donc on choisit $s - 1 = 489$, c'est-à-dire $s = 490$.

7. b) Ici, on approche la loi binomiale de paramètres $(928, 0.5)$ par une loi normale d'espérance 464 de variance 232. On veut $P_{0.5}(\{x,\ T(x) \geq s\}) \leq 0.05$, ce qui équivaut à

$$P_{0.5}\left(\left\{x,\ \frac{T(x) - 464}{\sqrt{232}} \geq \frac{s - 464}{\sqrt{232}}\right\}\right) \leq 0.05,$$

donc on choisit s tel que $(s - 464)/\sqrt{232} \geq 1.645$, c'est-à-dire $s \geq 489.056$ ou encore $s = 490$ également.

8. Ici, $T(x) = 497$ donc on rejette (H_0) au profit de (H_1) pour un niveau $\alpha = 0.05$.

3.5.3 Fermeture de Megaupload

Problème

Depuis la fermeture, le 19 janvier 2012, du site de téléchargement direct et gratuit Megaupload ordonnée par le gouvernement américain par l'intermédiaire du FBI, les sites légaux de télévision en replay (télévision de rattrapage) et de vente de VOD (Video On Demand) se félicitent d'une importante hausse de leur trafic. S'il est vrai que cette hausse est flagrante dans les jours qui ont suivi la fermeture de Megaupload, plus d'un mois après, cette tendance n'est pas si évidente.
En étudiant la fréquentation journalière du site `m6replay.fr` (en pourcentage d'usagers internet) sur 6 mois, on peut supposer qu'elle s'éléve en moyenne avant le 19 janvier à 0.02.
Les fréquentations journalières de `m6replay.fr` (en pourcentage d'usagers internet) relevées à intervalles réguliers entre le 20 janvier et le 29 février 2012 s'établissent de la façon suivante (source : `Alexa`).

0.026	0.016	0.027	0.015	0.024	0.016	0.023	0.018	0.028	0.016	0.026

On suppose que la fréquentation journalière de `m6replay.fr` après le 20 janvier suit une loi normale d'espérance m, d'écart-type σ.

1. Décrire la variable aléatoire X modélisant les données et le modèle statistique considéré.

. . . . . .
. . . . . .
. . . . . .

2. Donner l'estimateur empirique usuel $S(X)$ de σ et calculer sa valeur $S(x)$.

. . . . . .

On en déduit qu'il n'est pas aberrant de supposer que $\sigma = 0.005$. Que faudrait-il faire pour pouvoir l'affirmer avec plus de certitude ?

. . . . . .

On suppose donc dans les questions suivantes que $\sigma = 0.005$.

3. Décrire le nouveau modèle statistique considéré.

. . . . . .
. . . . . .

4. On souhaite tester l'hypothèse nulle $(H_0) : m = 0.02$ contre l'alternative $(H_1) : m > 0.02$.

a) De quel risque se prémunit-on en priorité en considérant ces hypothèses ?

. . . . . .
. . . . . .

b) Déterminer une statistique de test $T(X)$, et donner sa loi sous (H_0).

. . . . . .
. . . . . .

c) Choisir la forme de la région critique du test :

- ☐ $\mathcal{R}_{(H_0)} = \{x, T(x) \leq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, |T(x)| \geq s\}$
- ☐ $\mathcal{R}_{(H_0)} = \{x, |\sqrt{11}\frac{T(x)-0.02}{0.005}| \geq s\}$

Justifier la réponse.

. . . . . .
. . . . . .

d) On veut construire un test de niveau $\alpha = 5\%$. Représenter sur un graphe commenté la valeur de s qu'il faut prendre pour que le test soit bien du niveau choisi.

Déterminer précisément la valeur de s à l'aide des tables statistiques de la loi normale centrée réduite.

. . . . . .
. . . . . .

e) Quelle est la conclusion du test ?

. . . . . .

f) Donner une valeur approchée de la p-valeur du test et expliquer pourquoi cette valeur corrobore la conclusion précédente.

. . . . . .
. . . . . .

g) Donner une valeur approchée de la puissance du test pour $m = 0.024$.

. . . . . .
. . . . . .

4. Construire maintenant un test de l'hypothèse nulle $(H_0) : m > 0.02$ contre l'alternative $(H_1) : m = 0.02$, de niveau 5%. Quelle est la conclusion du test ? Que peut-on en dire ?

. . . . . .
. . . . . .
. . . . . .
. . . . . .

Extraits des tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0, 1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0, 1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi gaussienne : on donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0, 1)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(N \leq q)$	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(N \leq q)$	0.90	0.92	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	0.99

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0, 1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{10,\alpha}$	1.372	1.812	2.228
$t_{11,\alpha}$	1.363	1.796	2.201
$t_{12,\alpha}$	1.356	1.782	2.179
$t_{13,\alpha}$	1.350	1.771	2.160

Correction

1. Soit $X = (X_1, \dots, X_{11})$ un 11 échantillon de la loi $\mathcal{N}(m, \sigma^2)$ modélisant les fréquentations journalières pour les 11 jours relevés. Soit $x = (x_1, \dots, x_{11})$ l'observation de cet échantillon. Le modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, avec $\mathcal{X} = \mathbb{R}^{11}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{11})$, $\Theta = [0, 100] \times \mathbb{R}_+^*$, et pour $\theta = (m, \sigma^2) \in \Theta$, $P_\theta = \mathcal{N}(m, \sigma^2)^{\otimes 11}$, c'est-à-dire la loi d'un 11-échantillon de la loi normale d'espérance m et de variance σ^2.

2. L'estimateur empirique de l'écart-type σ est donné par $S(X) = \sqrt{\frac{1}{11}\sum_{i=1}^{11}(X_i - \bar{X})^2}$. Sa valeur est $S(x) = 0.00516$. Il n'est donc pas aberrant de supposer que $\sigma = 0.005$. Pour pouvoir l'affirmer avec plus de certitude, on peut envisager un test de $(H_0) : \sigma = 0.005$ contre $(H_1) : \sigma \neq 0.005$.

3. Le nouveau modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_m)_{m\in\Theta})$, avec $\mathcal{X} = \mathbb{R}^{11}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{11})$, $\Theta = [0, 100]$, et pour $m \in \Theta$, $P_m = \mathcal{N}(m, 0.005^2)^{\otimes 11}$, c'est-à-dire la loi d'un 11-échantillon de la loi normale d'espérance m et de variance 0.005^2.

4. a) On se prémunit en priorité du risque de déclarer que la fréquentation journalière de `m6replay.fr` a augmenté depuis la fermeture de Megaupload à tort.

b) On peut choisir par exemple $T(X) = \sqrt{11}\frac{\bar{X}-0.02}{0.005}$ dont la loi lorsque X suit la loi $P_{0.02}$ est une loi normale centrée réduite.

c) On choisit la forme 2. De grandes valeurs de $T(X)$ impliquent de grandes valeurs de $\bar{X}$. Or, puisque $\bar{X}$ est un estimateur sans biais et asymptotiquement normal de m, de grandes valeurs de $\bar{X}$ favorisent plutôt le choix de (H_1) donc le rejet de (H_0). On choisit donc de rejeter (H_0) au profit de (H_1) lorsque $T(x)$ dépasse une certaine valeur critique s.

d) L'équation du niveau s'écrit $P_{0.02}(\mathcal{R}_{(H_0)}) = 0.05$. Si $X \sim P_{0.02}$, on a vu que $T(X) \sim \mathcal{N}(0, 1)$, donc on prend $s = 1.645$.

e) Ici, $T(x) = 0.9045$, donc on ne rejette pas (H_0) au profit de (H_1) pour un niveau 5%.

f) La p-valeur du test est donnée par $p(x) = P_{0.02}(\{x, T(x) \geq 0.9045\}) \simeq 1 - F(0.9) \simeq 0.18$, où F est la fonction de répartition de la loi normale centrée réduite. On a $p(x) > 0.05$, ce qui veut dire que $0.9045 < s$ donc on ne rejette pas (H_0) pour un niveau 5%.

g) Puissance du test pour $m = 0.024$:

$$\gamma(0.024) = P_{0.024}\left(\left\{x,\ \sqrt{11}\frac{\bar{x} - 0.024}{0.005} \geq 1.645 - \sqrt{11}\frac{0.024 - 0.02}{0.005}\right\}\right) \simeq 1 - F(-1) \simeq 0.84.$$

4. Cette fois, $\mathcal{R}'_{(H_0)} = \{x, T(x) \leq -1.645\}$, donc on ne rejette pas (H_0) au profit de (H_1) pour un niveau 5%. Les deux conclusions semblent contradictoires. Elles ne le sont pas en fait, car on ne place pas le meilleur niveau de confiance dans la même conclusion. On se prémunit ici en priorité du risque de déclarer que la fréquentation n'a pas bougé alors qu'elle a augmenté.

3.5.4 Crise publicitaire pour M6

Problème

D'après l'Agence France-Presse, en 2010, la crise publicitaire semble dépassée pour la chaîne de télévision M6. En 2009, la chaîne enregistrait des recettes publicitaires mensuelles de 50.5 millions d'euros en moyenne. Ces recettes s'établissent pour les douze mois de l'année 2010 (en millions d'euros) à

50.6	62.6	51.7	55.1	52.3	62.3	60.2	66.8	42.6	53.2	51.2	61.6

Peut-on conclure à une réelle reprise des investissements publicitaires sur M6 ?

On suppose que les recettes publicitaires mensuelles (en millions d'euros) pour 2009 suivent une loi normale d'espérance 50.5 de variance σ^2, et que ces recettes pour 2010 suivent une loi normale, d'espérance m inconnue, de même variance σ^2.

On souhaite tester l'hypothèse nulle $(H_0) : m = 50.5$ contre l'alternative $(H_1) : m > 50.5$.

1. De quel risque se prémunit-on en priorité en considérant ces hypothèses de test ?

. . . . . .
. . . . . .

2. On suppose dans un premier temps que σ^2 est connue égale à 9. Décrire la variable aléatoire $X = (X_1, \ldots, X_{12})$ modélisant les données et le modèle statistique $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ considéré.

. . . . . .
. . . . . .
. . . . . .
. . . . . .

3. Quelle statistique de test notée $T(X)$ choisira-t-on ? (Entourer la réponse)

a) $T(X) = \bar{X}$, avec $\bar{X} = \frac{1}{12}\sum_{i=1}^{12} X_i$.
b) $T(X) = \sqrt{12}\frac{\bar{X}-50.5}{3}$.
c) $T(X) = \sqrt{12}\frac{\bar{X}-50.5}{\sqrt{S^2(X)}}$, où $S^2(X) = \frac{1}{11}\sum_{i=1}^{12}(X_i - \bar{X})^2$.
d) $T(X) = \sqrt{12}\frac{\bar{X}}{\sqrt{S^2(X)}}$.

4. Quelle est la loi de cette statistique de test sous l'hypothèse (H_0) ?

. . . . . .

5. On considère deux tests distincts dont les régions critiques et les fonctions de tests sont respectivement $\mathcal{R}^1_{(H_0)} = \{x, T(x) \leq s_1\}$, $\phi_1(x) = \mathbb{1}_{\{x,T(x)\leq s_1\}}$ et $\mathcal{R}^2_{(H_0)} = \{x, T(x) \geq s_2\}$, $\phi_2(x) = \mathbb{1}_{\{x,T(x)\geq s_2\}}$.
Déterminer les valeurs de s_1 et s_2 pour que les deux tests ϕ_1 et ϕ_2 soient de taille 5%. Justifier les réponses en s'aidant d'un graphe.

Détermination de s_1 :
.....

Détermination de s_2 :
.....

6. Donner une approximation des puissances de ces deux tests en $m = 52$.

.....
.....
.....
.....

7. Quel test choisira-t-on sur la base de ces puissances ? Ce choix est-il en accord avec l'intuition ? Expliquer.

.....
.....
.....

8. Dans un second temps, on ne suppose plus la variance σ^2 connue. Décrire le nouveau modèle statistique $(\mathcal{X}', \mathcal{A}', (P'_\theta)_{\theta\in\Theta'})$ considéré.

.....
.....
.....

9. Parmi les statistiques de test proposées dans la question 3, laquelle choisira-t-on ici ? Quelle est sa loi sous (H_0) ?

.....
.....

10. Déterminer la fonction de test ϕ correspondant à un test intuitif construit sur $T(X)$, de

taille 5%. Donner la conclusion de ce test.

.....
.....
.....
.....
.....

Extraits des tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi gaussienne : on donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0,1)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(N \leq q)$	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(N \leq q)$	0.90	0.92	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	0.99

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{10,\alpha}$	1.372	1.812	2.228
$t_{11,\alpha}$	1.363	1.796	2.201
$t_{12,\alpha}$	1.356	1.782	2.179
$t_{13,\alpha}$	1.350	1.771	2.160

Correction

1. On privilégie l'hypothèse (H_0) dans le sens où l'on souhaite la conserver tant qu'on n'a pas suffisamment de preuves pour affirmer qu'elle est fausse. On souhaite ainsi se prémunir en priorité du risque de déclarer que les recettes publicitaires de M6 ont augmenté de façon significative alors qu'elles ont stagné.

2. Soit $X = (X_1, \dots, X_{12})$ un 12 échantillon de la loi $\mathcal{N}(m, \sigma^2)$ modélisant les recettes publicitaires pour les 12 mois de l'année 2010. Soit $x = (x_1, \dots, x_{12})$ l'observation de cet échantillon. Le modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, avec $\mathcal{X} = \mathbb{R}^{12}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{12})$, $\Theta = \mathbb{R}_+$, et pour $\theta \in \Theta$, $P_\theta = \mathcal{N}(\theta, 9)^{\otimes 12}$, c'est-à-dire la loi d'un 12-échantillon de la loi normale d'espérance θ et de variance 9.

3. a) et b) sont possibles. On choisira ici b).

4. Sous (H_0), $T(X)$ suit une loi normale centrée réduite.

5. $s_1 = -1.645$ et $s_2 = 1.645$.

6. Si F désigne la fonction de répartition d'une loi normale centrée réduite, la puissance du test ϕ_1 est égale à $\gamma_1 = P_{52}(\{x, T(x) \leq -1.645\}) = F(-1.5/(3/\sqrt{(12)}) - 1.645) \simeq 0$ et la puissance du test ϕ_2 est égale à $\gamma_2 = P_{52}(\{x, T(x) \geq 1.645\}) = 1 - F(-1.5/(3/\sqrt{(12)}) + 1.645) \simeq 0.535$.

7. On choisit le deuxième test. En effet, si $T(x)$ prend de grandes valeurs, $\bar{x}$ prend aussi de grandes valeurs, et puisque $\bar{X}$ est un estimateur sans biais asymptotiquement normal de m, on rejettera (H_0) au profit de (H_1).

8. Modèle statistique : $\mathcal{X}' = \mathcal{X} = \mathbb{R}^{12}$, $\mathcal{A}' = \mathcal{A}$, $\Theta' = \mathbb{R}_+ \times \mathbb{R}_+^*$ et pour $\theta = (m, \sigma^2) \in \Theta'$, on note $P'_\theta = \mathcal{N}(m, \sigma^2)^{\bigotimes 12}$.

9. c) Sous (H_0), $T(X)$ suit une loi de Student à 11 degrés de liberté.

10. Ici $\mathcal{R}_{(H_0)} = \{x, T(x) \geq s\}$, $\phi(x) = \mathbb{1}_{\{x, T(x)\geq s\}}$ avec $s = 1.796$. On a $T(x) = 2.69$ donc on rejette bien l'hypothèse (H_0) au profit de (H_1) pour un niveau 5%.

Chapitre 4

Tests de Neyman-Pearson, tests uniformément plus puissants

Pour ce chapitre, on se réfère en particulier à l'ouvrage de Lehmann et Romano [15].

4.1 Tests randomisés

Le contexte des tests non randomisés est en quelque sorte idéal dans le sens où, au vu d'une observation x, on décide si l'on rejette ou non (H_0) au profit de (H_1). On se place ici dans un contexte où, pour certaines valeurs de x, il peut être difficile de prendre une décision (exemple des tests construits sur une statistique de loi discrète). On est alors amené à utiliser un test randomisé, où l'on considère que pour certaines valeurs de x, la valeur $\phi(x)$ de sa fonction de test peut être un nombre strictement compris entre 0 et 1 définissant la probabilité de rejeter (H_0) au profit de (H_1) avec l'observation x.

Définition 17 (Test randomisé)**.** *On appelle* test randomisé *un test dont la fonction de test est une statistique ϕ de $\mathcal{X}$ dans* $[0,1]$. *Pour une observation x de $\mathcal{X}$ telle que $\phi(x) = 1$, on décide de rejeter (H_0) au profit de (H_1), pour x telle que $\phi(x) = 0$, on décide de ne pas rejeter (H_0) au profit de (H_1), enfin, pour x telle que $\phi(x) \in]0,1[$, on décide de rejeter (H_0) au profit de (H_1) avec probabilité $\phi(x)$.*

Remarque. L'intérêt des tests randomisés est avant tout théorique. Dans la pratique, ils ne sont que très rarement utilisés.

Interprétation. On peut voir un test randomisé comme la première partie d'un processus de décision en deux étapes. Au vu de l'observation x,

1. On définit une probabilité $\phi(x)$ de rejeter (H_0) au profit de (H_1).
2. On réalise un tirage aléatoire dans $\{0,1\}$ selon une loi de Bernoulli de paramètre $\phi(x)$. On note y la valeur obtenue, Y la variable aléatoire dont elle est issue.

On décide finalement de ne pas rejeter (H_0) au profit de (H_1) si $y = 0$, et de rejeter (H_0) au profit de (H_1) si $y = 1$.

La loi de Y est définie par sa loi conditionnelle sachant $X = x$ (qui ne dépend pas de θ), i.e.

$$\mathcal{L}(Y|X = x) = \mathcal{B}(\phi(x)).$$

On a donc :

$$E[Y|X = x] = \phi(x) \quad \text{i.e. } E[Y|X] = \phi(X),$$

et

$$E[Y] = E[E[Y|X]] = E[\phi(X)],$$

d'où

$$Y \sim \mathcal{B}(E[\phi(X)]).$$

Risque de première espèce, niveau et taille

Définition 18 (Risque de première espèce)**.** *Le* risque de première espèce *d'un test randomisé ϕ est l'application, notée α qui à chaque $\theta_0 \in \Theta_0$ associe la probabilité P_{θ_0} de rejeter (H_0) avec ϕ :*

$$\begin{aligned} \alpha \ : \Theta_0 &\to [0,1] \\ \theta_0 &\mapsto \alpha(\theta_0) = E_{\theta_0}[\phi(X)]. \end{aligned}$$

Définition 19 (Risque de première espèce maximal)**.** *Le* risque de première espèce maximal *d'un test randomisé ϕ est donné par* $\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0)$.

Définition 20 (Test randomisé de niveau/taille α_0)**.** *Soit $\alpha_0 \in [0,1]$.*
Un test randomisé ϕ est de niveau α_0 *si son risque de première espèce maximal est inférieur ou égal à α_0 i.e. :*

$$\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0) \leq \alpha_0 \quad (\textit{inéquation du niveau}).$$

Un test ϕ est de taille α_0 *si son risque de première espèce maximal est égal à α_0 i.e. :*

$$\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0) = \alpha_0 \quad (\textit{équation de la taille}).$$

Risque de deuxième espèce, puissance

Définition 21 (Risque de deuxième espèce)**.** *Le* risque de deuxième espèce *d'un test randomisé ϕ est l'application, notée β qui à chaque $\theta_1 \in \Theta_1$ associe la probabilité P_{θ_1} de ne pas rejeter (H_0) au profit de (H_1) avec ϕ :*

$$\begin{aligned} \beta \ : \Theta_1 &\to [0,1] \\ \theta_1 &\mapsto \beta(\theta_1) = E_{\theta_1}[1 - \phi(X)]. \end{aligned}$$

Définition 22 (Fonction puissance)**.** *La* fonction puissance *d'un test randomisé ϕ est l'application, notée γ, qui à chaque $\theta_1 \in \Theta_1$ associe la probabilité P_{θ_1} de rejeter (H_0) au profit de (H_1) avec ϕ :*

$$\begin{aligned} \gamma \ : \Theta_1 &\to [0,1] \\ \theta_1 &\mapsto 1 - \beta(\theta_1) = E_{\theta_1}[\phi(X)]. \end{aligned}$$

4.2 Tests uniformément plus puissants (UPP), tests sans biais

Définition 23 (Test UPP)**.** *Soit ϕ et ψ deux tests de niveau α, dont les fonctions puissance sont respectivement notées γ_ϕ et γ_ψ. Le test ϕ est* uniformément plus puissant (UPP, en anglais UMP) *que le test ψ si pour tout $\theta_1 \in \Theta_1$,*

$$\gamma_\phi(\theta_1) \geq \gamma_\psi(\theta_1).$$

Définition 24 (Test UPP(α))**.** *Un test est dit* uniformément plus puissant parmi les tests de niveau α, noté UPP(α), *s'il est de niveau α et s'il est uniformément plus puissant que tout test de niveau α.*

On montre dans la proposition suivante qu'un test UPP(α) est nécessairement de taille α.

Proposition 2. *Soit $0 < \alpha \leq 1$ et soit ϕ^* un test de taille $\alpha^* < \alpha$. Alors, il existe un test ϕ de taille α uniformément plus puissant que ϕ^*.*

Preuve. On pose

$$\phi(x) = \phi^*(x) + \frac{\alpha - \alpha^*}{1 - \alpha^*}(1 - \phi^*(x)). \tag{4.1}$$

On a $0 \leq \frac{\alpha-\alpha^*}{1-\alpha^*} \leq 1$, donc $\phi(x) \in [0,1]$ pour tout $x \in \mathcal{X}$: ϕ définit bien un test, et pour tout $\theta_0 \in \Theta_0$, $E_0[\phi] \leq \left(1 - \frac{\alpha-\alpha^*}{1-\alpha^*}\right) E_{\theta_0}[\phi^*] + \frac{\alpha-\alpha^*}{1-\alpha^*} \leq \left(1 - \frac{\alpha-\alpha^*}{1-\alpha^*}\right)\alpha^* + \frac{\alpha-\alpha^*}{1-\alpha^*} \leq \alpha$. Par ailleurs, étant donné $\varepsilon > 0$, puisque ϕ^* est de taille α^*, il existe $\theta_{0,\varepsilon} \in \Theta_0$ tel que $\alpha^* - \varepsilon \leq E_{0,\varepsilon}[\phi^*] \leq \alpha^*$. Par (4.1), on a alors $\alpha^* - \varepsilon + \frac{\alpha-\alpha^*}{1-\alpha^*}(1-\alpha^*) \leq E_{0,\varepsilon}[\phi]$, d'où $\alpha - \varepsilon \leq E_{0,\varepsilon}[\phi]$. Le test ϕ est donc de taille α. De plus, comme $\phi^*(x) \leq \phi(x)$ pour tout x, pour tout $\theta_1 \in \Theta_1$,

$$E_{\theta_1}[\phi^*] \leq E_{\theta_1}[\phi].$$

Par conséquent, ϕ est UPP que ϕ^*.

Remarque. Sauf dans quelques cas particuliers (que l'on étudiera dans la suite du cours), il n'existe pas en général de tests UPP(α). On exigera alors d'un test qu'il soit au minimum sans biais.

Définition 25 (Test sans biais)**.** *Un test ϕ de taille α, de puissance γ_ϕ, est* sans biais *si pour tout $\theta_1 \in \Theta_1$,*

$$\gamma_\phi(\theta_1) \geq \alpha.$$

Remarque. Un *test constant* ϕ de taille α est défini par $\phi(x) = \alpha$, pour tout $x \in \mathcal{X}$. En d'autres termes, quelle que soit l'observation $x \in \mathcal{X}$, le test ϕ conclut au rejet de (H_0) avec une probabilité α. Un test sans biais est donc un test UPP que le test constant - qui ne tient pas compte de l'observation.

Proposition 3. *Un test UPP(α) est sans biais.*

4.3 Tests d'hypothèses simples - Lemme fondamental de Neyman-Pearson

On considère ici un modèle statistique $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ dominé par une mesure μ, et on note $L(x, \theta)$ sa vraisemblance.

On étudie le problème de test de

$$(H_0) \ : \ \theta = \theta_0 \text{ contre } (H_1) \ : \ \theta = \theta_1,$$

avec $\theta_0, \theta_1 \in \Theta$.
Contrairement au cadre général, les puissances de deux tests différents sont dans ce cadre toujours comparables. On peut donc espérer pouvoir trouver un test UPP(α).
Ce problème de test d'une hypothèse simple contre une hypothèse simple peut paraître simpliste et irréaliste, et c'est très souvent le cas. Cependant, l'étude de ce problème s'étendra à des cas plus complexes - et plus réalistes.

Définition 26 (Test de Neyman-Pearson)**.** *Un* test de Neyman-Pearson *est un test de la forme :*

$$\phi(x) = \begin{cases} 1 & si \quad L(x,\theta_1) > kL(x,\theta_0) \\ c(x) \in [0,1] & si \quad L(x,\theta_1) = kL(x,\theta_0) \\ 0 & si \quad L(x,\theta_1) < kL(x,\theta_0) \end{cases}$$

Remarques.
- Si $P_{\theta_0}(\{x, L(x,\theta_1) = kL(x,\theta_0)\}) = 0$, $c = 0$ ou $c = 1$, le test de Neyman-Pearson est un test non randomisé.
- On ne considère dans la suite que des tests de Neyman-Pearson avec c constante.

Proposition 4 (Existence)**.** *Pour tout $\alpha \in]0,1[$, il existe un test de Neyman-Pearson de taille α avec c constante.*

Preuve. Soit ϕ un test de Neyman-Pearson avec $k \geq 0$ et c constante. ϕ est de taille α si et seulement si

$$E_{\theta_0}[\phi(X)] = P_{\theta_0}(\{x, L(x,\theta_1) > kL(x,\theta_0)\}) + cP_{\theta_0}(\{x, L(x,\theta_1) = kL(x,\theta_0)\}) = \alpha.$$

Puisque $P_{\theta_0}(\{x, L(x,\theta_0) = 0\}) = 0$,

$$E_{\theta_0}[\phi(X)] = P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} > k\right\}\right) + cP_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} = k\right\}\right).$$

La fonction $t \mapsto P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} > t\right\}\right)$ étant le complément à 1 d'une fonction de répartition, est décroissante de 1 à 0. Par conséquent, on peut prendre

$$k = \inf\left\{t > 0, P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} > t\right\}\right) \leq \alpha\right\}.$$

On a alors

$$P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} > k\right\}\right) \leq \alpha \leq P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} \geq k\right\}\right).$$

Ainsi, si $P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} = k\right\}\right) = 0$, en prenant $c = 0$, on obtient bien un test de taille α. Sinon, on prend

$$c = \frac{\alpha - P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} > k\right\}\right)}{P_{\theta_0}\left(\left\{x, \frac{L(x,\theta_1)}{L(x,\theta_0)} = k\right\}\right)}.$$

Remarque. Si $\alpha = 1$, le test constant $\phi = 1$ est de taille α et correspond à un test de Neyman-Pearson avec $k = 0$, et $c = 1$. Si $\alpha = 0$, le test $\phi(x) = \mathbb{1}_{L(x,\theta_0)=0}$ est un test de Neyman-Pearson de taille α avec $k = +\infty$ et $c = 1$, à condition de prendre comme convention $(+\infty) \times 0 = 0$.
Le théorème suivant donne une condition nécessaire et suffisante pour qu'un test soit UPP(α) pour le problème de test de

$$(H_0) \ : \ \theta = \theta_0 \text{ contre } (H_1) \ : \ \theta = \theta_1$$

considéré.

Thèorème 2 (Lemme fondamental de Neyman-Pearson)**.** *Soit $\alpha \in]0, 1[$.*
(i) Un test de Neyman-Pearson de taille α est UPP(α).
(ii) Réciproquement, un test UPP(α) est un test de Neyman-Pearson de taille α.

Preuve.
(*i*) Soit ϕ un test de Neyman-Pearson avec $k \geq 0$ et c constante, de taille α, et ϕ^* un test de niveau $\alpha : E_{\theta_0}[\phi^*] \leq \alpha$.
On a $\phi(x) = 1 \geq \phi^*(x)$ si $L(x, \theta_1) > kL(x, \theta_0)$, et $\phi(x) = 0 \leq \phi^*(x)$ si $L(x, \theta_1) < kL(x, \theta_0)$. Donc pour tout $x \in \mathcal{X}$, $(\phi(x) - \phi^*(x))(L(x, \theta_1) - kL(x, \theta_0)) \geq 0$, et $\int_{\mathcal{X}}(\phi(x) - \phi^*(x))(L(x, \theta_1) - kL(x, \theta_0))d\mu(x) \geq 0$. D'où

$$\begin{aligned} E_{\theta_1}[\phi(X)] - E_{\theta_1}[\phi^*] &\geq k(E_{\theta_0}[\phi] - E_{\theta_0}[\phi^*]) && (4.2)\\ &\geq k(\alpha - E_{\theta_0}[\phi^*]) && (4.3)\\ &\geq 0. && (4.4)\end{aligned}$$

En conclusion, $E_{\theta_1}[\phi(X)] \geq E_{\theta_1}[\phi^*(X)]$, ce qui signifie que ϕ est UPP que ϕ^*.
(*ii*) D'après la proposition 4, il existe un test de Neyman-Pearson ϕ de taille α. Soit par ailleurs ϕ^* un test de niveau α, UPP(α). D'après (*i*), ϕ est UPP(α). Par conséquent, $\mathbb{E}_{\theta_1}[\phi(X)] = E_{\theta_1}[\phi^*(X)]$. On a vu par ailleurs dans la preuve de (*i*) que $(\phi(x) - \phi^*(x))(L(x, \theta_1) - kL(x, \theta_0)) \geq 0$. D'où

$$\begin{aligned} \int_{\mathcal{X}}(\phi(x) - \phi^*(x))(L(x, \theta_1) - kL(x, \theta_0))d\mu(x) &= \mathbb{E}_{\theta_1}[\phi] - E_{\theta_1}[\phi^*] + k(E_{\theta_0}[\phi^*] - E_{\theta_0}[\phi]) && (4.5)\\ &= k(E_{\theta_0}[\phi^*] - E_{\theta_0}[\phi]) && (4.6)\\ &= k(E_{\theta_0}[\phi^*] - \alpha) && (4.7)\\ &\geq 0. && (4.8)\end{aligned}$$

On a donc $E_{\theta_0}[\phi^*(X)] \geq \alpha$. Comme ϕ^* est de niveau α, ϕ^* est de taille α et

$$\int_{\mathcal{X}}(\phi(x) - \phi^*(x))(L(x, \theta_1) - kL(x, \theta_0))d\mu(x) = 0,$$

ce qui veut dire que μ-p.p. $(\phi(x) - \phi^*(x))(L(x, \theta_1) - kL(x, \theta_0)) = 0$ ou que $\phi(x) = \phi^*(x)$ pour μ-presque tout x tel que $L(x, \theta_1) \neq kL(x, \theta_0)$. On conclut que ϕ^* est un test de Neyman-Pearson.

Remarque. Si T est une statistique exhaustive pour le modèle, d'après le théorème de factorisation, on a $L(x, \theta_0) = h_{\theta_0}(T(x))h(x)$ et $L(x, \theta_1) = h_{\theta_1}(T(x))h(x)$, donc un test de Neyman-Pearson dans ce cadre ne dépend des observations qu'au travers de $T(x)$.

Tests de Neyman-Pearson pour des modèles à rapport de vraisemblance monotone

Dans des cas où $\Theta \subset \mathbb{R}$, les familles de lois à rapport de vraisemblance monotone permettent de simplifier l'écriture des tests de Neyman-Pearson, et de se ramener souvent à un test qu'on aurait pu construire de façon complètement intuitive.

Définition 27 (Modèle à rapport de vraisemblance monotone)**.** *Le modèle statistique* $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ *avec* $\Theta \subset \mathbb{R}$ *est dit* à rapport de vraisemblance (strictement) monotone en une statistique T *si pour* $\theta < \theta'$*, il existe une fonction* $h_{\theta,\theta'} : \mathbb{R} \to \bar{\mathbb{R}}$*, (strictement) monotone, telle que*

$$\frac{L(x,\theta')}{L(x,\theta)} = h_{\theta,\theta'}(T(x)).$$

Il est dit à rapport de vraisemblance croissant, strictement croissant, décroissant, strictement décroissant en $T(x)$ *si* $h_{\theta,\theta'}$ *l'est.*

Remarque. Un modèle à rapport de vraisemblance (strictement) croissant est aussi un modèle à rapport de vraisemblance (strictement) décroissant (il suffit de considérer la statistique $-T$ et la fonction $x \mapsto h_{\theta,\theta'}(-x)$) et vice versa.

Dans la suite, on ne considérera donc que des modèles à rapport de vraisemblance croissant.

Exemple fondamental : modèles exponentiels de statistique privilégiée T.

Proposition 5. *Si le modèle* $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ *avec* $\Theta \subset \mathbb{R}$ *est à rapport de vraisemblance croissant en* $T(x)$*, le test défini de la façon suivante est un test de Neyman-Pearson.*
Si $\theta_0 < \theta_1$*,*

$$\phi(x) = \begin{cases} 1 & si \quad T(x) > k \\ c & si \quad T(x) = k \\ 0 & si \quad T(x) < k. \end{cases}$$

Si $\theta_0 > \theta_1$*,*

$$\phi(x) = \begin{cases} 1 & si \quad T(x) < k \\ c & si \quad T(x) = k \\ 0 & si \quad T(x) > k. \end{cases}$$

4.4 Tests d'hypothèses composites

4.4.1 Extension du Lemme de Neyman-Pearson

Si le lemme fondamental de Neyman-Pearson s'applique à un problème de test d'hypothèses simples, ses conséquences s'étendent en fait à certains problèmes de tests d'hypothèses composites.

Proposition 6. *Soit* $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ *un modèle statistique dominé par une mesure* μ*, de vraisemblance notée* $L(x,\theta)$*. Soit* $\Theta_0 \subset \Theta$ *et* $\Theta_1 \subset \Theta$*, avec* $\Theta_0 \cap \Theta_1 = \emptyset$*. Soit* ϕ *un test de niveau* α *de* (H_0) $\theta \in \Theta_0$ *contre* (H_1) $\theta \in \Theta_1$*. S'il existe* $\theta_0 \in \Theta_0$ *tel que* $E_{\theta_0}[\phi] = \alpha$ *et si, pour tout* $\theta_1 \in \Theta_1$*, il existe un nombre k positif vérifiant :*
- $\phi(x) = 1$ *si* $L(x,\theta_1) > kL(x,\theta_0)$*,*
- $\phi(x) = 0$ *si* $L(x,\theta_1) < kL(x,\theta_0)$*.*
Alors ϕ *est UPP*(α)*.*

Preuve. Soit ϕ^* un test de niveau α de $(H_0)\ \theta \in \Theta_0$ contre $(H_1)\ \theta \in \Theta_1$. Soit $\theta_1 \in \Theta_1$. On considère le problème de test de

$$(H_0')\ :\ \theta = \theta_0 \text{ contre } (H_1')\ :\ \theta = \theta_1.$$

Pour ce problème, ϕ est un test de Neyman-Pearson de taille α. Il est donc UPP(α). On a par ailleurs, $E_{\theta_0}[\phi^*(X)] \leq \sup_{\theta\in\Theta_0} E_\theta[\phi^*(X)] \leq \alpha$ d'où $E_{\theta_1}[\phi(X)] \geq E_{\theta_1}[\phi^*(X)]$.

4.4.2 Tests unilatères

On considère dans les paragraphes suivants un modèle $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$, avec $\Theta \subset \mathbb{R}$, à rapport de vraisemblance monotone.

Tests unilatères de $(H_0)\ :\ \theta = \theta_0$ **contre** $(H_1)\ :\ \theta > \theta_0$.

Thèorème 3. *Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de taille α UPP(α) de la forme :*

$$\phi(x) = \begin{cases} 1 & si \quad T(x) > k \\ c & si \quad T(x) = k \\ 0 & si \quad T(x) < k. \end{cases}$$

Preuve. Soit $\theta' > \theta_0$. On considère le test de Neyman-Pearson de taille α de

$$(H_0)\ :\ \theta = \theta_0 \text{ contre } (H_1')\ :\ \theta = \theta'.$$

On a vu que ce test s'écrit sous la forme

$$\phi(x) = \begin{cases} 1 & \text{si} \quad T(x) > k \\ c & \text{si} \quad T(x) = k \\ 0 & \text{si} \quad T(x) < k. \end{cases}$$

Soit maintenant $\theta_1 > \theta_0$, puisque $\frac{L(x,\theta_1)}{L(x,\theta_0)} = h_{\theta_0,\theta_1}(T(x))$ avec h_{θ_0,θ_1} strictement croissante, il existe k_1 tel que $T(x) > k$ équivaut à $L(x,\theta_1) > k_1 L(x,\theta_0)$ et $T(x) < k$ équivaut à $L(x,\theta_1) < k_1 L(x,\theta_0)$. D'après la proposition précédente, on a donc que ϕ est UPP(α) de $(H_0)\ :\ \theta = \theta_0$ contre $(H_1)\ :\ \theta > \theta_0$.

Tests unilatères de $(H_0)\ :\ \theta = \theta_0$ **contre** $(H_1)\ :\ \theta < \theta_0$.

De la même façon, on a le théorème suivant.

Thèorème 4. *Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de taille α UPP(α) de la forme :*

$$\phi(x) = \begin{cases} 1 & si \quad T(x) < k \\ c & si \quad T(x) = k \\ 0 & si \quad T(x) > k. \end{cases}$$

Tests unilatères de $(H_0) : \theta \leq \theta_0$ **contre** $(H_1) : \theta > \theta_0$.

Thèorème 5 (Théorème de Lehmann)**.** *Soit* $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ *un modèle statistique à rapport de vraisemblance strictement croissant en* $T(x)$*. Pour tout* $\alpha \in]0,1[$*, il existe un test de taille* α *UPP(*α*) de la forme :*

$$\phi(x) = \begin{cases} 1 & si \quad T(x) > k \\ c & si \quad T(x) = k \\ 0 & si \quad T(x) < k. \end{cases}$$

De plus, la taille α *de* ϕ *est atteinte pour* $\theta = \theta_0$ *i.e.* $\sup_{\theta\leq\theta_0} E_\theta[\phi(X)] = E_{\theta_0}[\phi(X)] = \alpha$.

Preuve. On sait qu'il existe un test ϕ tel que $E_{\theta_0}[\phi(X)] = \alpha$ de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad T(x) < k \\ c & \text{si} \quad T(x) = k \\ 0 & \text{si} \quad T(x) > k. \end{cases}$$

Soit $\theta' < \theta''$. Puisque $\frac{L(x,\theta'')}{L(x,\theta')}$ est une fonction strictement croissante de $T(x)$, ϕ est un test de Neyman-Pearson pour le problème de test de $(H_0') : \theta = \theta'$ contre $(H_1') : \theta = \theta''$. D'après le lemme fondamental de Neyman-Pearson, pour tout test ϕ^* tel que $E_{\theta'}[\phi^*(X)] \leq E_{\theta'}[\phi(X)]$, $E_{\theta''}[\phi(X)] \geq E_{\theta''}[\phi^*(X)]$.
Soit $\theta' = \theta_0$, $\theta'' = \theta > \theta_0$, et ϕ^* un test de niveau α pour le problème de test inital de $(H_0) : \theta \leq \theta_0$ contre $(H_1) : \theta > \theta_0$. Alors $E_{\theta_0}[\phi^*(X)] \leq \sup_{\theta\in\Theta_0} E_\theta[\phi^*(X)] \leq \alpha = E_{\theta_0}[\phi(X)]$ donc $E_\theta[\phi^*(X)] \leq E_\theta[\phi(X)]$. Le test ϕ est UPP que ϕ^*.
Il reste à montrer que ϕ est de taille α pour le problème de test initial de $(H_0) : \theta \leq \theta_0$ contre $(H_1) : \theta > \theta_0$.
Soit maintenant $\theta' < \theta_0$, $\theta'' = \theta_0$ et ϕ^* le test constant égal à $E_{\theta'}[\phi(X)]$. Alors $E_{\theta'}[\phi^*(X)] = E_{\theta'}[\phi(X)]$ donc $E_{\theta_0}[\phi(X)] \geq E_{\theta_0}[\phi^*(X)]$. c'est-à-dire $E_{\theta_0}[\phi(X)] \geq E_{\theta'}[\phi(X)]$. On a donc pour tout $\theta' < \theta_0$ $E_{\theta'}[\phi(X)] \leq E_{\theta_0}[\phi(X)] = \alpha$, et $\sup_{\theta\leq\theta_0} E_\theta[\phi(X)] = E_{\theta_0}[\phi(X)] = \alpha$.

Remarque. Dans cette preuve, on a montré que ϕ est de Neyman-Pearson pour tout problème de test de $(H_0') : \theta = \theta'$ contre $(H_1') : \theta = \theta''$, avec $\theta' < \theta''$, donc il est sans biais et $E_{\theta'}[\phi(X)] < E_{\theta''}[\phi(X)]$ sauf si $P_{\theta'} = P_{\theta''}$.
En effet, si $E_{\theta'}[\phi(X)] = E_{\theta''}[\phi(X)]$ et si ϕ^* est le test constant de taille $E_{\theta'}[\phi(X)]$, ϕ^* a la même puissance que ϕ donc il est UPP parmi les tests de niveau $E_{\theta'}[\phi(X)]$. Il est donc de Neyman-Pearson et $L(x,\theta') = kL(x,\theta'')$ μ p.p. Comme L est une vraisemblance, $k = 1$ d'où $P_{\theta'} = P_{\theta''}$.
On en déduit que si le modèle est identifiable, la fonction $\theta \mapsto E_\theta[\phi(X)]$ est strictement croissante.

4.4.3 Tests bilatères

On construit ici des tests bilatères "optimaux" (en un certain sens) dans le cadre de modèles exponentiels à rapport de vraisemblance strictement monotone.

Thèorème 6 (Lemme de Neyman-Pearson généralisé (admis))**.** *Soit* $P_1, P_2, \ldots, P_{m+1}$ *des probabilités sur* $(\mathcal{X}, \mathcal{A})$*. Il existe une mesure* σ*-finie* μ *telle que* $dP_i = f_i d\mu$ *(par exemple,* $\mu = \sum_{i=1}^{m+1} P_i$*). On note* $E_i[\phi(X)] = \int \phi(x) f_i(x) d\mu(x)$*. On note* $\mathcal{C}_m$ *l'ensemble des tests* ϕ *vérifiant les contraintes*

$$E_1[\phi(X)] = c_1, E_2[\phi(X)] = c_2, \ldots, E_m[\phi(X)] = c_m,$$

pour des réels $c_1, c_2, \ldots, c_m$ fixés.
(i) Il existe un test ϕ dans C_m maximisant $E_{m+1}[\phi(X)]$.
(ii) Tout élément ϕ de C_m de la forme

$$\phi(x) = \begin{cases} 1 & si \quad f_{m+1}(x) > \sum_{i=1}^{m} k_i f_i(x) \\ 0 & si \quad f_{m+1}(x) < \sum_{i=1}^{m} k_i f_i(x) \end{cases}$$

maximise $E_{m+1}[\phi(X)]$.
(iii) Tout élément ϕ de C_m de la forme

$$\phi(x) = \begin{cases} 1 & si \quad f_{m+1}(x) > \sum_{i=1}^{m} k_i f_i(x) \\ 0 & si \quad f_{m+1}(x) < \sum_{i=1}^{m} k_i f_i(x) \end{cases}$$

avec $k_1 \geq 0, \ldots, k_m \geq 0$ maximise $E_{m+1}[\phi(X)]$ parmi l'ensemble de tous les tests tels que

$$E_1[\phi(X)] \leq c_1, E_2[\phi(X)] \leq c_2, \ldots, E_m[\phi(X)] \leq c_m.$$

(iv) L'ensemble $C_m = \{(E_1[\phi(X)], \ldots, E_m[\phi(X)]),\ \phi : (\mathcal{X}, \mathcal{A}) \to ([0,1], \mathcal{B}([0,1]))\}$ est convexe fermé. Si $(c_1, \ldots, c_m)$ est un point intérieur de C_m, il existe un test de Neyman-Pearson généralisé dans C_m et tout test de C_m maximisant $E_{m+1}[\phi(X)]$ est un test de Neyman-Pearson généralisé.

Preuve. On trouvera la preuve de ce théorème dans [15].

Tests bilatères de $(H_0)\ :\ \theta \leq \theta_1$ ou $\theta \geq \theta_2$ contre $(H_1)\ :\ \theta \in]\theta_1, \theta_2[$.

Thèorème 7. *Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta \in \Theta})$ un modèle exponentiel général dominé par une mesure μ, dont la vraisemblance est donnée par :*

$$L(x, \theta) = C(\theta) h(x) e^{\eta(\theta) T(x)}.$$

On suppose que η est strictement croissante, de telle façon que le modèle considéré soit à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de taille α UPP(α) de la forme :

$$\phi(x) = \begin{cases} 1 & si \quad k_1 < T(x) < k_2 \\ c_1 \text{ ou } c_2 & si \quad T(x) = k_1 \text{ ou } k_2 \\ 0 & si \quad T(x) < k_1 \text{ ou } T(x) > k_2. \end{cases}$$

De plus, la taille α de ϕ est atteinte pour $\theta = \theta_1$ et $\theta = \theta_2$ i.e. $\sup_{\theta \leq \theta_1} E_\theta[\phi(X)] = E_{\theta_1}[\phi(X)] = \alpha$ et $\sup_{\theta \geq \theta_2} E_\theta[\phi(X)] = E_{\theta_2}[\phi(X)] = \alpha$.

Preuve.
L'ensemble $C_2 = \{(E_{\theta_1}[\phi], E_{\theta_2}[\phi]),\ \phi : (\mathcal{X}, \mathcal{A}) \to ([0,1], \mathcal{B}([0,1]))\}$ (appelé diagramme des puissances) est convexe et contient la diagonale de $[0,1]$.
Pour tout test ϕ de Neyman-Pearson de taille α pour le problème de test de $(H_0)\ :\ \theta = \theta_1$ contre $(H_1)\ :\ \theta = \theta_2$, la puissance $E_{\theta_2}[\phi(X)]$ est strictement supérieure à α sauf si $P_{\theta_1} = P_{\theta_2}$ (voir remarque qui suit le théorème de Lehmann) ce qui est exclu car η est supposée strictement croissante. On peut faire le même raisonnement pour le problème de test de $(H_0)\ :\ \theta = \theta_2$ contre $(H_1)\ :\ \theta = \theta_1$ ce qui permet de conclure avec la convexité de C_2 que pour tout α, le point (α, α) est intérieur à C_2.

Soit $\theta_1 < \theta' < \theta_2$. D'après le lemme de Neyman-Pearson généralisé, tout test qui maximise $E_{\theta'}[\phi(X)]$ sous les contraintes $E_{\theta_1}[\phi(X)] = E_{\theta_2}[\phi(X)] = \alpha$ est de la forme :

$$\phi_\alpha(x) = \begin{cases} 1 & \text{si} \quad L(x,\theta') > k_1 L(x,\theta_1) + k_2 L(x,\theta_2) \\ 0 & \text{si} \quad L(x,\theta') < k_1 L(x,\theta_1) + k_2 L(x,\theta_2) \end{cases},$$

ou encore, au vu de l'expression de la vraisemblance,

$$\phi_\alpha(x) = \begin{cases} 1 & \text{si} \quad a_1 e^{b_1 T(x)} + a_2 e^{b_2 T(x)} < 1 \\ 0 & \text{si} \quad a_1 e^{b_1 T(x)} + a_2 e^{b_2 T(x)} > 1 \end{cases},$$

avec $a_1 = k_1 \frac{C(\theta_1)}{C(\theta')}$, $a_2 = k_2 \frac{C(\theta_2)}{C(\theta')}$, $b_1 = \eta(\theta_1) - \eta(\theta') < 0$, $b_2 = \eta(\theta_2) - \eta(\theta') > 0$.
On ne peut avoir $a_1 < 0$ et $a_2 < 0$, sinon, k_1 et k_2 seraient tous les deux négatifs, et $\phi_\alpha \equiv 1$, ce qui est impossible puisque $E_{\theta_1}[\phi(X)] = \alpha$ et que $\alpha < 1$.
On ne peut pas avoir $a_1 > 0$ et $a_2 < 0$, ni l'un des deux égal à 0 car sinon $a_1 e^{b_1 T(x)} + a_2 e^{b_2 T(x)}$ serait strictement monotone en $T(x)$, donc la fonction $\theta \mapsto E_\theta[\phi_\alpha(X)]$ serait elle aussi strictement monotone (voir raisonnement de la remarque qui suit le théorème de Lehmann), ce qui contredirait $E_{\theta_1}[\phi_\alpha(X)] = E_{\theta_2}[\phi_\alpha(X)] = \alpha$.
On a donc $a_1 > 0$ et $a_2 > 0$. La fonction g définie par $g(y) = a_1 e^{b_1 y} + a_2 e^{b_2 y}$ est alors strictement décroissante sur $]-\infty, y_0]$ puis strictement croissante sur $[y_0, +\infty[$ avec $y_0 = \frac{1}{b_2 - b_1} \ln\left(\frac{-a_1 b_1}{a_2 b_2}\right)$.
On a donc

$$\phi_\alpha(x) = \begin{cases} 1 & \text{si} \quad k_1' < T(x) < k_2' \\ 0 & \text{si} \quad T(x) < k_1' \text{ ou } T(x) > k_2', \end{cases}$$

ce qui veut dire qu'il est du type du test énoncé (à noter qu'on peut le prendre constant sur $\{x, T(x) = c_i\}$ puisque $T(x)$ est exhaustive).
Montrons maintenant que le test obtenu est bien de taille α.
Soit $\theta'' < \theta_1$ (le raisonnement est similaire pour $\theta'' > \theta_2$).
D'après le lemme de Neyman-Pearson généralisé, un test ϕ_α' tel que $E_{\theta_1}[\phi_\alpha'(X)] = E_{\theta_2}[\phi_\alpha'(X)] = \alpha$ et

$$\phi_\alpha'(x) = \begin{cases} 1 & \text{si} \quad L(x,\theta'') < k_1' L(x,\theta_1) + k_2' L(x,\theta_2) \\ 0 & \text{si} \quad L(x,\theta') > k_1' L(x,\theta_1) + k_2' L(x,\theta_2) \end{cases},$$

minimise $E_{\theta''}[\phi(X)]$ sous les contraintes $E_{\theta_1}[\phi(X)] = E_{\theta_2}[\phi(X)] = \alpha$ (appliquer le théorème à $1 - \phi_\alpha'$). On montre par le même raisonnement que précédemment que ϕ_α' est de la même forme que ϕ_α, et donc que ϕ_α minimise aussi $E_{\theta''}[\phi(X)]$ sous les contraintes $E_{\theta_1}[\phi(X)] = E_{\theta_2}[\phi(X)] = \alpha$. En considérant le test constant égal à α, on obtient que $E_{\theta''}[\phi(X)] \leq \alpha$, et ce, pour tout $\theta'' < \theta_1$, d'où $\sup_{\theta'' \leq \theta_1} E_{\theta''}[\phi(X)] = E_{\theta_1}[\phi(X)] = \alpha$.

Remarque. La difficulté pratique de mise en oeuvre de ce test réside dans la détermination des constantes k_1 et k_2 telles que $E_{\theta_1}[\phi(X)] = E_{\theta_2}[\phi(X)] = \alpha$.

Tests bilatères de $(H_0) : \theta \neq \theta_0$ **contre** $(H_1) : \theta = \theta_0$.

Ce problème est proche du problème précédent. On montre comme ci-dessus qu'il existe dans le cadre des modèles exponentiels généraux un test UPP(α), mais le théorème se trouve légèrement modifié dans l'écriture de l'équation de la taille.

Thèorème 8. *Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle exponentiel général dominé par une mesure μ, dont la vraisemblance est donnée par :*

$$L(x,\theta) = C(\theta)h(x)e^{\eta(\theta)T(x)}.$$

On suppose que η est strictement croissante, de telle façon que le modèle considéré soit à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de taille α UPP(α) de la forme :

$$\phi(x) = \begin{cases} 1 & si \quad k_1 < T(x) < k_2 \\ c_1 \text{ ou } c_2 & si \quad T(x) = k_1 \text{ ou } k_2 \\ 0 & si \quad T(x) < k_1 \text{ ou } T(x) > k_2. \end{cases}$$

De plus, la taille α de ϕ est atteinte pour $\theta = \theta_0$ et les constantes k_1 et k_2 sont déterminées par

$$\begin{cases} E_{\theta_0}[\phi(X)] = \alpha, \\ E_{\theta_0}[T(X)\phi(X)] = \alpha E_{\theta_0}[T(X)]. \end{cases}$$

Remarque. Le calcul des constantes est simplifié si la loi de $T(X)$ est symétrique par rapport à un nombre a lorsque $X \sim P_{\theta_0}$. Si on choisit un test $\phi = h(T)$, avec h symétrique par rapport à a (i.e. $(k_1 + k_2)/2 = a$ et $c_1 = c_2 = c$), et tel que $E_{\theta_0}[\phi(X)] = \alpha$, alors $E_{\theta_0}[T(X)\phi(X)] = E_{\theta_0}[(T(X) - a)h(T(X))] + aE_{\theta_0}[\phi(X)] = a\alpha = \alpha E_{\theta_0}[T(X)]$. La deuxième équation est vérifiée.

Tests bilatères de $(H_0) : \theta \in [\theta_1, \theta_2]$ contre $(H_1) : \theta < \theta_1$ ou $\theta > \theta_2$.

Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique à rapport de vraisemblance strictement croissant en $T(x)$. On montre qu'il n'existe pas de test UPP(α).
Si un tel test existait, on aurait en effet pour tout test ϕ^* tel que $\sup_{\theta\in[\theta_1,\theta_2]} E_\theta[\phi^*(X)] \leq \alpha$,

$$E_\theta[\phi(X)] \geq E_\theta[\phi^*(X)] \quad \text{pour } \theta > \theta_2 \text{ ou } \theta < \theta_1.$$

Alors ϕ serait aussi UPP(α) pour le problème de test de $(H_0) : \theta \in [\theta_1, \theta_2]$ contre $(H_1') : \theta < \theta_1$ ou contre $(H_1') : \theta > \theta_2$. D'après la remarque qui suit la preuve du théorème de Lehmann, on en déduirait que la fonction $\theta \mapsto E_\theta[\phi(X)]$ est strictement décroissante sur $\Theta \cap]-\infty, \theta_2]$, et strictement croissante sur $\Theta \cap [\theta_1, +\infty[$, ce qui est impossible.
On cherche alors pour ce problème de test des tests "optimaux" dans une classe de tests plus restreinte que celle des tests de niveau fixé.

Définition 28 (Test sans biais au niveau α)**.** *Un test ϕ de $(H_0) : \theta \in \Theta_0$ contre $(H_1) : \theta \in \Theta_1$ est dit* sans biais au niveau *α si pour tout $\theta_0 \in \Theta_0$,*

$$E_{\theta_0}[\phi(X)] \leq \alpha,$$

et si pour tout $\theta_1 \in \Theta_1$,

$$E_{\theta_1}[\phi(X)] \geq \alpha.$$

Définition 29 (Test UPPSB(α))**.** *Un test est dit* uniformément plus puissant parmi les tests sans biais au niveau *α, noté UPPSB(α), s'il est sans biais au niveau α et s'il uniformément plus puissant que tout test sans biais au niveau α.*

Thèorème 9. *Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle exponentiel général dominé par une mesure μ, dont la vraisemblance est donnée par :*

$$L(x,\theta) = C(\theta)h(x)e^{\eta(\theta)T(x)}.$$

On suppose que η est strictement croissante, de telle façon que le modèle considéré soit à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de taille α UPPSB(α) de la forme :

$$\phi(x) = \begin{cases} 1 & si \quad T(x) < k_1 \text{ ou } T(x) > k_2 \\ c_1 \text{ ou } c_2 & si \quad T(x) = k_1 \text{ ou } k_2 \\ 0 & si \quad k_1 < T(x) < k_2. \end{cases}$$

De plus, la taille α de ϕ est atteinte pour $\theta = \theta_1$ et $\theta = \theta_2$ i.e. $\sup_{\theta\in[\theta_1,\theta_2]} E_\theta[\phi(X)] = E_{\theta_1}[\phi(X)] = E_{\theta_2}[\phi(X)] = \alpha$.

Preuve. Il est connu (ou admis) que dans le cadre du modèle exponentiel considéré, la fonction $\theta \mapsto E_\theta[\phi(X)]$ est continue pour tout test ϕ. On a montré dans le théorème 7 que le test $1-\phi(x)$ défini ci-dessus existe et est UPP$(1-\alpha)$ pour le problème de test de (H_0) : $\theta \leq \theta_1$ ou $\theta \geq \theta_2$ contre (H_1) : $\theta \in]\theta_1, \theta_2[$, et que pour tout $\theta < \theta_1$ ou $\theta > \theta_2$, il minimise $E_\theta[\psi(X)]$ sous les contraintes $E_{\theta_1}[\psi(X)] = E_{\theta_2}[\psi(X)] = \alpha$ (voir preuve). Par conséquent, pour le problème de test posé, ϕ est UPP parmi les tests ψ tels que $E_{\theta_1}[\psi(X)] = E_{\theta_2}[\psi(X)] = \alpha$.
Soit ϕ^* un test sans biais au niveau α pour le problème de test posé.
Alors, $E_\theta[\phi^*(X)] \leq \alpha$ pour tout $\theta \in [\theta_1, \theta_2]$ et $E_\theta[\phi^*(X)] \geq \alpha$ pour tout $\theta < \theta_1$ ou $\theta > \theta_2$. Par continuité, on a alors $E_{\theta_1}[\phi^*(X)] = E_{\theta_2}[\phi^*(X)] = \alpha$. D'après ce qui précède, ϕ est donc UPP que ϕ^*.

Tests bilatères de (H_0) : $\theta = \theta_0$ **contre** (H_1) : $\theta \neq \theta_0$.

Ce problème est proche du problème précédent. On montre comme ci-dessus qu'il n'existe pas de test UPP(α). En revanche, pour l'existence d'un test UPPSB(α), on a un théorème qui est légèrement modifié.

Thèorème 10. *Soit $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle exponentiel général dominé par une mesure μ, dont la vraisemblance est donnée par :*

$$L(x,\theta) = C(\theta)h(x)e^{\eta(\theta)T(x)}.$$

On suppose que η est strictement croissante, de telle façon que le modèle considéré soit à rapport de vraisemblance strictement croissant en $T(x)$. Pour tout $\alpha \in]0,1[$, il existe un test de taille α UPPSB(α) de la forme :

$$\phi(x) = \begin{cases} 1 & si \quad T(x) < k_1 \text{ ou } T(x) > k_2 \\ c_1 \text{ ou } c_2 & si \quad T(x) = k_1 \text{ ou } k_2 \\ 0 & si \quad k_1 < T(x) < k_2. \end{cases}$$

Les constantes c_1, c_2, k_1, k_2 sont déterminées par

$$\begin{cases} E_{\theta_0}[\phi(X)] = \alpha, \\ E_{\theta_0}[T(X)\phi(X)] = \alpha E_{\theta_0}[T(X)]. \end{cases}$$

Remarque. Le calcul des constantes est là encore simplifié si la loi de $T(X)$ est symétrique par rapport à un nombre a lorsque $X \sim P_{\theta_0}$.

4.4.4 Tests avec paramètres de nuisance

Les théorèmes d'optimalité pour les problèmes de test d'hypothèses composites vus dans les paragraphes précédents sont valables pour des modèles paramétriques à un paramètre réel, à rapport de vraisemblance strictement monotone, voire seulement pour des modèles exponentiels à un paramètre réel, à rapport de vraisemblance strictement monotone. Des résultats similaires peuvent exister dans certains modèles à plusieurs paramètres réels $(\theta^*, \lambda_1, \dots, \lambda_{p-1})$, lorsque l'on fait un test d'hypothèses composites sur un seul de ces paramètres θ^*, ou lorsque l'on fait un test sur une forme linéaire de ces paramètres.
Soit Θ un ouvert convexe (non vide) de $\mathbb{R}^p$, et $(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta})$ un modèle exponentiel général dominé par une mesure μ, dont la vraisemblance est donnée par :

$$L(x, (\theta^*, \lambda)) = C(\theta)e^{\theta^* T(x) + \sum_{i=1}^{p-1} \lambda_i S_i(x)},$$

avec $\lambda = (\lambda_1, \dots, \lambda_{p-1})$. On note $S = (S_1, \dots, S_{p-1})$.
La statistique (T, S) est exhaustive et le modèle image par (T, S) est le modèle exponentiel $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), (P_\theta^{(T,S)})_{\theta\in\Theta})$, avec

$$dP_\theta^{(T,S)}(t, s) = C(\theta)e^{\theta^* t + \sum_{i=1}^{p-1} \lambda_i s_i} d\nu(t, s).$$

De plus, on sait que les lois conditionnelles $P_\theta^{T|S=s}$ forment une famille exponentielle qui ne dépend que de θ^* :

$$dP_\theta^{T|S=s}(y) = C_s(\theta^*)e^{\theta^* t} d\nu_s(t).$$

Pour s fixé, on se retrouve dans un modèle exponentiel ne dépendant que du paramètre θ^*. Dans ce modèle conditionnel, on sait trouver des tests UPP(α) ou UPPSB(α). On rapporte ensuite ces résultats dans le modèle initial. Par exemple, pour le problème de test de $(H_0) : \theta^* \leq \theta_0^*$ contre $(H_1) : \theta^* > \theta_0^*$, on a le théorème suivant :

Thèorème 11. *Dans le cadre du modèle décrit ci-dessus, pour le problème de test de* $(H_0) : \theta^* \leq \theta_0^*$ *contre* $(H_1) : \theta^* > \theta_0^*$, *il existe un test de taille* α *UPPSB(*α*) défini dans le modèle image par :*

$$\phi(t, s) = \begin{cases} 1 & si \quad t > k(s) \\ c(s) & si \quad t = k(s) \\ 0 & si \quad t < k(s), \end{cases}$$

avec

$$E_{\theta_0^*}[\phi(T(X), S(X))|S(X) = s] = \alpha \quad \forall s.$$

Dans certains cas, et notamment dans le cas gaussien, ce théorème peut se simplifier.

Proposition 7. *Dans le cadre du modèle décrit ci-dessus, on suppose qu'il existe une statistique réelle* $U = f(T, S)$ *libre dans* $\Theta_0^* = \{\theta, \theta^* = \theta_0^*\}$ *et telle que* $f(t, s)$ *soit croissante en t pour tout s. Pour le problème de test de* $(H_0) : \theta^* \leq \theta_0^*$ *contre* $(H_1) : \theta^* > \theta_0^*$, *il existe un test de taille* α *UPPSB(*α*) défini par :*

$$\phi(u) = \begin{cases} 1 & si \quad u > k \\ c & si \quad u = k \\ 0 & si \quad u < k, \end{cases}$$

avec

$$E_\theta[\phi(U(X))] = \alpha \quad \forall \theta \in \Theta_0^*.$$

Remarque. On peut parfois montrer en plus que ce test est UPP(α).
Pour le cas d'un test sur une forme linéaire, il suffit de transformer l'écriture du modèle pour se ramener au cas précédent.
Dans tous les cas, on se réfère pour plus de détails au livre d'A. Monfort. On répertorie ci-dessous les propriétés des tests paramétriques de base vus au début de ce cours dans les modèles gaussiens.

Application pour un échantillon gaussien

On considère un phénomène aléatoire modélisé par un n-échantillon $X = (X_1, \dots, X_n)$ d'une loi normale $\mathcal{N}(m, \sigma^2)$.
On se base sur une observation $x = (x_1, \dots, x_n)$ de cet échantillon.

Paramètres	Hypothèses	Statistiques / Forme des tests	Lois utilisées	Propriétés
σ^2 connue, $\theta = m \in \mathbb{R}$	$(H_0)\ m \leq m_0$ $(H_1)\ m > m_0$	$T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sigma}$ $\mathbb{1}_{T(x)>k}$	$\mathcal{N}(0,1)$	$UPP(\alpha)$
σ^2 connue, $\theta = m \in \mathbb{R}$	$(H_0)\ m = m_0$ $(H_1)\ m \neq m_0$	$T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sigma}$ $\mathbb{1}_{\lvert T(x)\rvert>k}$	$\mathcal{N}(0,1)$	$UPPSB(\alpha)$
σ^2 connue, $\theta = m \in \mathbb{R}$	$(H_0)\ m \leq m_1$ ou $m \geq m_2$ $(H_1)\ m \in]m_1, m_2[$	$T(x) = \bar{x}$ $\mathbb{1}_{k_1<T(x)<k_2}$	$\mathcal{N}\left(m_1, \frac{\sigma^2}{n}\right)$ $\mathcal{N}\left(m_2, \frac{\sigma^2}{n}\right)$	$UPP(\alpha)$
σ^2 connue, $\theta = m \in \mathbb{R}$	$(H_0)\ m \in [m_1, m_2]$ $(H_1)\ m < m_1$ ou $m > m_2$	$T(x) = \bar{x}$ $\mathbb{1}_{T(x)<k_1 \text{ ou } T(x)>k_2}$	$\mathcal{N}\left(m_1, \frac{\sigma^2}{n}\right)$ $\mathcal{N}\left(m_2, \frac{\sigma^2}{n}\right)$	$UPPSB(\alpha)$
$\theta = (m, \sigma^2)$ $\in \mathbb{R}\times\mathbb{R}_+^*$	$(H_0)\ m \leq m_0$ $(H_1)\ m > m_0$	$T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sqrt{S^2(x)}}$ avec $S^2(x) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ $\mathbb{1}_{T(x)>k}$	$\mathcal{T}(n-1)$	$UPP(\alpha)$
$\theta = (m, \sigma^2)$ $\in \mathbb{R}\times\mathbb{R}_+^*$	$(H_0)\ m = m_0$ $(H_1)\ m \neq m_0$	$T(x) = \sqrt{n}\frac{\bar{x}-m_0}{\sqrt{S^2(x)}}$ $\mathbb{1}_{\lvert T(x)\rvert>k}$	$\mathcal{T}(n-1)$	$UPPSB(\alpha)$
m connue, $\theta = \sigma^2 \in \mathbb{R}_+^*$	$(H_0)\ \sigma^2 \leq \sigma_0^2$ $(H_1)\ \sigma^2 > \sigma_0^2$	$T(x) = \sum_{i=1}^n \frac{(x_i-m)^2}{\sigma_0^2}$ $\mathbb{1}_{T(x)>k}$	$\chi^2(n)$	UPP(α)
m connue, $\theta = \sigma^2 \in \mathbb{R}_+^*$	$(H_0)\ \sigma^2 = \sigma_0^2$ $(H_1)\ \sigma^2 \neq \sigma_0^2$	$T(x) = \sum_{i=1}^n \frac{(x_i-m)^2}{\sigma_0^2}$ $\mathbb{1}_{T(x)<k_1 \text{ ou } T(x)>k_2}$	$\chi^2(n)$	UPPSB(α)
$\theta = (m, \sigma^2)$ $\in \mathbb{R}\times\mathbb{R}_+^*$	$(H_0)\ \sigma^2 \leq \sigma_0^2$ $(H_1)\ \sigma^2 > \sigma_0^2$	$T(x) = \sum_{i=1}^n \frac{(x_i-\bar{x})^2}{\sigma_0^2}$ $\mathbb{1}_{T(x)>k}$	$\chi^2(n-1)$	UPP(α)
$\theta = (m, \sigma^2)$ $\in \mathbb{R}\times\mathbb{R}_+^*$	$(H_0)\ \sigma^2 = \sigma_0^2$ $(H_1)\ \sigma^2 \neq \sigma_0^2$	$T(x) = \sum_{i=1}^n \frac{(x_i-\bar{x})^2}{\sigma_0^2}$ $\mathbb{1}_{T(x)<k_1 \text{ ou } T(x)>k_2}$	$\chi^2(n-1)$	UPPSB(α)

Application à la comparaison de deux échantillons gaussiens

On considère ici un phénomène aléatoire modélisé par deux échantillons indépendants : $(Y_1, \dots, Y_{n_1})$ n_1-échantillon de la loi $\mathcal{N}(m_1, \sigma_1^2)$ et $(Z_1, \dots, Z_{n_2})$ n_2-échantillon de la loi $\mathcal{N}(m_2, \sigma_2^2)$. On souhaite comparer, sur la base d'observations $(y_1, \dots, y_{n_1})$ et $(z_1, \dots, z_{n_2})$ de ces échantillons, m_1 et m_2, ou σ_1^2 et σ_2^2.

On ne décrit dans ce paragraphe que les tests bilatères. Les mêmes statistiques de test peuvent bien sûr être utilisées pour construire des tests unilatères. Dans ce cas, les tests seront tous UPP(α).

Paramètres	Hypothèses	Statistiques / Forme des tests	Lois utilisées	Propriétés
m_1, m_2 connues	(H_0) $\sigma_1^2 = \sigma_2^2$ (H_1) $\sigma_1^2 \neq \sigma_2^2$	$T(y,z) = \frac{\tilde{S}^2(y)}{\tilde{S}^2(z)}$ avec $\tilde{S}^2(y) = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - m_1)^2$ $\tilde{S}^2(z) = \frac{1}{n_2}\sum_{i=1}^{n_2}(z_i - m_2)^2$ $\mathbb{1}_{T(y,z)<k_1 \text{ ou } T(y,z)>k_2}$	$\mathcal{F}(n_1, n_2)$	$UPPSB(\alpha)$
m_1, m_2 inconnues	(H_0) $\sigma_1^2 = \sigma_2^2$ (H_1) $\sigma_1^2 \neq \sigma_2^2$	$T(y,z) = \frac{S^2(y)}{S^2(z)}$ avec $S^2(y) = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i - \bar{y})^2$ $S^2(z) = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(z_i - \bar{z})^2$ $\mathbb{1}_{T(y,z)<k_1 \text{ ou } T(y,z)>k_2}$	$\mathcal{F}(n_1 - 1, n_2 - 1)$	$UPPSB(\alpha)$
σ_1^2, σ_2^2 connues	(H_0) $m_1 = m_2$ (H_1) $m_1 \neq m_2$	$T(y,z) = \frac{\bar{y}-\bar{z}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\mathbb{1}_{\lvert T(y,z)\rvert>k}$	$\mathcal{N}(0,1)$	UPPSB(α)
σ_1^2, σ_2^2 inconnues mais égales	(H_0) $m_1 = m_2$ (H_1) $m_1 \neq m_2$	$T(y,z) = \frac{\bar{y}-\bar{z}}{\sqrt{S^2(y,z)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ avec $S^2(y,z) = \frac{(n_1-1)S^2(y) + (n_2-1)S^2(z)}{n_1+n_2-2}$ $\mathbb{1}_{\lvert T(y,z)\rvert>k}$	$\mathcal{T}(n_1 + n_2 - 2)$	UPPSB(α)
σ_1^2, σ_2^2 inconnues	(H_0) $m_1 = m_2$ (H_1) $m_1 \neq m_2$	$T(y,z) = \frac{\bar{y}-\bar{z}}{\sqrt{\frac{S^2(y)}{n_1} + \frac{S^2(z)}{n_2}}}$ $\mathbb{1}_{\lvert T(y,z)\rvert>k}$	$\approx \mathcal{N}(0,1)$ si n_1, n_2 grands ou approx.de Welch	

4.5 Exercices

Exercice 1 : Modèles à rapport de vraisemblance monotone

Montrer que les modèles statistiques suivants sont à rapport de vraisemblance monotone.
1. Le modèle de Bernoulli $\left(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}(\theta)^{\bigotimes n}\}_{\theta\in[0,1]}\right)$,
2. Le modèle gaussien $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathcal{N}(\theta,1)^{\otimes n}\}_{\theta\in\mathbb{R}})$,
3. Le modèle de Poisson $\left(\mathbb{N}^n, \mathcal{P}(\mathbb{N}^n), \{\mathcal{P}(\theta)^{\otimes n}\}_{\theta\in\mathbb{R}_+^*}\right)$,
4. Les modèles exponentiels généraux.

Exercice 2 : Résistance aux antibiotiques

Un laboratoire fabriquant un nouvel antibiotique contre une infection bactérienne résistante aux antibiotiques classiques affirme qu'il est efficace dans 90% des cas. Un expert prétend que la guérison n'intervient que dans 60% des cas. On décide de tester l'affirmation du laboratoire contre celle de l'expert : pour cela, on administre le médicament à 200 malades.
1. Construire un test UPP parmi les tests de niveau 1%.
2. Quelle est la taille de ce test ? Sa puissance ?
3. Ce test est-il sans biais ?
4. Sachant qu'après administration du médicament, 160 malades ont guéri, que peut-on conclure ?

Exercice 3 : Le 15

Les pouvoirs publics décident de réduire le nombre horaire de coups de téléphone au 15 qui ne sont pas justifiés. Ils envisagent d'adopter un dispositif de contrôle qui diminuerait de moitié ce nombre de communications. On admet que sans le dispositif, le nombre de communications non justifiées par poste de téléphone et par heure suit une loi de Poisson de paramètre $\lambda = 6$. On équipe 10 postes avec le dispositif. Soit $x_1, \dots, x_{10}$ les nombres de communications injustifiées pour chacun de ces postes.
1. Déterminer un test UPP parmi les tests de niveau $\alpha = 5\%$ de

$$(H_0) : \quad \lambda = 6 \quad \text{contre} \quad (H_1) : \quad \lambda = 3.$$

2. Calculer la puissance du test.
3. Quelle sera la conclusion du test si $\sum_{i=1}^{10} x_i = 51$?
4. On veut maintenant seulement tester la diminution du nombre d'appels injustifiés après mise en service du dispositif. Existe-t-il un test UPP parmi les tests de niveau $\alpha = 5\%$? Si oui, quelle est la conclusion de ce test ?

Exercice 4

Soit X une variable aléatoire réelle dont la loi a pour densité $f_\theta(x) = \frac{1}{2\theta\sqrt{x}} e^{-\frac{\sqrt{x}}{\theta}} \mathbb{1}_{\mathbb{R}_+}(x)$, avec $\theta > 0$, et $X_1, \dots, X_n$ un n-échantillon de cette loi.

1. Montrer que la variable aléatoire $2\sqrt{X}/\theta$ suit une loi du Khi-Deux à 2 degrés de liberté. En déduire la loi de $\frac{2}{\theta}\sum_{i=1}^{n}\sqrt{X_i}$.
2. On souhaite tester pour $0 < \theta_0 < \theta_1$, $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta = \theta_1$.
a) Etant donné $\alpha \in]0,1[$, déterminer un test UPP parmi les tests de niveau α.
b) Expliciter la puissance de ce test.
c) Décrire tous les tests UPP parmi les tests de niveau α.
3. On souhaite maintenant tester $(H_0) : \theta \leq \theta_0$ contre $(H_1) : \theta > \theta_0$. Existe-t-il un test UPP parmi les tests de niveau α pour ce nouveau problème de test ?

Exercice 5 : Présence d'un polluant

La limite du taux de présence d'un polluant contenu dans des déchets d'usine est 6mg/kg. On effectue un dosage sur 12 prélèvements de 1kg, pour lesquels on observe les taux x_i, $1 \leq i \leq 12$ de présence du polluant. On trouve $\sum_{i=1}^{12} x_i = 84$ et $\sum_{i=1}^{12} x_i^2 = 1413$. On admet que le taux de présence du polluant suit une loi normale $\mathcal{N}(m, \sigma^2)$ avec $\sigma = 8$.
1. Donner un test UPP parmi les tests de niveau 5% de $(H_0) : m \leq 6$ contre $(H_1) : m > 6$. Déterminer la fonction puissance de ce test. L'usine se conforme-t-elle à la législation ?
2. Envisager le cas où l'écart-type σ est inconnu.

Exercice 6

On dispose de l'observation $(x_1, \ldots, x_n)$ d'un échantillon de taille $n = 15$ d'une loi normale $\mathcal{N}(0, 1/\theta)$, $\theta > 0$.
1. Construire un test UPP parmi les tests de niveau $\alpha = 5\%$ de $(H_0) : \theta = 1$ contre $(H_1) : \theta > 1$.
2. Déterminer la puissance de ce test.
3. Quelle décision prend-on si $\sum_{i=1}^{n} x_i^2 = 6.8$? Pour quelles valeurs du niveau α prendrait-on la décision contraire ? Qu'a-t-on calculé alors ?
4. Existe-t-il un test UPP parmi les tests de niveau $\alpha = 5\%$ de $(H_0) : \theta = 1$ contre $(H_1) : \theta \neq 1$? Expliquer.

4.6 Problèmes corrigés

4.6.1 Campagne d'e-mailing

Problème

Un site de e-commerce généraliste lance une nouvelle campagne marketing par e-mail (campagne d'e-mailing). Soit θ la probabilité que le mail reçu par un client inscrit sur la liste de diffusion du site soit ouvert (taux d'ouverture). On considère X la variable aléatoire modélisant le nombre de mails à envoyer avant d'obtenir l'ouverture d'un de ces mails. En supposant les clients indépendants entre eux, X suit une loi géométrique de paramètre θ.
La campagne d'e-mailing est considérée comme inefficace si $\theta < 1/4$.
Sur la base de l'observation x de X, on souhaite tester :

$$(H_0) : \theta = 1/4 \quad \text{contre} \quad (H_1) : \theta < 1/4.$$

1. Décrire le modèle statistique considéré.
2. Montrer que ce modèle est à rapport de vraisemblance monotone.
3. Montrer que la fonction de répartition de la loi géométrique de paramètre θ est donnée par

$$F_\theta(x) = P(X \leq x) = 1 - (1-\theta)^x \text{ pour } x \in \mathbb{N}^*.$$

4. Construire de façon intuitive un test de (H_0) contre (H_1) de niveau 5%.
5. Expliquer pourquoi ce test ne peut pas être uniformément plus puissant parmi les tests de niveau 5%.
6. Construire un test uniformément plus puissant parmi les tests de niveau 5% de (H_0) contre (H_1).
7. Quelle est la conclusion de ce test si $x = 10$?
8. Existe-t-il un test uniformément plus puissant parmi les tests de niveau 5% des hypothèses composites

$$(H_0) : \theta \geq 1/4 \quad \text{contre} \quad (H_1) : \theta < 1/4 \text{ ?}$$

Justifier la réponse.

Correction

1. Modèle statistique considéré : soit X une variable aléatoire de loi géométrique de paramètre θ, avec $\theta \in \Theta =]0,1[$, modélisant le nombre de mails à envoyer pour que l'un de ces mails soit ouvert par un client. Soit x l'observation de X, $\mathcal{X} = \mathbb{N}^*$, et $\mathcal{A}$ l'ensemble des parties de $\mathcal{X}$. Pour $\theta \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{G}(\theta)$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in]0,1[})$.
2. Ce modèle est dominé par la mesure de comptage sur $\mathbb{N}^*$ et sa vraisemblance est donnée par $L(x,\theta) = \theta(1-\theta)^{x-1}$. Pour $\theta' > \theta$,

$$\frac{L(x,\theta')}{L(x,\theta)} = \frac{\theta'}{\theta}\left(\frac{1-\theta'}{1-\theta}\right)^{x-1} = h_{\theta,\theta'}(x),$$

où $h_{\theta,\theta'}$ est strictement décroissante. Le modèle est donc à rapport de vraisemblance strictement décroissant en $T(x) = x$.
3. On a pour $x = 1$, $F_\theta(x) = \theta$, et pour $x \geq 2$, $F_\theta(x) = \sum_{k=1}^{x-1} \theta(1-\theta)^{k-1} = \theta\frac{1-(1-\theta)^x}{1-(1-\theta)} = 1-(1-\theta)^x$.
4. L'espérance de X est égale à $1/\theta$, donc si l'on estime cette espérance par x, un test de (H_0) contre (H_1) a par exemple pour région critique :

$$\mathcal{R}_{(H_0)} = \{x \in \mathbb{N}^*, x > s\} = \{s+1, \ldots\}.$$

Calcul de la constante s : D'après la question 3, $P_{1/4}(\mathcal{R}_{(H_0)}) \leq 5\%$ si et seulement si $1 - F_{1/4}(s) \leq 5\%$ i.e. $(3/4)^s \leq 5\%$ ou encore $s \geq \ln(0.05)/\ln(3/4)$, d'où $s = 11$.
5. On a $P_{1/4}(\{x \in \mathbb{N}^*, x > 11\}) = 1 - F_{1/4}(11) = (3/4)^{11} = 0.0422$ donc le test construit à la question précédente n'est pas de taille 5%. En conséquence, il ne peut pas être UPP parmi les tests de niveau 5%.
6. Le modèle considéré étant un modèle à rapport de vraisemblance strictement décroissant en $T(x) = x$, d'après l'extension du lemme fondamental de Neyman-Pearson, un test UPP parmi les tests de niveau 5% est de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad x > k \\ c & \text{si} \quad x = k \\ 0 & \text{si} \quad x < k, \end{cases}$$

avec $E_{1/4}[\phi] = 5\%$.
Calcul des constantes k et c : on choisit $k = s = 11$ car $P_{1/4}(\{x \in \mathbb{N}^*, x > 11\}) = 0.0422$ et $P_{1/4}(\{x \in \mathbb{N}^*, x > 10\}) = 0.0563$, puis c telle que $P_{1/4}(\{x \in \mathbb{N}^*, x > 11\}) + cP_{1/4}(11) = 0.05$ d'où $c = \frac{0.05-0.0422}{0.0563-0.0422} = 0.5532$, c'est-à-dire $\phi(x) = \mathbb{1}_{x>11} + 0.5532\,\mathbb{1}_{x=11}$.
7. Ici, $x = 10$ donc $\phi(x) = 0$ donc on ne rejette pas l'hypothèse (H_0) au profit de (H_1) pour un niveau 5%.
8. D'après le théorème de Lehmann, le test construit dans la question précédente est aussi un test UPP parmi les tests de niveau 5% des hypothèses composites $(H_0) : \theta \geq 1/4$ contre $(H_1) : \theta < 1/4$.

4.6.2 Presse écrite en danger

Problème

On souhaite étudier les effets d'une campagne publicitaire sur la vente d'un journal hebdomadaire. En dehors de toute campagne publicitaire, on admet que le produit de la vente hebdomadaire de ce journal suit, en milliers d'euros, une loi normale $\mathcal{N}(m, \sigma^2)$, où m vaut 22. après la mise en place de la campagne publicitaire, on observe le produit de la vente du journal sur 8 semaines. On obtient alors les valeurs suivantes (en milliers d'euros) :

$$18.7, \quad 25.3, \quad 23.1, \quad 27.7, \quad 26.2, \quad 19.2, \quad 21.9, \quad 19.9,$$

et on veut savoir si la campagne de publicité est efficace ou non. On admet que la variance σ^2 du produit de la vente ne varie pas.
1. On suppose σ^2 connue et égale à 12.
a) Lorsqu'on teste

$$(H_0) \ : \ m = 22 \quad \text{contre} \quad (H_1) \ : \ m > 22,$$

se place-t-on du point de vue de l'éditeur du journal ou du concepteur de la campagne publicitaire ? Justifier la réponse.
b) Construire le test du rapport de vraisemblance maximale de (H_0) contre (H_1) pour un niveau 5%. Quelle est la conclusion du test ?
c) Pour quel niveau la conclusion du test serait-elle différente ?
d) Tracer la courbe de puissance du test.
e) Ce test est-il UPP parmi les tests de niveau 5% ?
f) Que doit-on faire si l'on souhaite augmenter la puissance du test tout en gardant un niveau 5% ?
g) Construire le test du rapport de vraisemblance maximale de taille 5% de

$$(H_0) \ : \ m = 22 \quad \text{contre} \quad (H_1') \ : \ m \neq 22.$$

Ce test est-il UPP parmi les tests de niveau 5% ?
2. On suppose maintenant que la variance σ^2 est inconnue.
a) Construire le test du rapport de vraisemblance maximale de taille 5% de

$$(H_0) \ : \ m \leq 22 \quad \text{contre} \quad (H_1) \ : \ m > 22.$$

b) Quelle est la conclusion de ce test ?
c) Ce test est-il sans biais ?
d) Quel résultat du cours utilisera-t-on si l'on veut montrer qu'il est UPP ?

Correction

1. a) Modèle statistique : $(\mathcal{X}, \{P_\theta\}_{\theta\in\Theta})$ où $\mathcal{X} = \mathbb{R}^8$, $\Theta = [22, +\infty[$, et pour $\theta \in \Theta$, P_θ est la loi d'un 8-échantillon $X = (X_1, \dots, X_8)$ d'une loi $\mathcal{N}(\theta, 12)$. On dispose d'une observation $x = (x_1, \dots, x_8)$ de cet échantillon.
Lorsqu'on teste $(H_0) : \theta = 22$ contre $(H_1) : \theta > 22$, on contrôle en premier lieu le risque de décider que la campagne publicitaire est efficace alors qu'elle ne l'est pas (rejeter (H_0) à tort). On se place donc du point de vue de l'éditeur du journal.
b) Le modèle statistique considéré est dominé par la mesure de Lebesgue sur $\mathbb{R}^8$. On note $L(x, \theta)$ sa vraisemblance. La statistique du test du rapport de vraisemblance est

$$\lambda(x) = \frac{L(x, 22)}{\sup_{\theta\geq 22} L(x, \theta)} = \frac{e^{-\frac{\sum_{i=1}^8 (x_i-22)^2}{24}}}{\sup_{\theta\geq 22} e^{-\frac{\sum_{i=1}^8 (x_i-\theta)^2}{24}}}.$$

Pour déterminer $\sup_{\theta\geq 22} e^{-\frac{\sum_{i=1}^8 (x_i-\theta)^2}{24}}$, on étudie les variations de la fonction $\theta \mapsto e^{-\frac{\sum_{i=1}^8 (x_i-\theta)^2}{24}}$ ou, de façon équivalente puisque la fonction ln est croissante, les variations de $\theta \mapsto -\frac{\sum_{i=1}^8 (x_i-\theta)^2}{24}$. On trouve alors que cette fonction est croissante sur $]-\infty, \bar{x}]$, puis décroissante sur $[\bar{x}, +\infty[$. Deux cas se présentent donc. Si $\bar{x} \leq 22$, $\sup_{\theta\geq 22} e^{-\frac{\sum_{i=1}^8 (x_i-\theta)^2}{24}} = e^{-\frac{\sum_{i=1}^8 (x_i-22)^2}{24}}$ et si $\bar{x} \geq 22$, $\sup_{\theta\geq 22} e^{-\frac{\sum_{i=1}^8 (x_i-\theta)^2}{24}} = e^{-\frac{\sum_{i=1}^8 (x_i-\bar{x})^2}{24}}$. On a donc

$$\lambda(x) = \begin{cases} e^{\frac{\sum_{i=1}^8 (x_i-\bar{x})^2 - \sum_{i=1}^8 (x_i-22)^2}{24}} & \text{si } \bar{x} \geq 22, \\ 1 & \text{si } \bar{x} < 22. \end{cases}$$

Comme $\sum_{i=1}^8 (x_i - \bar{x})^2 - \sum_{i=1}^8 (x_i - 22)^2 = -8(\bar{x} - 22)^2$,

$$\lambda(x) = \begin{cases} e^{\frac{-(\bar{x}-22)^2}{3}} & \text{si } \bar{x} \geq 22, \\ 1 & \text{si } \bar{x} < 22. \end{cases}$$

Le rapport $\lambda(x)$ est décroissant en $\bar{x}$, donc rejeter (H_0) lorsque $\lambda(x) \leq c$ revient à rejeter (H_0) lorsque $\bar{x} \geq c'$. Un test du rapport de vraisemblance maximale est donc de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si } \bar{x} \geq c', \\ 0 & \text{si } \bar{x} < c'. \end{cases}$$

Le test est de taille 5% si $P_{22}(\bar{X} \geq c') = 0.05$. Or, lorsque $X \sim P_{22}$, $\bar{X} \sim \mathcal{N}(22, 3/2)$, et comme $P_{22}(\bar{X} \geq c') = P_{22}\left(\frac{\bar{X}-22}{\sqrt{3/2}} \geq \frac{c'-22}{\sqrt{3/2}}\right)$, on choisit $c' = 22 + 1.645\sqrt{3/2} = 24.015$. Le test du rapport de vraisemblance de taille 5% est finalement donné par :

$$\phi(x) = \begin{cases} 1 & \text{si } \bar{x} \geq 24.015, \\ 0 & \text{si } \bar{x} < 24.015. \end{cases}$$

Puisqu'ici, $\bar{x} = 22.75$, on ne rejette pas (H_0) au profit de (H_1) pour un niveau 5%.
c) On aurait rejeté l'hypothèse (H_0) pour $\alpha = P_{22}(\bar{X} \geq 22.75) = P_{22}(\frac{\bar{X}-22}{\sqrt{3/2}} \geq 0.612)$. Pour un niveau égal à 27, 03%, on aurait rejeté (H_0).

d) Pour $\theta > 22$, $\pi(\theta) = P_\theta(\bar{X} \geq 24.015) = P_\theta\left(\frac{\bar{X}-\theta}{\sqrt{3/2}} \geq \frac{24.015-\theta}{\sqrt{3/2}}\right) = 1 - F\left(\frac{24.015-\theta}{\sqrt{3/2}}\right)$, où F est la fonction de répartition de la loi normale centrée réduite. π est croissante, et pour la tracer, on peut par exemple utiliser les valeurs :

θ	20	22	22.5	23	23.5	24.015	24.5	25	27
$\pi(\theta)$	0.0005	0.05	0.11	0.20	0.34	0.5	0.65	0.79	0.99

e) Pour $22 \leq \theta_1 < \theta_2$, le rapport $\frac{L(x,\theta_2)}{L(x,\theta_1)}$ est croissant en $\bar{x}$. Un corollaire du lemme de Neyman-Pearson nous dit donc que tout test de taille 5% de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si } \bar{x} > k, \\ c(x) & \text{si } \bar{x} = k, \\ 0 & \text{si } \bar{x} < k. \end{cases}$$

est UPP parmi les tests de niveau 5%. Le test du rapport de vraisemblance maximale précédent est donc bien UPP parmi les tests de niveau 5%.
f) A taille d'échantillon fixée, les risques de première et deuxième espèce varient en sens inverse. Par conséquent, si l'on souhaite augmenter la puissance du test tout en gardant un niveau 5%, on devra augmenter la taille de l'échantillon.
g) Modèle statistique : le même, mais avec $\Theta = \mathbb{R}$. La statistique du test du rapport de vraisemblance maximale est ici

$$\lambda(x) = \frac{L(x,22)}{\sup_{\theta\in\mathbb{R}} L(x,\theta)} = e^{\frac{\sum_{i=1}^8 (x_i-\bar{x})^2}{24} - \frac{\sum_{i=1}^8 (x_i-22)^2}{24}} = e^{-\frac{(\bar{x}-22)^2}{3}}.$$

Le rapport $\lambda(x)$ est croissant en $\bar{x}$ sur $]-\infty, 22]$, puis décroissant en $\bar{x}$ sur $[22, +\infty[$, donc rejeter (H_0) lorsque $\lambda(x) \leq c$ revient à rejeter (H_0) lorsque $\bar{x} \leq c_1$ ou lorsque $\bar{x} \geq c_2$, avec $c_1 < c_2$. Un test du rapport de vraisemblance maximale est donc de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si } \bar{x} \leq c_1 \text{ ou } \bar{x} \geq c_2, \\ 0 & \text{sinon.} \end{cases}$$

Il est de taille 5% si $P_{22}(\bar{X} \geq c_2) = P_{22}(\bar{X} \leq c_1) = 0.025$. Or, lorsque $X \sim P_{22}$, $\bar{X} \sim \mathcal{N}(22, 3/2)$, et comme $P_{22}(\bar{X} \geq c_2) = P_{22}\left(\frac{\bar{X}-22}{\sqrt{3/2}} \geq \frac{c_2-22}{\sqrt{3/2}}\right)$, on choisit $c_2 = 22 + 1.96\sqrt{3/2} = 24.4$. De la même façon, on choisira $c_1 = 22 - 1.96\sqrt{3/2} = 19.6$.
Un test de rapport de vraisemblance de taille 5% est finalement donné par :

$$\phi(x) = \begin{cases} 1 & \text{si } \bar{x} \leq 19.6 \text{ ou } \bar{x} \geq 24.4, \\ 0 & \text{sinon.} \end{cases}$$

Il n'est pas UPP(0.05).
2. On suppose maintenant que la variance σ^2 est inconnue.
a) Modèle statistique : $(\mathcal{X}, \{P_\theta\}_{\theta\in\Theta})$ où $\mathcal{X} = \mathbb{R}^8$, $\Theta = \mathbb{R} \times \mathbb{R}_+^*$, et pour $\theta = (m, \sigma^2) \in \Theta$, P_θ est la loi d'un 8-échantillon $X = (X_1, \ldots, X_8)$ d'une loi $\mathcal{N}(m, \sigma^2)$. On dispose d'une observation $x = (x_1, \ldots, x_8)$ de cet échantillon. Le modèle statistique considéré est dominé par la mesure de Lebesgue sur $\mathbb{R}^8$. On note $L(x, \theta)$ sa vraisemblance. La statistique du test du rapport de vraisemblance maximale est

$$\lambda(x) = \frac{\sup_{\theta\in]-\infty,22]\times\mathbb{R}_+^*} L(x,\theta)}{\sup_{\theta\in\Theta} L(x,\theta)}.$$

On sait que $L(x,.)$ atteint son maximum sur $\mathcal{M}\times\mathbb{R}_+^*$ en un point $(\hat{m},\hat{\sigma}^2)$ tel que $\hat{\sigma}^2 = \frac{1}{8}\sum_{i=1}^8 (x_i - \hat{m})^2$, et que si $\mathcal{M} = \mathbb{R}$, $\hat{m} = \bar{x}$. On a donc deux cas.
Si $\bar{x} \leq 22$,

$$\sup_{\theta\in]-\infty,22]\times\mathbb{R}_+^*} L(x,\theta) = L(x,(\bar{x},\frac{1}{n}\sum_{i=1}^8 (x_i-\bar{x})^2)),$$

et $\lambda(x) = 1$.
Si par contre, $\bar{x} \geq 22$,

$$\sup_{\theta\in]-\infty,22]\times\mathbb{R}_+^*} L(x,\theta) = L(x,(22,\frac{1}{8}\sum_{i=1}^8 (x_i-22)^2)),$$

et

$$\lambda(x) = \frac{L(x,(22,\frac{1}{8}\sum_{i=1}^8 (x_i-22)^2))}{L(x,(\bar{x},\frac{1}{8}\sum_{i=1}^8 (x_i-\bar{x})^2))} = \left(\frac{\sum_{i=1}^8 (x_i-\bar{x})^2}{\sum_{i=1}^8 (x_i-22)^2}\right)^4.$$

En posant $T(x) = \frac{\bar{x}-22}{\sqrt{\sum_{i=1}^8 (x_i-\bar{x})^2/8}}$, on a $\lambda(x) = \left(\frac{7}{T(x)^2+7}\right)^4$ si $\bar{x} \geq 22$, 1 sinon. Ainsi, $\lambda(x)$ est décroissant en $T(x)$. Par conséquent, rejeter (H_0) lorsque $\lambda(x) \leq c$ revient à rejeter (H_0) lorsque $T(x) \geq c'$. Un test du rapport de vraisemblance maximale est donc de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si } T(x) \geq c', \\ 0 & \text{si } T(x) < c'. \end{cases}$$

Il est de taille 5% si $P_{22}(T(X) \geq c') = 0.05$. Or, lorsque $X \sim P_{22}$, $T(X) \sim \mathcal{T}(7)$, donc on choisit $c' = 1.895$. Le test de rapport de vraisemblance de taille 5% est finalement donné par :

$$\phi(x) = \begin{cases} 1 & \text{si } T(x) \geq 1.895, \\ 0 & \text{si } T(x) < 1.895. \end{cases}$$

b) Puisqu'ici, $T(x) = 0.236$, on ne rejette pas (H_0) au profit de (H_1) pour un niveau 5%.
c) Pour $m > 22$, $\sigma^2 > 0$, $\pi(m,\sigma^2) = P_{(m,\sigma^2)}(T(X) \geq 1.895) =$
$P_{(m,\sigma^2)}\left(\frac{\bar{X}-m}{\sqrt{\sum_{i=1}^8 (X_i-\bar{X})^2/8}} + \frac{m-22}{\sqrt{\sum_{i=1}^8 (X_i-\bar{X})^2/8}} \geq 1.895\right) \geq P_{(m,\sigma^2)}\left(\frac{\bar{X}-m}{\sqrt{\sum_{i=1}^8 (X_i-\bar{X})^2/8}} \geq 1.895\right) = 0.05$. Le test est sans biais.
d) On utilisera le résultat d'extension du lemme de Neyman-Pearson. En effet, on ne peut utiliser aucun des résultats sur les modèles à rapport de vraisemblance monotone, puisque le paramètre n'est plus dans $\mathbb{R}$.

4.6.3 Ensemencement des nuages

Problème

Les techniques d'ensemencement des nuages par iodure d'argent, utilisées depuis de nombreuses années dans de nombreux pays pour réduire les dégâts causés par la grêle, sont aujourd'hui assez controversées. En France, une équipe de physiciens a choisi d'étudier l'efficacité d'un dispositif d'ensemencement mis en place par l'Association Nationale d'Etude et de Lutte contre les Fléaux Atmosphériques dans une région du Sud-Ouest. Ils souhaitent savoir en particulier si ce dispositif augmente de façon significative la probabilité qu'une précipitation solide soit seulement du grésil et non de la grêle. On admet que sans ensemencement, cette probabilité est inférieure à 0.3, et on note θ la probabilité qu'une précipitation solide après la mise en place du dispositif soit seulement du grésil.
Les physiciens ont étudié 30 épisodes de précipitations solides après la mise en place du dispositif, et regardé si ces précipitations correspondaient à du grésil ou de la grêle.

1. Introduire un modèle statistique permettant de décrire le problème posé.
2. Montrer que ce modèle est à rapport de vraisemblance monotone.
3. L'équipe de physiciens décide de tester l'hypothèse (H_0) $\theta \leq 0.3$ contre (H_1) $\theta > 0.3$.
a) Justifier le choix des hypothèses (H_0) et (H_1) par les physiciens. De quel risque se prémunissent-ils en faisant ce choix ?
b) Construire de façon intuitive un test de (H_0) contre (H_1) de niveau 5%.
c) Ce test est-il uniformément plus puissant parmi les tests de niveau 5% ?
Si oui, pourquoi ? Si non, comment faut-il modifier ce test pour qu'il le devienne ? Justifier précisément la réponse.
d) Sur les 30 précipitations solides étudiées par les physiciens, seulement 15 correspondaient à du grésil. Quelle peut être la conclusion ?

Extraits de tables statistiques

Pour $N \sim \mathcal{N}(0,1)$, on donne pour α, $q_{1-\alpha}$ tel que $P(N \leq q_{1-\alpha}) = 1-\alpha$.

α	0.01	0.025	0.05	0.1
$q_{1-\alpha}$	2.33	1.96	1.645	1.28

Pour $K \sim \mathcal{B}(30, 0.3)$, on donne pour k, la probabilité $P(K \leq k)$.

k	4	5	6	12	13	14
$P(K \leq k)$	0.03	0.077	0.16	0.915	0.96	0.98

Correction

1. Modèle statistique considéré : soit $X = (X_1, \dots, X_{30})$ un 30-échantillon d'une loi de Bernoulli de paramètre θ, avec $\theta \in \Theta = [0, 1]$, modélisant la nature de 30 précipitations solides ($X_i = 1$ si la précipitation i est du grésil, 0 sinon), et $x = (x_1, \dots, x_{30})$ l'observation de cet échantillon. Soit $\mathcal{X} = \{0, 1\}^{30}$, et $\mathcal{A}$ l'ensemble des parties de $\mathcal{X}$. Pour $\theta \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{B}(\theta)^{\otimes 30}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in[0,1]})$.

2. Ce modèle est dominé par la mesure de comptage sur $\{0, 1\}^{30}$ et sa vraisemblance est donnée par $L(x, \theta) = \theta^{\sum_{i=1}^{30} x_i}(1 - \theta)^{30-\sum_{i=1}^{30} x_i}$. Il est à rapport de vraisemblance croissant en $T(x) = \sum_{i=1}^{30} x_i$.

3. a) En faisant ce choix d'hypothèses, les physiciens privilégient l'hypothèse que le dispositif d'ensemencement n'est pas efficace tant qu'on ne leur a pas prouvé le contraire. Ils se prémunissent du risque de déclarer que l'ensemencement est efficace à tort.

b) Statistique de test et fonction de test : on prend comme statistique de test $T(x) = \sum_{i=1}^{30} x_i$. $T(X)/30 = \bar{X}$ est un estimateur sans biais de θ donc on choisit de rejeter (H_0) lorsque $T(x)$ prend de grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{T(x)>s}$.
Calcul de la constante s : en admettant que $\sup_{\theta\leq 0.3} P_\theta(\phi(X) = 1) = P_{0.3}(\phi(X) = 1)$, on cherche s telle que $P_{0.3}(\{x, T(x) > s\}) \leq 0.05$. Si X suit la loi $P_{0.3}$, $T(X) = \sum_{i=1}^{30} X_i$ suit la loi binomiale de paramètres $(30, 0.3)$. D'après les tables fournies, on choisit alors $s = 13$ et $\phi(x) = \mathbb{1}_{\sum_{i=1}^{30} x_i>13}$.

c) On a $P_{0.3}\left(\left\{x, \sum_{i=1}^{30} x_i > 13\right\}\right) = 0.04$ donc le test construit à la question précédente n'est pas de taille 5%. En conséquence, il ne peut pas être UPP parmi les tests de niveau 5%.
Le modèle considéré étant un modèle à rapport de vraisemblance croissant en $T(x) = \sum_{i=1}^{30} x_i$, d'après le théorème de Lehmann, un test UPP parmi les tests de niveau 5% est de la forme :

$$\tilde{\phi}(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{30} x_i > k \\ c & \text{si} \quad \sum_{i=1}^{30} x_i = k \\ 0 & \text{si} \quad \sum_{i=1}^{30} x_i < k, \end{cases}$$

avec $\sup_{\theta\leq 0.3} E_\theta[\tilde{\phi}(X)] = E_{0.3}[\tilde{\phi}(X)] = 5\%$.
Calcul des constantes k et c : on choisit $k = s = 13$ car $P_{0.3}\left(\left\{x, \sum_{i=1}^{30} x_i > 13\right\}\right) = 0.04$ et $P_{0.3}\left(\left\{x, \sum_{i=1}^{30} x_i > 12\right\}\right) = 0.085$, puis c telle que $P_{0.3}\left(\left\{x, \sum_{i=1}^{30} x_i > 13\right\}\right) + cP_{0.3}\left(\left\{x, \sum_{i=1}^{30} x_i = 13\right\}\right) = 0.05$ d'où $c = \frac{0.05-0.04}{0.085-0.04} = 0.22$, c'est-à-dire $\tilde{\phi}(x) = \mathbb{1}_{\sum_{i=1}^{30} x_i>13} + 0.22\, \mathbb{1}_{\sum_{i=1}^{30} x_i=13}$.

d) Ici, $\sum_{i=1}^{30} x_i = 15$ donc $\tilde{\phi}(x) = 1$ et on rejette l'hypothèse (H_0) au profit de (H_1) pour un niveau 5%.

Chapitre 5

Tests non paramétriques du Khi-Deux et de Kolmogorov-Smirnov

5.1 Les tests du Khi-Deux de Pearson

Le test du Khi-Deux est à l'origine un test d'adéquation (ou d'ajustement) d'une loi totalement inconnue à une loi donnée, mais il peut être utilisé pour vérifier l'indépendance ou l'homogénéité de deux variables aléatoires. Il est fondé sur une propriété asymptotique de la loi multinomiale.

5.1.1 La (pseudo) distance du Khi-Deux

Soit $X = (X_1, \dots, X_n)$ un n-échantillon de la loi P d'une variable aléatoire prenant ses valeurs dans un ensemble O. On considère une partition $\{O_1, \dots, O_m\}$ de O : $O = \cup_{k=1}^{m} O_k$ avec $O_k \cap O_l = \emptyset \ \forall k \neq l \in \{1, \dots, m\}$.
On définit, pour tout $k \in \{1, \dots, m\}$,

$$N_k(n) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in O_k\}} \text{ (Nombre de } X_i \text{ appartenant à } O_k).$$

Si $p_1, \dots, p_m$ désignent les probabilités pour X_i d'appartenir à $O_1, \dots, O_m$, alors le vecteur $(N_1(n), \dots, N_m(n))$ suit une loi multinomiale $\mathcal{M}(n, p_1, \dots, p_m)$:

$$P(N_1(n) = n_1, \dots, N_m(n) = n_m) = \frac{n!}{n_1! \dots n_m!} p_1^{n_1} \dots p_m^{n_m}.$$

On note P_n la loi empirique estimant P sur la base de l'échantillon $(X_1, \dots, X_n)$: $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$.

Définition 30. *La* (pseudo) distance du Khi-Deux entre P_n et P *est définie par*

$$D(P_n, P) = \sum_{k=1}^{m} \frac{(N_k(n) - np_k)^2}{np_k}.$$

Thèorème 12. *Lorsque n tend vers $+\infty$, la statistique $D(P_n, P)$ suit asymptotiquement une loi du Khi-Deux à $(m-1)$ degrés de liberté.*

Preuve. On utilise le théorème central limite vectoriel. On introduit les variables $Y_i = (\mathbb{1}_{X_i \in O_1}, \dots, \mathbb{1}_{X_i \in O_m})$ pour $i = 1 \dots n$. Alors $N(n) = \sum_{i=1}^n Y_i$, et $cov(Y_{i,k}, Y_{i,l}) = E[\mathbb{1}_{X_i \in O_k} \mathbb{1}_{X_i \in O_l}] - p_k p_l = \delta_k^l p_k - p_k p_l$. La matrice des variances-covariances de Y est donc donnée par $\Sigma = \Delta_\pi - \pi\pi'$, où $\pi = (p_1, \dots, p_m)'$ et Δ_π est la matrice diagonale dont les éléments diagonaux sont les composantes de π. Le TCL vectoriel donne alors

$$\frac{N(n) - n\pi}{\sqrt{n}} \to^{(\mathcal{L})} \mathcal{N}_m(0, \Sigma).$$

Si $f : \mathbb{R}^m \to \mathbb{R}_+, t \mapsto \sum_{k=1}^m \frac{t_k^2}{p_k}$, alors $D(P_n, P) = f\left(\frac{N(n)-n\pi}{\sqrt{n}}\right)$, d'où $D(P_n, P) \to^{(\mathcal{L})} f(Z)$, avec $Z = (Z_1, \dots, Z_m) \sim \mathcal{N}_m(0, \Sigma)$.
Loi de $f(Z)$? $f(Z) = \|AZ\|^2$, avec

$$A = \begin{pmatrix} 1/\sqrt{p_1} & & \\ & \ddots & \\ & & 1/\sqrt{p_m} \end{pmatrix}.$$

Pour toute transformation orthogonale U de $\mathbb{R}^m$ dans $\mathbb{R}^m$, alors $f(Z) = \|AZ\|^2 = \|UAZ\|^2$. Or AZ suit la loi normale centrée de matrice de variances-covariances $A\Sigma A' = I - \sqrt{\pi}\sqrt{\pi}'$, donc UAZ suit la loi normale centrée de matrice de variances-covariances $U(I - \sqrt{\pi}\sqrt{\pi}')U' = I - (U\sqrt{\pi})(U\sqrt{\pi})'$. En prenant U telle que $U\sqrt{\pi} = (0, \dots, 0, 1)'$ (on peut car $\|\sqrt{\pi}\| = 1$), alors

$$U(I - \sqrt{\pi}\sqrt{\pi}')U' = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix}.$$

On a donc $f(Z) = \|\tilde{Z}\|^2$ avec $\tilde{Z} \sim \mathcal{N}\left(0, \begin{pmatrix} I_{m-1} & 0 \\ 0 & 1 \end{pmatrix}\right)$. Ce qui signifie que les $\tilde{Z}_k$ sont indépendantes, que $\tilde{Z}_k \sim \mathcal{N}(0,1)$ pour $k = 1 \dots (m-1)$, et $\tilde{Z}_m = 0$ p.s. La loi de $f(Z)$ est donc la loi $\chi^2(m-1)$.

Remarque. Le raisonnement de cette preuve est similaire à celui de la preuve du théorème de Cochran, qu'on peut également utiliser ici directement.

5.1.2 Le test du Khi-Deux d'adéquation

On dispose d'une observation $x = (x_1, \dots, x_n)$ de l'échantillon X (ou éventuellement d'une observation $(n_1, \dots, n_m)$ du vecteur $N(n) = (N_1(n), \dots, N_m(n))$). On se donne une loi P^0 (parfaitement spécifiée) sur O, et on considère ici le problème de test de

$$(H_0) \ : \ P = P^0 \text{ contre } (H_1) \ : \ P \neq P^0.$$

On note $p_k^0 = P^0(O_k)$.
Intuitivement, si les X_i suivent la loi P^0, la (pseudo) distance du Khi-Deux $D(P_n, P^0)$ entre P_n et P^0 sera petite. Par ailleurs, on sait que si les X_i suivent la loi P^0, $D(P_n, P^0)$ suit asymptotiquement une loi du χ^2 à $(m-1)$ degrés de liberté, et que s'il existe k tel que $p_k \neq p_k^0$, d'après la loi forte des grands nombres, $N_k(n)/n \to p_k \neq p_k^0$ p.s. $\Rightarrow D(P_n, P^0) \to +\infty$ p.s.
On peut alors utiliser $D(P_n, P^0)$ comme statistique de test. Cela donne le test suivant :

Statistique de test :

$$T(X) = D(P_n, P^0) = \sum_{k=1}^{m} \frac{(N_k(n) - np_k^0)^2}{np_k^0}.$$

Fonction de test : $\phi(x) = \mathbb{1}_{T(x)\geq s}$.

Choix de la valeur critique s : si tous les np_k^0 sont supérieurs ou égaux à 5, on considère en pratique l'approximation de la loi de $D(P_n, P^0)$ lorsque $P = P^0$ par la loi $\chi^2(m-1)$ valide, et on choisira s telle que que $P(K \geq s) = \alpha$, avec $K \sim \chi^2(m-1)$.

Remarques importantes.

— Le test du Khi-Deux est un test asymptotique.

— En pratique, si les np_k^0 ne sont pas tous supérieurs ou égaux à 5, on modifie la partition de O en regroupant par exemple certaines de ses classes.

— Si l'on veut tester l'appartenance à une famille de lois paramètrée par $\theta \in \mathbb{R}^r$, θ sera estimé sur la base de l'échantillon $(X_1, \ldots, X_n)$ (par exemple par l'estimateur du maximum de vraisemblance, ou en tout cas, par un estimateur consistant et asymptotiquement normal), et le test reste inchangé, à ceci prés que s sera choisie tel que $P(K \geq s) = \alpha$, avec $K \sim \chi^2(m-r-1)$.

5.1.3 Le test du Khi-Deux d'indépendance

Le test du Khi-Deux peut également être utilisé dans le cadre suivant : on considère un couple de variables aléatoires (Y, Z), où Y est à valeurs dans $\{a_1, \ldots, a_r\}$ et Z dans $\{b_1, \ldots, b_t\}$. On dispose de l'observation $x = ((y_1, z_1), \ldots, (y_n, z_n))$ d'un n-échantillon $X = ((Y_1, Z_1), \ldots, (Y_n, Z_n))$ de la loi de ce couple.
On veut tester

$$(H_0) \ : \ Y \text{ et } Z \text{ sont indépendantes contre } (H_1) \ : \ Y \text{ et } Z \text{ ne sont pas indépendantes.}$$

Dans ce cas, l'hypothèse (H_0) s'écrit $P((Y, Z) = (a_i, b_j)) = p_{i,*}p_{*,j}$, donc on peut voir l'hypothèse (H_0) comme une hypothèse d'appartenance à une famille paramétrique de lois. En estimant le paramètre inconnu

$$\theta = (p_{1,*}, \ldots, p_{r-1,*}, p_{*,1}, \ldots, p_{*,t-1})$$

par maximum de vraisemblance (maximisation sous contrainte à l'aide des multiplicateurs de Lagrange), on obtient le test suivant.

Statistique de test :

$$T(X) = \sum_{i=1}^{r}\sum_{j=1}^{t} \frac{(\frac{N_{i,*}N_{*,j}}{n} - N_{i,j})^2}{\frac{N_{i,*}N_{*,j}}{n}},$$

où

— $N_{i,j}$ est le nombre d'éléments de $\{(Y_1, Z_1), \ldots, (Y_n, Z_n)\}$ qui prennent la valeur (a_i, b_j),

— $N_{i,*}$ est le nombre d'éléments de $\{Y_1, \ldots, Y_n\}$ qui prennent la valeur a_i,

— $N_{*,j}$ est le nombre d'éléments de $\{Z_1, \ldots, Z_n\}$ qui prennent la valeur b_j.

Fonction de test : $\phi(x) = \mathbb{1}_{T(x)\geq s}$.

Choix de la valeur critique s : sous l'hypothèse (H_0), on peut montrer que $T(X)$ suit asymptotiquement la loi $\chi^2(rt-(r+t-2)-1=(r-1)(t-1))$. s sera donc choisie telle que $P(K \geq s) = \alpha$, avec $K \sim \chi^2((r-1)(t-1))$.

Remarque. Si les variables considérées ne sont pas des variables discrètes à support fini, on peut considérer comme pour le test du Khi-Deux d'adéquation des partitions finies des supports des lois de Y et Z.

5.1.4 Le test du Khi-Deux d'homogénéité

Enfin, le test du Khi-Deux peut être utilisé dans le cadre de la comparaison des lois de deux échantillons indépendants.
On considère un couple de variables aléatoires (Y, Z), où Y et Z sont à valeurs dans $\{a_1, \dots, a_m\}$. On suppose que Y et Z sont indépendantes. On considère un n_1-échantillon $(Y_1, \dots, Y_{n_1})$ de la loi de Y et un n_2-échantillon $(Z_1, \dots, Z_{n_2})$ de la loi de Z, supposés indépendants, et on dispose de l'observation $x = (y_1, \dots, y_{n_1}, z_1, \dots, z_{n_2})$ de $X = (Y_1, \dots, Y_{n_1}, Z_1, \dots, Z_{n_2})$.
On veut tester

(H_0) : Y et Z suivent la même loi contre (H_1) : Y et Y ne suivent pas la même loi.

Statistique de test :

$$T(X) = \sum_{k=1}^{m} n_1 \frac{(\frac{N_k+M_k}{n_1+n_2} - \frac{N_k}{n_1})^2}{\frac{N_k+M_k}{n_1+n_2}} + n_2 \frac{(\frac{N_k+M_k}{n_1+n_2} - \frac{M_k}{n_2})^2}{\frac{N_k+M_k}{n_1+n_2}},$$

où
— N_k est le nombre d'éléments de $\{Y_1, \dots, Y_{n_1}\}$ qui prennent la valeur a_k,
— M_k est le nombre d'éléments de $\{Z_1, \dots, Z_{n_2}\}$ qui prennent la valeur a_k,

Fonction de test : $\phi(x) = \mathbb{1}_{T(x) \geq s}$.

Choix de la valeur critique s : sous l'hypothèse (H_0), $T(X)$ suit asymptotiquement une loi $\chi^2(m-1)$. s sera donc choisie telle que $P(K \geq s) = \alpha$, avec $K \sim \chi^2(m-1)$.

5.2 Test de Kolmogorov-Smirnov, extensions et généralisations

5.2.1 Le test de Kolmogorov-Smirnov d'adéquation

Soit $X = (X_1, \dots, X_n)$ un n-échantillon d'une loi P absolument continue par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ inconnue. Soit $x = (x_1, \dots, x_n)$ une observation de cet échantillon. La fonction de répartition notée F associée à P, définie pour tout $t \in \mathbb{R}$, $F(t) = P(X_i \leq t)$, est inconnue. On peut l'estimer par F_n la fonction de répartition empirique associée à l'échantillon X, définie pour $t \in \mathbb{R}$ par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(X_i).$$

Notons que pour chaque $t \in \mathbb{R}$ fixé, $F_n(t)$ est une variable aléatoire à valeurs dans $[0, 1]$, et que par la loi forte des grands nombres, $F_n(t) \to F(t)$ p.s.
En fait, l'estimation de F par F_n est justifiée par le résultat de convergence uniforme (plus fort) suivant.

Thèorème 13 (Glivenko-Cantelli (admis))**.**

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to_{n\to\infty} 0 \; p.s.$$

Comme pour le test du Khi-Deux, on introduit alors une (pseudo) distance entre la mesure empirique P_n associée à l'échantillon X et P, qui est tout simplement la distance en norme infinie entre la fonction de répartition empirique F_n et la fonction de répartition F :

$$D_{KS}(P_n, P) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|.$$

Proposition 8. *Si $(X_{(1)}, \dots, X_{(n)})$ est la statistique d'ordre associée à l'échantillon X, alors*

$$D_{KS}(P_n, P) = \max_{1\le i\le n} \max \left\{ \left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\}.$$

Thèorème 14. *Soit $D_n(X) = \sqrt{n} D_{KS}(P_n, P)$.*
(i) Si $U_n = \sqrt{n} \sup_{t\in[0,1]} |F_{U,n}(t) - t|$, où $F_{U,n}$ est la fonction de répartition empirique associée à n-échantillon de la loi uniforme sur $[0,1]$, alors $D_n(X)$ suit la même loi que U_n. On dira que la loi de la statistique $D_n(X)$ est libre de P (elle est entièrement connue). Cette loi est tabulée pour des petites valeurs de n.
(ii) (Massart 1990) Pour tout n, $P(D_n(X) \le t) \ge 1 - 2e^{-2t^2}$.
(iii) (Kolmogorov 1933) $D_n(X)$ converge en loi vers une loi (tabulée) dont la fonction de répartition est donnée par

$$H(t) = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2t^2}.$$

(iv) Si P^0 est une loi donnée absolument continue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, telle que $P \neq P^0$, alors $D_n^0(X) = \sqrt{n} D_{KS}(P_n, P^0) \to^{(P)} \infty$.

Preuve. (*i*) On rappelle la définition de la fonction quantile (ou inverse généralisée de la fonction de répartition) : $F^{-1}(u) = \inf\{t, F(t) \ge u\}$.

Les fonctions F et F^{-1} vérifient les propriétés suivantes :

1. Pour tout $t \in \mathbb{R}$, $F^{-1}(F(t)) \le t$.
2. Pour tout $u \in (0,1)$, $F\left(F^{-1}(u)\right) \ge u$.
3. $F^{-1}(u) \le t$ si et seulement si $u \le F(t)$.
4. Si $U \sim \mathcal{U}(0,1)$, $F^{-1}(U)$ est de même loi que X.

Pour démontrer ces propriétés, on note $\mathcal{F}_u$ l'ensemble $\mathcal{F}_u = \{t,\ F(t) \ge u\}$. Alors $F^{-1}(u) = \inf \mathcal{F}_u$.
1. Soit $t \in \mathbb{R}$, alors $t \in \mathcal{F}_{F(t)}$ donc $F^{-1}(F(t)) \le t$.
2. Par définition, il existe une suite $(t_n)_n$ décroissante d'éléments de $\mathcal{F}_u$ telle que $t_n \to F^{-1}(u)$. Puisque $F(t_n) \ge u$ pour tout n et que F est continue à droite, $F(F^{-1}(u)) \ge u$.
3. Supposons $u \le F(t)$. Alors $t \in \mathcal{F}_u$ donc $F^{-1}(u) \le t$.
Pour la réciproque, supposons $F^{-1}(u) \le t$. Alors par croissance de F, $F(F^{-1}(u)) \le F(t)$. Par le point 2 ci-dessus, cela entraîne que $F(t) \ge u$.
4. Pour $t \in \mathbb{R}$, par le point 3 ci-dessus, $P(F^{-1}(U) \le t) = P(U \le F(t)) = F(t)$, et la fonction F caractérise la loi de X.

Donc, si U_i suit une loi uniforme sur $[0,1]$, $F^{-1}(U_i)$ suit la même loi que X_i et $F_{U,n}(F(t)) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i \leq F(t)} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{F^{-1}(U_i)\leq t} =^{(\mathcal{L})} F_n(t)$. Donc $\sqrt{n}\sup_{t\in\mathbb{R}}|F_{U,n}(F(t)) - F(t)|$ suit la même loi que $D_n(X)$ et comme F est continue, lorsque t parcourt $\mathbb{R}$, $F(t)$ parcourt $[0,1]$. Finalement, on a bien que U_n suit la même loi que $D_n(X)$.
(*iii*) Admis. Voir l'ouvrage de Billingsley par exemple.
(*iv*) Il existe $\varepsilon > 0$, $t > 0$ tels que $F(t) - F^0(t) \geq \varepsilon$ ou $F^0(t) - F(t) > \varepsilon$. Supposons (sans perte de généralité) que $F(t) - F^0(t) > \varepsilon$.
Alors

$$D_n^0(X) \geq \sqrt{n}|F_n(t) - F^0(t)| \geq \sqrt{n}(F_n(t) - F(t)) + \sqrt{n}(F(t) - F^0(t)) \geq \sqrt{n}(F_n(t) - F(t)) + \sqrt{n}\varepsilon.$$

D'après la loi des grands nombres, $F_n(t) - F(t) \to^{(P)} 0$, donc $P(F_n(t) - F(t) \leq -\varepsilon/2) \to 0$ et pour tout $M > 0$,

$$P(D_n^0(X) \geq M) \to 1.$$

Application de ces résultats à la construction du test de Kolmogorov-Smirnov d'adéquation.

On se donne maintenant une loi P^0 sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, dont la fonction de répartition F^0 est continue et on considère le problème de test de

$$(H_0) \ : \ P = P^0 \text{ contre } (H_1) \ : \ P \neq P^0.$$

Statistique de test : au vu du théorème précédent, il est naturel de considérer comme statistique de test la statistique

$$D_n^0(X) = \sqrt{n}D_{KS}(P_n, P^0) = \sqrt{n}\max_{1\leq i\leq n}\max\left\{\left|F^0(X_{(i)}) - \frac{i}{n}\right|, \left|F^0(X_{(i)}) - \frac{i-1}{n}\right|\right\}.$$

Forme de la région critique et fonction de test : puisque sous (H_1), $D_n^0(X) \to +\infty$, on choisit $\phi(x) = \mathbb{1}_{D_n^0(x)\geq s}$.

Choix de la valeur critique s : pour n petit (en pratique $n \leq 100$), on utilise les tables de la loi de U_n, pour n plus grand, on utilise les tables de la loi asymptotique de fonction de répartition H.

Remarques.

- Le test de Kolmogorov-Smirnov n'est utilisable que pour les lois absolument continues !
- On constate souvent en pratique que le test de Kolmogorov-Smirnov est plus puissant que le test du Khi-Deux.
- Comme pour le test du Khi-Deux, on peut dans certains cas adapter le test de Kolmogorov-Smirnov pour tester l'appartenance à une famille paramétrique de lois $\{P_\theta, \theta \in \Theta\}$: familles de lois uniformes, gaussiennes, exponentielles. Dans ce cas, on doit montrer qu'en prenant l'estimateur empirique classique $\hat{\theta}$ du paramètre θ de la famille de lois, la loi de la statistique $\sqrt{n}D_{KS}(P_n, P_{\hat{\theta}})$ est libre de P. On verra dans la suite l'exemple du test de normalité (Lilliefors).

5.2.2 Le test de Kolmogorov-Smirnov d'homogénéité

On considère un couple de variables aléatoires (Y, Z), tel que Y et Z sont supposées indépendantes, de lois respectives P_Y et P_Z. On considère un n_1-échantillon $(Y_1, \dots, Y_{n_1})$ de la loi de Y et un n_2-échantillon $(Z_1, \dots, Z_{n_2})$ de la loi de Z, supposés indépendants, et on dispose de l'observation $x = (y_1, \dots, y_{n_1}, z_1, \dots, z_{n_2})$ de $X = (Y_1, \dots, Y_{n_1}, Z_1, \dots, Z_{n_2})$.
On note F_{n_1} la fonction de répartition empirique associée à $(Y_1, \dots, Y_{n_1})$ et G_{n_2} celle associée à $(Z_1, \dots, Z_{n_2})$, et on introduit $D_{n_1,n_2}(X) = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{t\in\mathbb{R}} |F_{n_1}(t) - G_{n_2}(t)|$.

Thèorème 15. (*i*) *La loi de la statistique $D_{n_1,n_2}(X)$ est libre de P_Y et P_Z lorsque $P_Y = P_Z$. Cette loi est tabulée. Elle ne dépend que de n_1 et n_2.*
(*ii*) *Si $P_Y \neq P_Z$, alors $D_{n_1,n_2}(X) \to^{(P)} +\infty$.*

Application de ces résultats à la construction du test de Kolmogorov-Smirnov d'homogénéité.

On veut tester

$$(H_0) \ : \ Y \text{ et } Z \text{ suivent la même loi contre } (H_1) \ : \ Y \text{ et } Z \text{ ne suivent pas la même loi.}$$

Statistique de test : $D_{n_1,n_2}(X)$.

Fonction de test : $\phi(x) = \mathbb{1}_{D_{n_1,n_2}(x)\geq s}$.

Choix de la valeur critique s : à partir des tables du test de Kolmogorov-Smirnov d'homogénéité.

5.2.3 Tests de Cramér-von Mises et Anderson-Darling

Soit $X = (X_1, \dots, X_n)$ un n-échantillon d'une loi inconnue P. On se donne, comme pour le test de Kolmogrov-Smirnov une loi P^0 sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, absolument continue par rapport à la mesure de Lebesgue, dont la fonction de répartition et la densité sont notées F^0 et f^0 respectivement. Sur la base d'une observation $x = (x_1, \dots, x_n)$ de X, on souhaite tester :

$$(H_0) \ : \ P = P^0 \text{ contre } (H_1) \ : \ P \neq P^0.$$

On considère différentes mesures de dissimilarité entre F_n et F^0.
Statistique de test de Cramér-von Mises :

$$C_n^0(X) = n \int_{\mathbb{R}} \left(F_n(t) - F^0(t)\right)^2 f^0(t)dt = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F^0(X_{(i)})\right)^2.$$

Fonction de test de Cramér-Von Mises : $\phi(x) = \mathbb{1}_{C_n^0(x)\geq s}$.

Statistique de test d'Anderson-Darling :

$$A_n^0(X) = n \int_{\mathbb{R}} \left(F_n(t) - F^0(t)\right)^2 \frac{f^0(t)}{F^0(t)(1-F^0(t))}dt = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left(\ln(F^0(X_{(i)})) + \ln(1 - F^0(X_{(n+1-i)})\right).$$

Fonction de test d'Anderson-Darling : $\phi(x) = \mathbb{1}_{A_n^0(x)\geq s}$.

Choix de la valeur critique s : sous (H_0), les lois de $C_n^0(X)$ et $A_n^0(X)$ sont libres de P, la valeur critique s est à lire dans les tables de Cramér-von Mises et Anderson-Darling.

Remarque : ces tests s'étendent à la comparaison de la loi de X à celle d'un autre échantillon indépendant en remplaçant F^0 par la fonction de répartition empirique G_n du deuxième échantillon et $f^0(t)dt$ par $dH(t)$ où H est la fonction de répartition empirique associée au vecteur aléatoire composé des variables issues des deux échantillons agrégés, et en renormalisant différemment la statistique de test (c.f. [1] pour plus de détails).

5.2.4 Test de normalité de Lilliefors

Soit $X = (X_1, \dots, X_n)$ un n-échantillon d'une loi inconnue P. Sur la base d'une observation $x = (x_1, \dots, x_n)$ de X, on souhaite tester :

$$(H_0) \ : \ X \text{ suit une loi gaussienne contre } (H_1) \ : \ X \text{ ne suit pas une loi gaussienne.}$$

Le test de normalité de Lilliefors [16] revient - intuitivement (attention une loi de paramètres aléatoires n'existe pas) - à appliquer le test de Kolmogorov-Smirnov d'adéquation à une loi $\hat{P}^0$ correspondant à une loi gaussienne d'espérance $\bar{X}$ et de variance $S^2(X) = \sum_{i=1}^n (X_i - \bar{X})^2$.

Statistique de test :

$$L_n(X) = \sqrt{n} D_{KS}(P_n, \hat{P}^0) = \sqrt{n} \max_{1 \le i \le n} \max \left\{ \left| F^0\left(\frac{X_{(i)} - \bar{X}}{S(X)}\right) - \frac{i}{n} \right|, \left| F^0\left(\frac{X_{(i)} - \bar{X}}{S(X)}\right) - \frac{i-1}{n} \right| \right\},$$

Fonction de test : $\phi(x) = \mathbb{1}_{L_n(x) \ge s}$.

Choix de la valeur critique s : la loi du vecteur aléatoire composé des $\frac{X_i - \bar{X}}{S(X)}$ étant libre de l'espérance et la variance des X_i sous (H_0), la loi de $L_n(X)$ sous (H_0) est libre de P, mais attention, elle diffère de celle de Kolmogorov-Smirnov ! Donc s est à lire dans la table spécifique de Lilliefors.

Remarque : on peut, de la même façon, construire un test de normalité à partir des tests de Cramér-von Mises et Anderson-Darling.

5.3 Exercices

Exercice 1 : La vie en couleurs des M&M's

On souhaite tester l'hypothèse suivant laquelle la répartition des couleurs de M&M's dans un paquet est bien en moyenne celle annoncée officiellement :

— Jaune : 15%
— Rouge : 12%
— Orange : 23%
— Bleu : 23%
— Marron : 12%
— Vert : 15%

Compter le nombre de M&M's de chaque couleur dans les paquets distribués et conclure sur l'observation de cet échantillon.

Exercice 2 : Attaque d'insectes

Les fruits d'un verger en agriculture biologique ont subi une attaque d'insectes. Sur un lot de quatre cents fruits, on compte le nombre d'insectes contenu dans chaque fruit. On obtient les résultats suivants :

Nombre d'insectes par fruit	0	1	2	3	4	5	6	7	≥ 8
Nombre de fruits	85	138	104	49	15	8	0	1	0

Les observations confirment-elles l'hypothèse selon laquelle le nombre d'insectes présents dans un fruit choisi arbitrairement est une variable aléatoire obéissant à une loi de Poisson ?

Exercice 3 : Tir à l'arc

Un statisticien, futur candidat de Koh-Lanta, s'entraîne pour l'épreuve de tir à l'arc. Il installe une cible dans son jardin et la vise 30 fois. A chaque tir i ($i \in \{1, \dots, 30\}$), il apprécie la précision de son tir en mesurant l'éloignement X_i entre le centre de la cible et le point d'impact de sa flèche. Si la mesure $x_i = 0$, il a tiré au centre de la cible, si $x_i < 0$, il a tiré dans la demi-cible du bas, si $x_i > 0$ il a tiré dans la demi-cible du haut.
Les mesures sont les suivantes (en dm) :
$-1.8, 1, -0.1, 3, -1.3, -1.4, 1.3, -0.8, 0.3, -1, -1, 0, 0.6, -0.2, 0.6, 0.8, 1, -1.4, -1.6, -2.3, -1.2, 4, 4.3, 0, 1, 1.2, 3.8, -1, 0.5, -2.9$. Les tirs sont considérés indépendants entre eux et on admet que le statisticien est un excellent tireur si les X_i suivent une loi normale $\mathcal{N}(0, 4)$.
1. On suppose ici que les X_i suivent une loi normale. Tester au niveau 5% les hypothèses :

$$(H_0) \ : \ \sigma^2 = 4 \text{ contre } (H_1) \ : \ \sigma^2 \neq 4,$$

puis

$$(H_0) \ : \ m = 0 \text{ contre } (H_1) \ : \ m \neq 0.$$

2. A l'aide d'un test du Khi-Deux de niveau asymptotique 5% sur les classes $]-\infty, -1[$, $[-1, 0[$, $[0, 1[$, $[1, 2[$, $[2, +\infty[$, dire si le statisticien peut être considéré comme un excellent tireur.
3. Quel test le statisticien aurait-il fait s'il avait seulement souhaité vérifier que les X_i suivent une loi gaussienne ?

Exercice 4 : Taille des hommes politiques

Afin de mener une étude sociologique sur la taille des hommes politiques et leur popularité, on souhaite s'assurer au préalable que la taille d'un homme politique peut être supposée de loi gaussienne. On relève pour cela la taille de 500 hommes politiques tirés au sort. On obtient les résultats suivants :

taille (cm)	≤ 161	$]161, 163]$	$]163, 165]$	$]165, 167]$	$]167, 169]$	$]169, 171]$
effectif	4	1	25	35	75	115
taille (cm)	$]171, 173]$	$]173, 175]$	$]175, 177]$	$]177, 179]$	$]179, 181]$	> 181
effectif	125	60	40	10	5	5

Que peut-on conclure ?

Exercice 5 : Bienvenue chez les Ch'tis

Dans un village du Nord de la France, 500 personnes sont allées voir le film "Bienvenue chez les Ch'tis" à sa sortie, dont 100 sont retraitées, 50 sont chômeuses, et 350 sont actives. Les impressions à la sortie sont les suivantes :

	Très satisfait	Assez satisfait	Déçu
Chômeurs	35	10	5
Actifs	90	210	50
Retraités	70	23	7

Dans ce village, l'opinion sur le film dépend-elle de l'activité ?

Exercice 6 : Aptitude à l'utilisation de SAS et R

Dans deux formations de Statistique, on a effectué des tests d'aptitude à l'utilisation des logiciels SAS et R. Pour chaque élève, le degré d'aptitude a été noté 1, 2, 3 et 4, du meilleur au moins bon.

	Formation 1					Formation 2			
SAS — R	1	2	3	4	SAS — R	1	2	3	4
1	38	72	108	112	1	42	50	104	108
2	58	80	127	141	2	76	84	122	136
3	98	138	202	161	3	96	132	186	184
4	131	146	166	222	4	112	138	199	231

1. Vérifier, pour les deux formations, l'indépendance entre les aptitudes à utiliser SAS et R au niveau asymptotique 5%.
2. On veut savoir si les formations conduisent au même degré d'aptitude. On sélectionne alors, dans les deux formations, les élèves ayant atteint un degré d'aptitude au moins égal à 2 pour chacun des deux logiciels. Tester l'hypothèse d'égalité des deux répartitions (selon la formation) à l'aide d'un test du Khi-Deux au niveau asymptotique 2.5%.

Exercice 7 : Tir à l'arc, la revanche

1. Reprendre l'exercice 3 ci-dessus et dire si le statisticien peut être considéré comme un excellent joueur à l'aide d'un test de Kolmogorov-Smirnov.
2. Rappeler les propriétés théoriques permettant de tabuler la loi de la statistique de test sous l'hypothèse nulle.

Exercice 8 : Cannabis thérapeutique

On souhaite comparer deux médicaments sensés soulager la douleur. On a observé sur 16 patients dont 8 ont pris le médicament A habituel et les 8 autres un médicament B expérimental à base de cannabis, les durées de soulagement suivantes (en heures) :

A	6.8	3.1	5.8	4.5	3.3	4.7	4.2	4.9
B	4.4	2.5	2.8	2.1	6.6	1.1	4.8	2.3

1. Proposer un test paramétrique de comparaison.
2. On veut maintenant effectuer un test non paramétrique de Kolmogorov-Smirnov. Expliquer pourquoi la statistique de test sous l'hypothèse nulle peut être tabulée. Conclure pour un niveau asymptotique 5%.

Exercice 9 : Test non paramétrique d'exponentialité

Soit $\theta > 0$ et $S_k = X_1 + \ldots + X_k$, où $X_1, \ldots, X_n$ sont des v.a.r. i.i.d. de loi exponentielle $\mathcal{E}(\theta)$.
1. Montrer que $\left(\frac{S_1}{S_n}, \ldots, \frac{S_{n-1}}{S_n}\right)$ suit la même loi que la statistique d'ordre $(U_{(1)}, \ldots, U_{(n-1)})$ d'un $(n-1)$ échantillon de la loi uniforme sur $[0, 1]$.
2. A partir du test de Kolmogorov-Smirnov, proposer un test non paramétrique d'exponentialité.

Chapitre 6

Tests non paramétriques basés sur les rangs ou les statistiques d'ordre

6.1 Symétrie : test des rangs signés de Wilcoxon

Soit $X = (X_1, \dots, X_n)$ un n-échantillon d'une loi P absolument continue par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ inconnue. On note $(X^a_{(1)}, \dots, X^a_{(n)})$ la statistique d'ordre associée aux $|X_i|$. Soit $x = (x_1, \dots, x_n)$ une observation de X.
La loi P étant continue, $P(|X_i| = |X_j|) = 0$ pour $i \neq j$, donc on considère qu'il n'y a pas d'ex æquo dans les $|X_i|$.

On désigne par R l'application qui, aux variables $|X_i|$, associe leur rang : $R(|X_i|) = k$ si $|X_i| = X^a_{(k)}$.

Test de symétrie. Sur la base de l'observation x, on souhaite tester

$$(H_0) \ : \ \text{La loi } P \text{ est symétrique contre } (H_1) \ : \ \text{La loi } P \text{ n'est pas symétrique.}$$

Statistique de test :

$$W_n^+(X) = \sum_{i=1}^n R(|X_i|)\mathbb{1}_{X_i>0} = \sum_{k=1}^n kB_k^+,$$

avec $B_k^+ = \mathbb{1}_{X_i \text{ telle que } |X_i|=X^a_{(k)} \text{ est strictement positive}}$.

Fonction de test : $\phi(x) = \mathbb{1}_{W_n^+(x)\leq s_1} + \mathbb{1}_{W_n^+(x)\geq s_2}$.

Choix des valeurs critiques s_1 et s_2 : sous (H_0), les variables B_k^+ sont i.i.d. de loi de Bernoulli de paramètre 0.5, donc la loi de $W_n^+(X)$ sous (H_0) est libre de P. Elle est symétrique par rapport à son espérance, qui est égale à $\frac{n(n+1)}{4}$. On peut donc déduire s_1 et s_2 de la table correspondante pour de faibles valeurs de n, ou utiliser l'approximation asymptotique de cette loi par une loi $\mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$ pour de grandes valeurs de n.

Test sur le centre de symétrie. Sur la base de l'observation x, sachant de façon sûre que la loi P est symétrique par rapport à un réel m inconnu, on souhaite tester

$$(H_0) \ : \ m = m_0 \text{ contre } (H_1) \ : \ m \neq m_0, \text{ ou } m > m_0, \text{ ou } m < m_0.$$

Statistique de test : $W_n^+(X_1 - m_0, \dots, X_n - m_0)$.

Fonction de test : $\phi(x) = \mathbb{1}_{W_n^+(x_1-m_0,\dots,x_n-m_0)\leq s_1} + \mathbb{1}_{W_n^+(x_1-m_0,\dots,x_n-m_0)\geq s_2}$, ou $\mathbb{1}_{W_n^+(x_1-m_0,\dots,x_n-m_0)\geq s}$, ou $\mathbb{1}_{W_n^+(x_1-m_0,\dots,x_n-m_0)\leq s}$.

6.2 Homogénéité : tests de Wilcoxon et Mann-Whitney

On considère un couple de variables aléatoires (Y, Z), tel que Y et Z sont supposées indépendantes, de lois respectives P_Y et P_Z. On considère un n_1-échantillon $(Y_1, \dots, Y_{n_1})$ de la loi de Y et un n_2-échantillon $(Z_1, \dots, Z_{n_2})$ de la loi de Z, supposés indépendants, et on dispose de l'observation $x = (y_1, \dots, y_{n_1}, z_1, \dots, z_{n_2})$ de $X = (Y_1, \dots, Y_{n_1}, Z_1, \dots, Z_{n_2})$.
On note F_{n_1} la fonction de répartition empirique associée à $(Y_1, \dots, Y_{n_1})$ et G_{n_2} celle associée à $(Z_1, \dots, Z_{n_2})$, et on introduit $D_{n_1,n_2}(X) = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{t\in\mathbb{R}} |F_{n_1}(t) - G_{n_2}(t)|$.
On suppose ici que P_Y et P_Z sont absolument continues par rapport à la mesure de Lebesgue sur $\mathbb{R}$ (probabilité nulle d'avoir des ex æquo dans les échantillons).
On note $n = n_1 + n_2$, et on redésigne par $(X_1, \dots, X_n)$ les variables composant X. Soit $(X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée à $(X_1, \dots, X_n)$.

On note R l'application qui, aux variables Y_i, associe leur rang dans $(X_{(1)}, \dots, X_{(n)})$: $R(Y_i) = k$ si $Y_i = X_{(k)}$.
Sur la base d'une observation $x = (y_1, \dots, y_{n_1}, z_1, \dots, z_{n_2})$ de X, on veut tester

$$(H_0) \ : \ Y \text{ et } Z \text{ suivent la même loi contre } (H_1) \ : \ Y \text{ et } Z \text{ ne suivent pas la même loi.}$$

Test de Wilcoxon. Statistique de test de Wilcoxon :

$$W_{n_1,n_2}(X) = \sum_{i=1}^{n_1} R(Y_i) = \sum_{k=1}^{n} kB_k,$$

où $B_k = \mathbb{1}_{X_{(k)}\in\{Y_1,\dots,Y_{n_1}\}}$.

Fonction de test : $\phi(x) = \mathbb{1}_{W_{n_1,n_2}(x)\leq s_1} + \mathbb{1}_{W_{n_1,n_2}(x)\geq s_2}$.

Choix des valeurs critiques s_1 et s_2 : sous (H_0), la loi de $W_{n_1,n_2}(X)$ est libre de P_Y et P_Z. Elle est symétrique par rapport à son espérance égale à $\frac{n_1(n_1+n_2+1)}{2}$. On peut donc déduire s_1 et s_2 de la table correspondante pour de faibles valeurs de n_1 et n_2, ou utiliser l'approximation asymptotique de cette loi par une loi $\mathcal{N}\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$ pour de grandes valeurs de n_1 et n_2.

Test de Mann-Whitney. Statistique de test de Mann-Whitney : considérant les $n_1 n_2$ couples (Y_i, Z_j), avec $n_1 \leq n_2$, la statistique de Mann-Whitney $U_{n_1,n_2}(X)$ correspond au nombre de couples (Y_i, Z_j) tels que $Y_i > Z_j$ ou encore

$$U_{n_1,n_2}(X) = W_{n_1,n_2}(X) - \frac{n_1(n_1+1)}{2}.$$

Fonction de test : $\phi(x) = \mathbb{1}_{U_{n_1,n_2}(x)\leq s_1} + \mathbb{1}_{U_{n_1,n_2}(x)\geq s_2}$.

6.3 Tests de Wilcoxon et Mann-Whitney : présence d'ex æquo

Les tests basés sur les rangs peuvent être également utilisés pour des lois discrètes. Dans ce cas, il faut gérer la présence possible d'ex æquo.
Deux méthodes sont alors possibles.

1. On range les variables par classe et on attribue à chacune des variables d'une classe le rang moyen correspondant à cette classe (méthode des rangs moyens), de façon à conserver la relation : $\sum_{i=1}^{n} R(X_i) = \frac{n(n+1)}{2}$. **Attention :** la loi de la statistique sous (H_0) s'en trouve modifiée. Pour plus de détails, on renvoie à [**?**] ou [21].
2. On attribue aux ex æquo des valeurs différenciées à l'aide d'une table de nombres au hasard. Les tables et les approximations asymptotiques habituelles peuvent alors être utilisées.

6.4 Normalité : test de Shapiro-Wilk

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon d'une loi inconnue P. Sur la base d'une observation $x = (x_1, \ldots, x_n)$ de X, on souhaite tester :

$$(H_0) \ : \ X \text{ suit une loi gaussienne contre } (H_1) \ : \ X \text{ ne suit pas une loi gaussienne.}$$

Statistique de test :

$$SW_n(X) = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2},$$

où

— les a_i sont des constantes dépendant de l'espérance m et de la matrice de variance-covariance V de la statistique d'ordre associée à un échantillon de taille n de loi gaussienne centrée réduite : $(a_1, \ldots, a_n) = \frac{m'V^{-1}}{\left(m'V^{-1}V^{-1}m\right)^{1/2}}$, et disponibles dans des tables spécifiques,
— $(X_{(1)}, \ldots, X_{(n)})$ est la statistique d'ordre associée à X.

Cette statistique de test peut s'interpréter comme le cœfficient de détermination entre le vecteur des quantiles générés à partir d'une loi gaussienne et celui des quantiles empiriques basés sur les données. Le tracé du premier vecteur en fonction du deuxième s'appelle un graphe quantile-quantile ou "Q-Q plot", que l'on compare graphiquement à la droite de Henry.

Fonction de test : $\phi(x) = \mathbb{1}_{SW_n(x) \leq s}$.

Choix de la valeur critique s : la loi de $SW_n(X)$ sous (H_0) est libre de P (c.f. [22] pour plus de détails), donc s est à lire dans la table de Shapiro-Wilk (comme les a_i).

Chapitre 7

Annales corrigées

7.1 Lutte contre la fraude fiscale

Sujet d'examen, année universitaire 2011-2012, durée : 1h30

Le problème suivant se compose de trois parties. Des extraits de tables statistiques sont donnés à la fin du sujet en Annexe 1.

Devant la crise de la dette publique en Europe, de nombreux états européens ont fait de la lutte contre la fraude fiscale une priorité nationale. Des dispositifs de détection de la fraude sont mis en place, et parmi eux, la détection de l'écart à la loi de Benford. Cette loi, découverte en 1881 par Simon Newcomb, alors passée inaperçue, fut redécouverte par Frank Benford en 1938, puis démontrée en 1996. Elle modélise la fréquence d'apparition du premier chiffre d'un nombre issu de résultats de mesures non falsifiées *(voir Annexe 2)*.
La table de la loi de Benford (en base 10) est donnée ci-dessous :

k	1	2	3	4	5	6	7	8	9
$P(C = k)$	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

C étant une variable aléatoire modélisant le premier chiffre d'un nombre issu de résultats de mesures.

Partie I : Test sur une probabilité

Lors du contrôle fiscal d'une entreprise, afin de détecter et d'identifier plus rapidement une éventuelle zone falsifiée dans la comptabilité de l'entreprise, un contrôleur relève au hasard 120 montants apparaissant dans cette zone.
On a remarqué que dans des comptes d'entreprises, les zones falsifiées comportent plus de montants commençant par le chiffre 6 que dans les zones non falsifiées, pour lesquelles la proportion théorique de montants commençant par 6 est égale à 6.7% selon la loi de Benford.
En première intention, le contrôleur décide donc de regarder la proportion de montants commençant par 6 dans les montants qu'il a rélevés.
On introduit un 120-échantillon $X = (X_1, \ldots, X_{120})$ de la loi de Bernoulli de paramètre θ modélisant l'apparition du chiffre 6 en première position sur 120 montants relevés au hasard

dans la zone de comptabilité considérée : pour $i = 1 \dots 120$, $X_i = 1$ si le i-éme montant commence par 6, 0 sinon.

1. Ecrire le modèle statistique correspondant. Quelles hypothèses ce modèle induit-il sur les montants relevés ? Qu'en pensez-vous ?
2. Le contrôleur souhaite tester sur la base de l'observation de X, $(H_0) : \theta = 0.067$ contre $(H_1) : \theta > 0.067$. Que signifie ce choix d'hypothèses ?
3. Construire un test intuitif de (H_0) contre (H_1) de niveau 5%.
4. Donner la taille de ce test, et calculer sa puissance pour $\theta = 0.1$.
5. Sur les 120 montants relevés par le contrôleur, 18 commencent par le chiffre 6. Que peut-il conclure ?
6. Calculer la p valeur du test et retrouver, à l'aide de cette p valeur, la conclusion ci-dessus.
7. Expliquer pourquoi ce test ne peut pas être uniformément plus puissant parmi les tests de niveau 5%.
8. Montrer que le modèle considéré est à rapport de vraisemblance croissant. En déduire la construction d'un nouveau test de (H_0) contre (H_1) uniformément plus puissant parmi les tests de niveau 5%.
9. Calculer la puissance de ce nouveau test pour $\theta = 0.1$.
10. La conclusion de ce nouveau test diffère-t-elle de celle obtenue à l'aide du test intuitif ?

Partie II : Test du Khi-Deux d'adéquation

Afin d'affiner son étude de première intention, le contrôleur s'intéresse en deuxième intention à la loi du premier chiffre d'un montant relevé dans la zone de comptabilité étudiée. Il souhaite tester l'hypothèse selon laquelle cette loi est la loi de Benford à l'aide d'un test du Khi-Deux d'adéquation. Il note pour cela le nombre de montants commençant par k parmi les 120 montants relevés, pour $k = 1 \dots 9$, et il obtient les résultats suivants.

k	1	2	3	4	5	6	7	8	9
Montants commençant par k	12	10	12	13	17	18	16	12	10

1. Donner la statistique de test du Khi-Deux d'adéquation à la loi de Benford.
2. Déterminer le test du Khi-Deux d'adéquation à la loi de Benford de niveau 5%.
3. Quelle est la conclusion du test ?

Partie III : Durée de contrôles fiscaux

Le temps (en jours) consacré habituellement au contrôle fiscal d'une petite entreprise peut être modélisé par une loi gaussienne d'espérance $m = 55$. Les études de première et deuxième intentions précédentes ayant pour but d'identifier rapidement d'éventuelles zones falsifiées dans les comptes d'une entreprise, on veut savoir si la mise en place d'un dispositif de contrôle intégrant de telles études permet de diminuer significativement la durée d'un contrôle fiscal en moyenne. On relève pour cela les durées $(d_1, \dots, d_{50})$ de 50 contrôles fiscaux après la mise en place d'un dispositif intégrant les études de première et deuxième

intentions précédentes. On obtient les résultats suivants : $\bar{d} = \frac{1}{50}\sum_{i=1}^{50} d_i = 52.78$ et $S^2(d) = \frac{1}{49}\sum_{i=1}^{50}\left(d_i - \bar{d}\right)^2 = 55.726$. On a mené l'étude statistique décrite ci-dessous.

---

1. Modèle statistique. Soit $D = (D_1, \ldots, D_{50})$ un 50-échantillon de la loi gaussienne d'espérance m inconnue, de variance σ^2 inconnue, modélisant les durées de 50 contrôles fiscaux après la mise en place du dispositif. Soit $\mathcal{X} = \mathbb{R}^{50}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{50})$, $\Theta = \mathbb{R}_+ \times \mathbb{R}_+^*$, et pour $\theta = (m, \sigma^2) \in \Theta$, $P_\theta = \mathcal{N}(m, \sigma^2)^{\otimes 50}$. Le modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.

2. hypothèses. On teste $(H_0) : m = 55$ contre $(H_1) : m < 55$. On souhaite ici se prémunir en priorité du risque de déclarer que la durée d'un contrôle fiscal a diminué avec la mise en place du dispositif alors que ce n'est pas le cas.

3. Statistique de test. On introduit la statistique suivante :

$$T(d) = \sqrt{50}\frac{\bar{d} - m}{\sqrt{S^2(d)}}.$$

Sous (H_0), $T(D)$ suit une loi de Student à 49 degrés de liberté.

4. Région de rejet du test de niveau 5% : $R_{(H_0)} = \{d, |T(d)| \geq 2.01\}$.

5. Conclusion. Ici, $T(d) = -2.1$ donc on rejette (H_0) au profit de (H_1) pour un niveau 5%.

Des erreurs se sont glissées dans cette étude statistique. Relevez-les et corrigez-les.

ANNEXE 1 : Extraits des tables statistiques

Table de la loi gaussienne centrée réduite : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi binomiale : on donne pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi binomiale de paramètres 120 et 0.067.

k	1	2	3	4	...	12	13	14	...	17	18
$P(X \leq k)$	0.002	0.011	0.037	0.090	...	0.941	0.970	0.985	...	0.999	1

Table de la loi binomiale : on donne pour différentes valeurs de k, les valeurs de $P(X \leq k)$ lorsque X suit une loi binomiale de paramètres 120 et 0.1.

k	3	4	...	12	13	14	...	17	18
$P(X \leq k)$	0.002	0.006	...	0.576	0.687	0.782	...	0.947	0.970

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.95	0.975
$t_{48,\alpha}$	1.677	2.011
$t_{49,\alpha}$	1.677	2.010
$t_{50,\alpha}$	1.676	2.009

Table de la loi du Khi-Deux : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que $P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{7,\alpha}$	1.690	2.167	14.067	16.013
$k_{8,\alpha}$	2.180	2.733	15.507	17.535
$k_{9,\alpha}$	2.700	3.325	16.919	19.023
$k_{118,\alpha}$	89.827	93.918	144.354	149.957
$k_{119,\alpha}$	90.700	94.811	145.461	151.084
$k_{120,\alpha}$	91.573	95.705	146.567	152.211

ANNEXE 2 : La loi de Benford

Extrait d'un article de la Revue Française de Comptabilité no 321,
Ecrit par X. et R. Labouze

A l'époque où les calculatrices n'existaient pas, les calculs se faisaient à la main, à l'aide de tables. Un jour de 1881, un astronome américain, Simon Newcomb, s'aperçut que les premières pages des tables de logarithmes étaient plus usées que les autres. Se pouvait-il que les données recherchées dans cette table commençaient plus souvent par le chiffre "1" ? Il tenta de résumer les résultats de son observation dans une formule simple pour mesurer la fréquence d'apparition du premier chiffre C, celui situé le plus à gauche, dans un ensemble de données : la fréquence du premier chiffre C est égale à $\log_{10}(1 + 1/C)$.
A l'époque, cette formule ne convainquit personne.

Cinquante plus tard, vers 1938, un physicien américain, Frank Benford, redécouvrit les mêmes fréquences que celles résultant de l'application de la formule de Newcomb, en répertoriant plus de 20000 données sélectionnées dans des domaines aussi divers que les longueurs de plus de 300 fleuves, les recensements démographiques de plus de 3000 régions, les masses atomiques des éléments chimiques, les cours de bourse, les constantes de la physique, les couvertures de journaux, etc.
Il constata donc que le premier chiffre était un "1" prés d'une fois sur trois ! Il en fit une loi, généralisant la formule de Newcomb, qui porte aujourd'hui son nom : la loi de Benford.

Ce n'est qu'en 1996 que Terence Hill démontra mathématiquement la loi de Benford, celle-ci ne s'appliquant qu'aux résultats de mesure.

Correction

Partie I : Test sur une probabilité

1. On a considéré $X = (X_1, \dots, X_{120})$ un 120-échantillon de la loi de Bernoulli de paramètre θ, modélisant l'apparition du chiffre 6 en première position sur 120 montants relevés au hasard dans la zone de comptabilité considérée : pour $i = 1 \dots 120$, $X_i = 1$ si le i-ème montant commence par 6, 0 sinon. Soit $x = (x_1, \dots, x_{120})$ l'observation de cet échantillon, $\mathcal{X} = \{0,1\}^{120}$, $\mathcal{A} = \mathcal{P}(\{0,1\}^{120})$. Pour $\theta \in \Theta =]0,1[$, on note $P_\theta = \mathcal{B}(\theta)^{\otimes 120}$. Le modèle statistique considéré est alors représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$. Cette modélisation implique qu'on suppose les apparitions du chiffre 6 en première position dans les 120 montants de même loi, et indépendantes, ce qui peut, dans le cas d'une même zone de comptabilité, être discuté...
2. Par ce choix d'hypothèses, le contrôleur souhaite se prémunir en priorité du risque de déclarer que la zone étudiée est falsifiée alors qu'elle ne l'est pas.
3. On considère la statistique de test $T(X) = \sum_{i=1}^{120} X_i$ correspondant au nombre de montants commençant par un 6 parmi 120 montants relevés au hasard dans la zone de comptabilité étudiée. Lorsque $\theta = 0.067$, $T(X)$ suit une loi binomiale de paramètres $(120, 0.067)$. On sait que $T(X)/120$ est un estimateur sans biais, asymptotiquement normal de θ, donc de grandes valeurs de $T(X)$ favorisent plutôt le choix de (H_1). Par conséquent, la région critique d'un test intuitif de (H_0) contre (H_1) est de la forme $\mathcal{R}_{(H_0)} = \{x,\ T(x) \geq s\}$. La valeur critique s doit vérifier l'inéquation du niveau : $P_{0.067}(\{x,\ T(x) \geq s\}) \leq 0.05$, avec $s \in \mathbb{N}$, ce qui est équivalent à $P_{0.05}(\{x,\ T(x) \leq s-1\}) \geq 0.95$. Or lorsque $X \sim P_{0.067}$, $T(X) \sim \mathcal{B}(120, 0.067)$, donc on choisit $s - 1 = 13$, c'est-à-dire $s = 14$.
4. Le niveau exact du test est donné par $P_{0.067}(\{x,\ T(x) \geq 14\}) = 1 - P_{0.067}(\{x,\ T(x) \leq 13\}) = 1 - 0.97 = 0.03$. La puissance du test en $\theta = 0.1$ est donnée par $P_{0.1}(\{x,\ T(x) \geq 14\}) = 1 - P_{0.1}(\{x,\ T(x) \leq 13\}) = 1 - 0.687 = 0.313$.
5. Ici, $T(x) = 18 \geq 14$ donc on rejette (H_0) au profit de (H_1) pour un niveau 5%.
6. La p valeur du test est donnée par $p(x) = P_{0.067}(\{x,\ T(x) \geq 18\}) = 1 - P_{0.067}(\{x,\ T(x) \leq 17\}) = 1 - 0.999 = 0.001$. On a très clairement $p(x) < 0.05$ donc on rejette bien là aussi (H_0) au profit de (H_1) pour un niveau 5%.
7. La taille du test intuitif construit ci-dessus est de $3\% < 5\%$. Or un test uniformément plus puissant de (H_0) contre (H_1) parmi les tests de niveau 5% ne peut être que de taille exactement 5%.
8. Le modèle considéré est dominé par la mesure de comptage sur $\mathcal{X}$ et sa vraisemblance est donnée par $L(x,\theta) = \theta^{\sum_{i=1}^{120} x_i}(1-\theta)^{120-\sum_{i=1}^{120} x_i}$. Pour $0 < \theta < \theta' < 1$,

$$\frac{L(x,\theta')}{L(x,\theta)} = \left(\frac{\theta'}{\theta}\right)^{\sum_{i=1}^{120} x_i} \left(\frac{1-\theta'}{1-\theta}\right)^{120-\sum_{i=1}^{120} x_i} = h_{\theta,\theta'}(T(x)),$$

avec $h_{\theta,\theta'}$ strictement croissante. Le modèle considéré est donc à rapport de vraisemblance croissant. D'après un corollaire du lemme de Neyman-Pearson, il existe un test uniformément plus puissant parmi les tests de niveau 5% de (H_0) contre (H_1) de la forme :

$$\Phi(x) = \begin{cases} 1 & \text{si } T(x) > k \\ c & \text{si } T(x) = k \\ 0 & \text{si } T(x) < k \end{cases},$$

avec $E_{0.067}[\Phi] = P_{0.067}(\{x,\ T(x) > k\}) + cP_{0.067}(\{x,\ T(x) = k\}) = 0.05$.

On choisit k tel que $P_{0.067}(\{x,\ T(x) > k\}) \leq 0.05$ c'est-à-dire $P_{0.067}(\{x,\ T(x) \leq k\}) \geq 0.95$. On prend donc $k = 13$. Ensuite, on résout l'équation précédente et on obtient pour c :

$$c = \frac{0.05 - P_{0.067}(\{x,\ T(x) > 13\})}{P_{0.067}(\{x,\ T(x) = 13\})} = \frac{0.05 - 1 + P_{0.067}(\{x,\ T(x) \leq 13\})}{P_{0.067}(\{x,\ T(x) \leq 13\}) - P_{0.067}(\{x,\ T(x) \leq 12\})} = 0.69.$$

On obtient donc finalement le test suivant :

$$\Phi(x) = \begin{cases} 1 & \text{si } T(x) > 13 \\ 0.69 & \text{si } T(x) = 13 \\ 0 & \text{si } T(x) < 13 \end{cases}.$$

9. La puissance de ce nouveau test en $\theta = 0.1$ est donnée par $E_{0.1}[\Phi] = P_{0.1}(\{x,\ T(x) > 13\}) + 0.69P_{0.1}(\{x,\ T(x) = 13\}) = 1 - 0.687 + 0.69(0.687 - 0.576) = 0.39$, donc la puissance du test est supérieure à celle du test intuitif précédent.
10. La conclusion de ce nouveau test avec $T(x) = 18$ reste inchangée.

Partie II : Test du Khi-Deux d'adéquation

1. On note pour $k = 1, \ldots, 9$, N_k le nombre de montants commençant par le chiffre k dans x, et p_k la probabilité que C soit égal à k lorsque C est une variable aléatoire modélisant le premier chiffre d'un nombre issu de résultats de mesure suivant la loi de Benford. La statistique du test du Khi-Deux d'adéquation est donnée par :

$$T(x) = \sum_{k=1}^{9} \frac{(N_k - 120p_k)^2}{120p_k}.$$

2. La fonction de test correspondante est de la forme : $\Phi(x) = \mathbb{1}_{T(x) \geq s}$, avec s égale au 0.95 quantile de la loi du Khi-Deux à $9 - 1 = 8$ degrés de liberté, c'est-à-dire $s = 15.507$.
3. On a les résultats suivants :

k	1	2	3	4	5	6	7	8	9
N_k	12	10	12	13	17	18	16	12	10
p_k	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
$120p_k$	36.12	21.12	15.00	11.64	9.48	8.04	6.96	6.12	5.52

On trouve après calculs $T(x) = 62.05 \geq 15.507$ (on a déjà $(N_1 - 120p_1)^2/120p_1 = 16.11 \geq 15.507$), donc pour un niveau asymptotique 5%, on rejette l'hypothèse selon laquelle le premier chiffre d'un montant relevé au hasard dans la zone de comptabilité étudiée suit la loi de Benford.

Partie III : Durée de contrôles fiscaux

1. Modèle statistique. Soit $D = (D_1, \ldots, D_{50})$ un 50-échantillon de la loi gaussienne d'espérance m inconnue, de variance σ^2 inconnue, modélisant les durées de 50 contrôles fiscaux après la mise en place du dispositif. Soit $\mathcal{X} = \mathbb{R}^{50}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{50})$, $\Theta = \mathbb{R}_+ \times \mathbb{R}_+^*$, et

pour $\theta = (m, \sigma^2) \in \Theta$, $P_\theta = \mathcal{N}(m, \sigma^2)^{\otimes 50}$. Le modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$.

2. hypothèses. On teste $(H_0) : m = 55$ contre $(H_1) : m < 55$. On souhaite ici se prémunir en priorité du risque de déclarer que la durée d'un contrôle fiscal a diminué avec la mise en place du dispositif alors que ce n'est pas le cas.

3. Statistique de test. On introduit la statistique suivante :

$$T(d) = \sqrt{50}\frac{\bar{d} - \mathbf{55}}{\sqrt{S^2(d)}}.$$

Sous (H_0), $T(D)$ suit une loi de Student à 49 degrés de liberté.

4. Région de rejet du test de niveau 5% : $R_{(H_0)} = \{d, \mathbf{T(d)} \leq \mathbf{-1.677}\}$.

5. Conclusion. Ici, $T(d) = -2.1$ donc on rejette (H_0) au profit de (H_1) pour un niveau 5%.

7.2 Dernière pensée pour Gregory House

Sujet d'examen, année universitaire 2010-2011, durée 2h30

Le problème suivant se compose de quatre parties et d'une conclusion. Des extraits de tables statistiques sont donnés à la fin du sujet en Annexe.

La sixième saison de la série télévisée américaine *House* est diffusée en France sur TF1 depuis le 19 avril dernier. Au Etats-Unis, les magazines de télévision ont noté une chute d'audience pour cette sixième saison diffusée en 2009-2010 par rapport à la cinquième diffusée en 2008-2009.
On s'intéresse de prés aux audiences des Etats-Unis et on mène ici différentes études statistiques.

La saison 5 contient 24 épisodes et la saison 6 en contient 22. On a relevé les audiences (en millions de téléspectateurs) de ces différents épisodes.

Pour la saison 5, on a les résultats suivants :

Episode	501	502	503	504	505	506	507	508	509	510	511	512
Audience	14.77	12.38	12.98	13.27	13.09	13.50	13.06	13.26	12.88	12.52	14.05	15.03
Episode	513	514	515	516	517	518	519	520	521	522	523	524
Audience	15.69	14.87	14.20	14.86	12.38	13.13	12.51	13.29	12.19	11.69	12.05	12.74

Pour la saison 6, on a les résultats suivants :

Episode	601	602	603	604	605	606	607	608	609	610	611
Audience	16.50	16.50	14.44	13.74	13.50	11.65	13.31	12.67	11.95	13.25	12.24
Episode	612	613	614	615	616	617	618	619	620	621	622
Audience	14.21	13.38	13.60	12.82	11.37	10.80	10.82	10.85	9.98	9.48	11.06

On note $Y = (Y_1, \ldots, Y_{24})$ la variable aléatoire modélisant les audiences des 24 épisodes de la saison 5, et $y = (y_1, \ldots, y_{24})$ l'observation de cette variable, puis $Z = (Z_1, \ldots, Z_{22})$ la variable aléatoire modélisant les audiences des 22 épisodes de la saison 6, et $z = (z_1, \ldots, z_{22})$ l'observation de cette variable.

Partie I : Comparaison d'échantillons gaussiens

On suppose que Y est un 24-échantillon d'une loi gaussienne $\mathcal{N}(m_1, \sigma_1^2)$ et que Z est un 22-échantillon d'une loi gaussienne $\mathcal{N}(m_2, \sigma_2^2)$. On suppose également ici que les audiences des deux saisons sont indépendantes, c'est-à-dire que les échantillons Y et Z sont indépendants.

On note que $\sum_{i=1}^{24} y_i = 320.39$, $\sum_{i=1}^{24} y_i^2 = 4303.229$, $\sum_{i=1}^{22} z_i = 278.12$, $\sum_{i=1}^{22} z_i^2 = 3588.408$.

Pour juger d'une éventuelle réévaluation des tarifs de diffusion d'un spot publicitaire pendant les épisodes de la saison 6 (par rapport aux tarifs de la saison 5) en France, on se base sur les audiences américaines.
On souhaite tester l'hypothèse (H_0) "$m_1 \leq m_2$" contre l'alternative (H_1) "$m_1 > m_2$".

1. Que signifie ce choix d'hypothèses ? Quel point de vue adopte-t-on, celui des annonceurs publicitaires ou celui de TF1 qui diffuse la saison 6 ?

2. Montrer à l'aide d'un test statistique (intuitif) de niveau 5% que l'on ne peut raisonnablement pas supposer que $\sigma_1^2 = \sigma_2^2$.

3. Construire un test asymptotique de (H_0) "$m_1 \leq m_2$" contre (H_1) "$m_1 > m_2$" de niveau 5%. Quelle est la conclusion du test ? Quel est l'inconvénient de ce test ?

Partie II : Tests sur l'espérance et la variance en modèle gaussien

Le test précédent n'étant pas satisfaisant, on décide de considérer un autre modèle. On introduit pour $i = 1, \ldots, 22$, la variable aléatoire $X_i = Y_i - Z_i$ modélisant la différence entre l'audience du $i$ème épisode de la saison 5 et celle du $i$ème épisode de la saison 6 (on ne considère ici que les 22 premiers épisodes de la saison 5). On suppose que $X = (X_1, \ldots, X_{22})$ est un 22-échantillon d'une loi gaussienne $\mathcal{N}(m, \sigma^2)$, et on note $x = (x_1, \ldots, x_{22})$ avec pour $i = 1, \ldots, 22$, $x_i = y_i - z_i$.

On souhaite tester sur la base de l'observation x l'hypothèse (H_0) "$m \leq 0$" contre l'alternative (H_1) "$m > 0$".

On note que $\sum_{i=1}^{22} x_i = 17.48$ et $\sum_{i=1}^{22} x_i^2 = 81.986$.

1. Montrer, à l'aide d'un test intuitif de niveau 5% que l'on peut supposer : $\sigma^2 = 3$.
Le test construit est-il uniformément plus puissant parmi les tests de niveau 5% ?

2. On suppose donc désormais que $\sigma^2 = 3$.
a) Montrer que le modèle statistique considéré est à rapport de vraisemblance monotone.
b) En déduire la construction détaillée d'un test uniformément plus puissant parmi les tests de niveau 5% de (H_0) contre (H_1).
c) Quelle est la conclusion de ce test ?
d) A l'aide de la table de la loi gaussienne 2 fournie en Annexe, donner un encadrement de la p-valeur du test. Qu'en conclut-on pour un niveau 1% ?

3. Quel test peut-on mettre en œuvre si l'on ne veut plus faire l'hypothèse que $\sigma^2 = 3$? Comparer ce test au test asymptotique de la partie I.

Partie III : Test d'adéquation du Khi-Deux

On considère toujours pour $i = 1, \ldots, 22$, la variable aléatoire $X_i = Y_i - Z_i$ modélisant la différence entre l'audience du $i$ème épisode de la saison 5 et celle du $i$ème épisode de la saison 6. On suppose que $X = (X_1, \ldots, X_{22})$ est un 22-échantillon d'une loi inconnue, et on note $x = (x_1, \ldots, x_{22})$ avec pour $i = 1, \ldots, 22$, $x_i = y_i - z_i$. Afin de valider (ou non) le modèle gaussien posé dans la partie II, on a recours à un test non paramétrique du Khi-Deux d'adéquation.

1. *Question préliminaire.* Détermination des effectifs sous l'hypothèse gaussienne.
a) Donner les estimateurs empiriques classiques $\hat{m}$ et $\hat{\sigma}^2$ de l'espérance et la variance de la loi des X_i et calculer les valeurs de ces estimateurs. On note $\hat{m}_{obs}$ et $\hat{\sigma}^2_{obs}$ les valeurs trouvées.
b) Montrer à l'aide de la table de la loi gaussienne 3 fournie en Annexe que si $\hat{P}$ désigne la loi gaussienne d'espérance $\hat{m}_{obs}$ et de variance $\hat{\sigma}^2_{obs}$, alors $\hat{P}(]-\infty, -0.45]) \simeq 0.245$, $\hat{P}(]-0.45, 0.85]) \simeq 0.267$, $\hat{P}(]0.85; 2]) \simeq 0.237$ et $\hat{P}(]2, +\infty[) \simeq 0.251$.

2. On note que l'observation $(x_{(1)}, \dots, x_{(22)})$ de la statistique d'ordre associée à $x = (x_1, \dots, x_{22})$ est égale à $(-4.12, -1.73, -1.46, -0.73, -0.47, -0.41, -0.25, 0.59, 0.63, 0.82, 0.93, 1.27, 1.38, 1.58, 1.66, 1.81, 1.85, 2.31, 2.31, 2.71, 3.31, 3.49)$.
On souhaite tester l'hypothèse que X est bien un 22-échantillon d'une loi gaussienne à l'aide d'un test du Khi-Deux de niveau asymptotique 5%, en considérant la partition de $\mathbb{R}$ composée des intervalles suivants : $]-\infty, -0.45]$, $]-0.45, 0.85]$, $]0.85; 2]$ et $]2, +\infty[$.
a) Donner la statistique de test, ainsi que la fonction de test.
b) Quelle est la conclusion du test ?
c) Que peut-on penser de la qualité de ce test ?

Partie IV : Test de normalité de Kolmogorov-Smirnov / Lilliefors

On cherche finalement à valider (ou non) le modèle gaussien posé dans la partie II, non plus à l'aide d'un test du Khi-Deux d'adéquation, mais à l'aide d'un test inspiré du test de Kolmogorov-Smirnov (test de Lilliefors, 1967).

Le problème de test considéré est donc (H_0) : X est un 22-échantillon d'une loi gaussienne contre (H_1) : X est n'est pas un 22-échantillon d'une loi gaussienne.

On rappelle que les estimateurs empiriques de l'espérance et la variance de la loi des X_i sont notés $\hat{m}$ et $\hat{\sigma}^2$ (c.f. question 1. a) de la partie III précédente).
On note alors $F_{(\hat{m},\hat{\sigma}^2)}$ la fonction de répartition de la loi gaussienne d'espérance $\hat{m}$ et de variance $\hat{\sigma}^2$ et on introduit F_{22} la fonction de répartition empirique associée au 22-échantillon $X = (X_1, \dots, X_{22})$.

1. Justifier l'introduction du test $\phi(x) = \mathbb{1}_{L(x)\geq s}$, où

$$L(X) = \sup_{t\in\mathbb{R}} \left| F_{(\hat{m},\hat{\sigma}^2)}(t) - F_{22}(t) \right|,$$

en se basant sur l'expression du test de Kolmogorov-Smirnov d'adéquation.

2. Expliquer pourquoi les tables de la loi de Kolmogorov-Smirnov ne peuvent pas exactement être utilisées ici.

3. On souhaite maintenant montrer que la loi de la statistique de test $L(X)$ sous l'hypothèse (H_0), appelée loi de Lilliefors, ne dépend pas des paramètres inconnus (i.e. l'espérance et la variance) de la loi des X_i.
a) Pour cela, montrer dans un premier temps que

$$\begin{aligned} L(X) &= \sup_{t\in\mathbb{R}} \left| F_{(\hat{m},\hat{\sigma}^2)}(\hat{m} + t\hat{\sigma}) - F_{22}(\hat{m} + t\hat{\sigma}) \right| \\ &= \sup_{t\in\mathbb{R}} |F(t) - G_{22}(t)|, \end{aligned}$$

où F est la fonction de répartition de la loi gaussienne centrée réduite et G_{22} est la fonction de répartition empirique associée au vecteur $(V_1, \dots, V_{22})$, défini par $V_i = (X_i - \hat{m})/\hat{\sigma}$ pour tout $i = 1 \dots 22$.
b) Montrer dans un deuxième temps que sous (H_0), la loi de G_{22} ne dépend pas des paramètres inconnus, et conclure.

4. Un extrait de la table de la loi de Lilliefors est donné en Annexe. En déduire la constante s lorsque le niveau du test est choisi égal à 5%.

5. On a ici, en notant $x = (x_1, \ldots, x_{22})$ l'observation de X, $L(x) = 0.1366$. Quelle est la conclusion du test ϕ ?

Conclusion

Dresser un bilan des différentes études menées, en donnant éventuellement de nouvelles pistes pour en mener d'autres plus pertinentes.

ANNEXE : Extraits des tables statistiques

Table de la loi gaussienne centrée réduite : on donne pour différentes valeurs de $\alpha \in [0, 1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0, 1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi gaussienne centrée réduite : on donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0, 1)$.

q	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3
$P(N \leq q)$	0.977	0.982	0.986	0.989	0.992	0.994	0.995	0.997	0.997	0.998	0.999

Table de la loi gaussienne centrée réduite : on donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0, 1)$.

q	0.01	0.02	0.03	0.04	0.05	0.65	0.66	0.67	0.68	0.69	0.7
$P(N \leq q)$	0.504	0.508	0.512	0.516	0.52	0.742	0.745	0.749	0.752	0.755	0.758

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0, 1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.95	0.975
$t_{21,\alpha}$	1.721	2.08
$t_{22,\alpha}$	1.717	2.074
$t_{23,\alpha}$	1.714	2.069

Table de la loi du Khi-Deux : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que $P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{1,\alpha}$	0.001	0.004	3.841	5.024
$k_{2,\alpha}$	0.051	0.103	5.991	7.378
$k_{3,\alpha}$	0.216	0.352	7.815	9.348
$k_{4,\alpha}$	0.484	0.711	9.488	11.143
$k_{21,\alpha}$	10.283	11.591	32.671	35.479
$k_{22,\alpha}$	10.982	12.338	33.924	36.781
$k_{23,\alpha}$	11.689	13.091	35.172	38.076

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{22,20,\alpha}$	0.419	0.483	2.102	2.434
$f_{22,21,\alpha}$	0.421	0.486	2.073	2.394
$f_{22,22,\alpha}$	0.424	0.488	2.048	2.358
$f_{23,20,\alpha}$	0.424	0.488	2.092	2.420
$f_{23,21,\alpha}$	0.427	0.491	2.063	2.380
$f_{23,22,\alpha}$	0.430	0.494	2.038	2.344
$f_{24,20,\alpha}$	0.430	0.493	2.082	2.408
$f_{24,21,\alpha}$	0.433	0.496	2.054	2.368
$f_{24,22,\alpha}$	0.436	0.499	2.028	2.331

Table de Kolmogorov-Smirnov : pour différentes valeurs de n, on donne $q_{0.95}$ tel que $P(\sup_{t\in[0,1]} |F_{U,n}(t) - t| \leq q_{n,0.95}) = 0.95$ (sans la racine carrée $\sqrt{n}$), lorsque $F_{U,n}$ est la fonction de répartition empirique associée à un n-échantillon de la loi uniforme sur $[0,1]$.

n	20	21	22
$q_{n,0.95}$	0.2941	0.2872	0.2809

Table de Lilliefors : pour différentes valeurs de n, on donne $q_{0.95}$ tel que $P(L(X_1, \ldots, X_n) \leq q_{n,0.95}) = 0.95$, lorsque $L(X_1, \ldots, X_n)$ est la statistique du test de Lilliefors.

n	20	21	22
$q_{n,0.95}$	0.1920	0.1881	0.1840

Correction

Partie I

1. On privilégie l'hypothèse (H_0) que l'audience de la saison 6 n'a pas été moins bonne que celle de la saison 5 en moyenne, dans le sens où l'on préfère se dire qu'elle n'a pas été moins bonne tant qu'on n'a pas suffisamment de preuves pour affirmer qu'elle l'a été. On se prémunit ici en priorité du risque de déclarer que l'audience de la saison 6 a été moins bonne que celle de la saison 5 à tort. On adopte plutôt le point de vue de TF1 qui ne souhaite pas voir ses tarifs de diffusion d'un spot publicitaire diminués sans raison valable.

2. Modèle statistique : Soit $X = (Y, Z)$, où Y est la variable aléatoire modélisant les audiences des 24 épisodes de la saison 5 et Z la variable aléatoire modélisant les audiences des 22 épisodes de la saison 6. On note $x = (y, z)$ l'observation de X. Soit $\mathcal{X} = \mathbb{R}^{46}$, et $\mathcal{A} = \mathcal{B}(\mathbb{R}^{46})$. Pour $\Theta = (\mathbb{R}_+^*)^4$ et $\theta = (m_1, \sigma_1^2, m_2, \sigma_2^2) \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{N}(m_1, \sigma_1^2)^{\bigotimes 24} \bigotimes \mathcal{N}(m_2, \sigma_2^2)^{\bigotimes 22}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.
On teste (H_0)' "$\sigma_1^2 = \sigma_2^2$" contre (H_1') "$\sigma_1^2 \neq \sigma_2^2$".
Statistique de test : $F(x) = \frac{S^2(y)}{S^2(z)}$, où $S^2(y) = \frac{1}{23}\sum_{i=1}^{24}(y_i - \bar{y})^2$ et $S^2(z) = \frac{1}{21}\sum_{i=1}^{22}(z_i - \bar{z})^2$.
Fonction de test : puisque $S^2(y)$ et $S^2(z)$ sont les estimateurs empiriques sans biais de σ_1^2 et σ_2^2, on choisit de rejeter (H_0) lorsque $F(x)$ prend de trop petites ou de trop grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{F(x)\leq s_1} + \mathbb{1}_{F(x)\geq s_2}$, avec $s_2 > s_1$.
Calcul de s_1 et s_2 : l'équation du niveau s'écrit

$$\sup_{\theta=(m_1,\sigma_1^2,m_2,\sigma_2^2)\in\Theta,\sigma_1^2=\sigma_2^2} \Big(P_\theta\left(\{x, F(x) \leq s_1\}\right) + P_\theta\left(\{x, F(x) \geq s_2\}\right)\Big) = 5\%.$$

Lorsque $\sigma_1^2 = \sigma_2^2$, on sait que $F(X) \sim \mathcal{F}(23, 21)$, donc en prenant $s_1 = 0.427, s_2 = 2.38$, l'équation du niveau est vérifiée.
La fonction de test d'égalité des variances s'écrit finalement $\phi(x) = \mathbb{1}_{F(x)\leq 0.427} + \mathbb{1}_{F(x)\geq 2.38}$.
Conclusion : ici, on a $F(x) = 0.3295$ (on peut obtenir ce résultat en notant que $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2/n$ ou en faisant les calculs directement sur l'observation des échantillons), donc pour un niveau 5%, on rejette l'hypothèse d'égalité des variances. Ainsi, on ne peut pas mettre en œuvre de test exact d'égalité des espérances.

3. Statistique de test asymptotique :

$$T(x) = \frac{\bar{y} - \bar{z}}{\sqrt{\frac{S^2(y)}{24} + \frac{S^2(z)}{22}}}.$$

Fonction de test : puisque $\bar{y}$ et $\bar{z}$ sont les estimateurs empiriques (des moments et du maximum de vraisemblance) de m_1 et m_2, on choisit de rejeter (H_0) lorsque $T(x)$ prend de grandes valeurs. La fonction de test correspondante s'écrit $\phi'(x) = \mathbb{1}_{T(x)\geq s}$.
Calcul de la constante s : lorsque $m_1 = m_2$, d'après le théorème central limite et le lemme de Slutsky, si les tailles d'échantillons sont suffisamment grandes, la loi de $T(X)$ est approchée par la loi $\mathcal{N}(0, 1)$. Donc pour un niveau asymptotique 5%, $s = 1.645$ convient. La fonction de test d'égalité des moyennes s'écrit finalement $\phi'(x) = \mathbb{1}_{T(x)\geq 1.645}$.

Conclusion : ici, on a $T(x) = 1.566$, donc on ne rejette pas l'hypothèse (H_0) au profit de (H_1). Pour un niveau de test asymptotique de 5%, on ne conclut pas à une chute d'audience significative, donc on choisit de ne pas diminuer les tarifs de diffusion d'un spot publicitaire. L'inconvénient du test est qu'il est asymptotique, c'est-à-dire construit sous la condition que les tailles d'échantillons sont suffisamment grandes. Ce n'est pas le cas ici...

Partie II

1. Test intuitif de "$\sigma^2 = 3$" contre "$\sigma^2 \neq 3$".
modèle statistique : soit $X = (X_1, \dots, X_{22})$ un 22-échantillon d'une loi $\mathcal{N}(m, \sigma^2)$, avec $(m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$, modélisant les différences d'audience entre la saison 5 et la saison 6 pour 22 épisodes de chaque saison, et $x = (x_1, \dots, x_{22})$ l'observation de cet échantillon. Pour $(m, \sigma^2) \in \Theta$, on note $P_{(m,\sigma^2)}$ la loi de X : $P_{(m,\sigma^2)} = \mathcal{N}(m, \sigma^2)^{\bigotimes 22}$. Le modèle statistique considéré est défini par $(\mathbb{R}^{22}, \mathcal{B}(\mathbb{R}^{22}), (P_{(m,\sigma^2)})_{(m,\sigma^2)\in\Theta})$.

Statistique de test et fonction de test : on prend comme statistique de test $F(x) = \sum_{i=1}^{22} \frac{(x_i - \bar{x})^2}{3}$ où $\bar{x} = \sum_{i=1}^{22} x_i/22$. On a $F(x) = 21S^2(x)/3$, où $S^2(X) = \frac{1}{21}\sum_{i=1}^{22}(X_i - \bar{X})^2$ est un estimateur sans biais de σ^2, donc on choisit de rejeter l'hypothèse "$\sigma^2 = 3$" au profit de "$\sigma^2 \neq 3$" lorsque $F(x)$ prend de trop grandes ou de trop petites valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{F(x)\leq s_1} + \mathbb{1}_{F(x)\geq s_2}$.
Calcul des constantes s_1 et s_2 pour un niveau 5% : pour que le test soit de niveau 5%, on choisit par exemple de prendre s_1 et s_2 telles que $\sup_{m\in\mathbb{R}} P_{(m,3)}(\{x, F(x) \leq s_1\}) = 0.025$ et $\sup_{m\in\mathbb{R}} P_{(m,3)}(\{x, F(x) \geq s_2) = 0.025$. Lorsque $X \sim P_{(m,3)}$, quelle que soit la valeur de m, $F(X) \sim \chi^2(21)$ donc $s_1 = 10.283$ et $s_2 = 35.479$.
La fonction de test s'écrit finalement $\phi(x) = \mathbb{1}_{F(x)\leq 10.283} + \mathbb{1}_{F(x)\geq 35.479}$.
Conclusion : on a ici $F(x) = 22.699$ donc pour un niveau 5%, on ne rejette pas l'hypothèse "$\sigma^2 = 3$" au profit de "$\sigma^2 \neq 3$".

Le test construit est un test bilatère, donc il n'est pas uniformément plus puissant parmi les tests de niveau 5%, mais puisque l'on est dans un modèle exponentiel, il est uniformément plus puissant parmi les tests sans biais au niveau 5%.

2. a) Modèle statistique : pour $m \in \mathbb{R}$, on note P_m la (nouvelle) loi de X : $P_m = \mathcal{N}(m, 3)^{\bigotimes 22}$. Le modèle statistique considéré est défini par $(\mathbb{R}^{22}, \mathcal{B}(\mathbb{R}^{22}), (P_m)_{m\in\mathbb{R}})$. Ce modèle est dominé par la mesure de Lebesgue sur $\mathbb{R}^{22}$, et sa vraisemblance est donnée par $L(x, m) = \frac{1}{\sqrt{6\pi}^{22}} e^{-\frac{1}{6}\sum_{i=1}^{22}(x_i - m)^2}$. En prenant $m_1 < m_2$, le rapport $L(x, m_2)/L(x, m_1)$ s'écrit :

$$\frac{L(x, m_2)}{L(x, m_1)} = e^{\frac{11}{3}(m_1^2 - m_2^2)} e^{\frac{m_2 - m_1}{3}\sum_{i=1}^{22} x_i}.$$

Le modèle considéré est donc à rapport de vraisemblance strictement croissant en la statistique $\varphi(x) = \sum_{i=1}^{22} x_i$.
b) D'après le théorème de Lehmann, il existe un test uniformément plus puissant parmi les tests de niveau 5% de (H_0) contre (H_1) de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \varphi(x) > k, \\ c & \text{si} \quad \varphi(x) = k, \\ 0 & \text{si} \quad \varphi(x) < k, \end{cases}$$

dont la taille, exactement égale à 5%, est atteinte en $m = 0$ i.e.

$$\sup_{m\leq 0} \mathbb{E}_m[\phi(X)] = \mathbb{E}_0[\phi(X)] = 5\%.$$

Calcul des constantes c et k : on doit trouver c et k telles que $P_0(\{x, \varphi(x) > k\}) + cP_0(\{x, \varphi(x) = k\}) = 0.05$. Si $X \sim P_0$, alors la loi de $\varphi(X)$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$, donc $P_0(\{x, \varphi(x) = k\}) = 0$. Par conséquent, on peut prendre pour c n'importe quelle valeur de $[0, 1]$, par exemple $c = 1$, de façon à obtenir un test non randomisé. L'équation précédente devient alors $P_0(\{x, \varphi(x) \geq k\}) = 0.05$ ou encore $P_0\left(\left\{x, \frac{\varphi(x)}{\sqrt{66}} \geq \frac{k}{\sqrt{66}}\right\}\right) = 0.05$. Sachant que si $X \sim P_0$, $\frac{\varphi(X)}{\sqrt{66}} \sim \mathcal{N}(0, 1)$, on obtient $k = 1.645\sqrt{66} = 13.364$. La fonction de test correspondante s'écrit alors $\phi(x) = \mathbb{1}_{\varphi(x)\geq 13.364}$.
c) On a ici $\varphi(x) = \sum_{i=1}^{22} x_i = 17.48$, donc pour un niveau de 5%, on rejette (H_0) au profit de (H_1). La conclusion diffère de celle du test de la partie précédente.
d) La p-valeur du test est égale à $p(x) = P_0(\{x, \varphi(x) \geq 17.48\}) = 1-F(17.48/\sqrt{66}) = 1-F(2.15) \in [0.014; 0.018]$. Par conséquent, pour un niveau 1%, on ne rejette pas (H_0) au profit de (H_1).

3. Modèle statistique : celui de la question 1 i.e. $(\mathbb{R}^{22}, \mathcal{B}(\mathbb{R}^{22}), (P_{(m,\sigma^2)})_{(m,\sigma^2)\in\Theta})$, $\Theta = \mathbb{R}\times\mathbb{R}_+^*$.
Statistique de test et fonction de test : on prend comme statistique de test $T(x) = \sqrt{22}\frac{\bar{x}}{\sqrt{S^2(x)}}$. $\bar{X}$ étant un estimateur sans biais de m, on choisit de rejeter (H_0) au profit de (H_1) lorsque $T(x)$ prend de trop grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{T(x)\geq s}$.
Calcul de s pour un niveau 5% : pour que le test soit de niveau 5%, on choisit s telle que $\sup_{m\leq 0} P_{(m,\sigma^2)}(\{x, T(x) \geq s\}) = 0.05$. On admet que le supremum est atteint en $m = 0$, or lorsque $X \sim P_{(0,\sigma^2)}$, quelle que soit la valeur de σ^2, $T(X) \sim \mathcal{T}(21)$ donc $s = 1.721$ convient.
La fonction de test s'écrit finalement $\phi(x) = \mathbb{1}_{T(x)\geq 1.721}$.
Conclusion : on a ici $T(x) = 2.0695$ donc pour un niveau 5%, on rejette (H_0) au profit de (H_1).

Si la taille de l'échantillon Y était égale à celle de Z, et suffisamment grande, ce test serait bien équivalent au test asymptotique de la partie I, mais à la condition que les deux échantillons Y et Z soient non corrélés, ce qui est peu probable en réalité !

Partie III

1. a) $\hat{m}_{obs} = \bar{x} = 0.795$ et $\hat{\sigma}^2_{obs} = S^2(x) = 3.24$.
b) $\hat{P}(]-\infty, -0.45]) = F((-0.45 - 0.795)/\sqrt{3.24}) \simeq 1 - F(0.69) \simeq 0.245$, où F est la fonction de répartition de la loi gaussienne centrée réduite. $\hat{P}(]-0.45, 0.85]) \simeq F(0.03) - F(-0.69) \simeq 0.267$, $\hat{P}(]0.85; 2]) \simeq F(0.67) - F(0.03) \simeq 0.237$ et $\hat{P}(]2, +\infty[) \simeq 1 - F(0.67) \simeq 0.251$.
2. On note pour $i = 1, \ldots, 4$, N_i le nombre de valeurs dans x appartenant au $i$ème intervalle de la partition choisie, et $\hat{p}_i$ la probabilité $\hat{P}$ de ce $i$ème intervalle.
a) La statistique du test est $T(x) = \sum_{i=1}^{4} \frac{(N_i - 22\hat{p}_i)^2}{22\hat{p}_i}$.
La fonction de test est de la forme : $\phi(x) = \mathbb{1}_{T(x)\geq s}$, avec s égal au 0.95 quantile de la loi du Khi-Deux à $4 - 1 - 2 = 1$ degré de liberté puisqu'on a estimé deux paramètres de la loi, c'est-à-dire $s = 3.841$.
b) On a les résultats suivants :

Intervalle de la partition	1	2	3	4
N_i	5	5	7	5
$\hat{p}_i$	0.245	0.267	0.237	0.251
$22\hat{p}_i$	5.390	5.874	5.214	5.522

On trouve après calculs $T(x) = 0.82$, donc pour un niveau asymptotique 5%, on ne rejette pas l'hypothèse gaussienne.
c) D'une part, la taille de l'échantillon ($n = 22$) et les effectifs théoriques sont faibles donc on a peu de chances d'être dans de bonnes conditions d'application du test du Khi-Deux. Le niveau du test ne sera pas de 5% et il sera peu puissant.
Le fait de n'avoir que 4 classes perd beaucoup d'information sur les données. Le test de Kolmogorov-Smirnov d'appartenance à une famille paramétrique de lois est plus approprié ici.

Partie IV

1. Puisque $L(x) = \sup_{t\in\mathbb{R}} |F_{(\hat{m}_{obs},\hat{\sigma}^2_{obs})}(t) - \frac{1}{22}\sum_{i=1}^{22} \mathbb{1}_{x_i\leq t}|$, la statistique $L(X)$ correspond précisément à la statistique du test de Kolmogorov-Smirnov qui permettrait de tester l'hypothèse "X est un 22 échantillon de la loi $\mathcal{N}(\hat{m}_{obs}, \hat{\sigma}^2_{obs})$.

2. L'introduction, dans la statistique de test, des estimateurs $\hat{m}$ et $\hat{\sigma}^2$, qui sont des variables aléatoires construites sur X, modifie nécessairement sa loi initiale sous (H_0).

3. a) Les égalités se déduisent des faits suivants : lorsque $\hat{m} + t\hat{\sigma}$ parcourt $\mathbb{R}$, t parcourt $\mathbb{R}$ également, puis, par une transformation de centrage et réduction de la loi gaussienne, $F_{(\hat{m},\hat{\sigma}^2)}(\hat{m} + t\hat{\sigma}) = F(t)$. Enfin, l'égalité $F_{22}(\hat{m} + t\hat{\sigma}) = G_{22}(t)$ se déduit de façon évidente de l'expression même de F_{22}.
b) Supposons que X est un 22 échantillon d'une loi gaussienne de paramètres inconnus (m, σ^2). On note $V_i = \frac{X_i - \hat{m}}{\hat{\sigma}}$ et on introduit $\tilde{X}_i = (X_i - m)/\sigma$. Alors on montre que

$$V_i = \frac{\tilde{X}_i - \bar{\tilde{X}}}{\sqrt{\frac{1}{21}\sum_{i=1}^{22}(\tilde{X}_i - \bar{\tilde{X}})^2}},$$

et comme la loi des $\tilde{X}_i$ est précisément la loi $\mathcal{N}(0, 1)$, la loi du vecteur V est libre de (m, σ^2). Par conséquent, la loi de G_{22} est libre de (m, σ^2) également.

4. Pour un niveau de 5%, on a $s = 0.1840$.

5. Puisqu'ici $L(x) = 0.1366$, on ne rejette pas l'hypothèse (H_0) de normalité.

Conclusion

Les tests mis en œuvre dans la partie II doivent être justifiés par un test de normalité. Le test du Khi-Deux accepte effectivement l'hypothèse de normalité, mais on n'est pas dans des conditions d'utilisation très satisfaisantes. L'introduction du test de Lilliefors est donc justifiée et valide plus sérieusement le modèle mis en place dans la partie II.
D'autres tests non paramétriques sont cependant plus puissants pour tester l'hypothèse de normalité que le test du Khi-Deux et celui de Lilliefors comme le test de Shapiro-Wilk... Et si l'on ne souhaite pas passer par la vérification d'une hypothèse gaussienne, on peut utiliser directement des tests non paramétriques de comparaison d'échantillons (ayant recours à des méthodes de bootstrap par exemple).

7.3 Vaccination contre la grippe A pandémique

Sujet d'examen, année universitaire 2009-2010, durée 2h30

Les trois parties du problème suivant peuvent être traitées de façon indépendante. Des extraits de tables statistiques sont donnés à la fin du sujet.

On s'intéresse à la campagne de vaccination contre la grippe A pandémique à la fin de l'année 2009 en France suivant plusieurs points de vue statistiques. On souhaite mener trois études : la première portant sur la teneur en l'un des composés les plus controversés d'un vaccin français, le thiomersal, la deuxième sur la vaccination des enfants, la troisiéme sur l'influence de la diffusion d'une campagne de promotion de la vaccination.

Partie I : Etude de la teneur en thiomersal d'un vaccin grippal pandémique

L'un des vaccins grippaux pandémiques produit par un laboratoire pharmaceutique français contient selon la notice par dose de 0.5 mL, 45 μg de thiomersal, conservateur permettant d'éviter la contamination bactérienne du vaccin.
On suppose ici que la teneur en thiomersal dans une dose de vaccin suit une loi normale, d'espérance m et de variance σ^2. On souhaite savoir si m est inférieure ou supérieure aux 45 μg annoncés par le laboratoire.
On mesure pour cela la teneur (en μg) en thiomersal de 16 doses de vaccin prises au hasard.

1. On suppose que σ^2 est connue et égale à 4.
a) Décrire le modèle statistique considéré.
b) On veut construire un test de (H_0) contre (H_1), avec $(H_0) : m \leq 45$ ou $m \geq 45$, et $(H_1) : m > 45$ ou $m < 45$. Quelles sont les hypothèses nulle et alternative à considérer si l'on souhaite se prémunir en priorité du risque :
   — lié à la contamination bactérienne du vaccin ?
   — lié à la toxicité du thiomersal ?

Justifier les réponses.
c) En utilisant le théorème de Lehmann, déterminer un test uniformément plus puissant parmi les tests de niveau $\alpha = 5\%$ de l'hypothèse nulle $(H_0) : m \leq 45$ contre l'alternative $(H_1) : m > 45$.
d) Les mesures effectuées donnent les résultats suivants :

45.8	44.7	47.4	44.4	46.1	45.9	44.9	46	44.2	45.7	45	39.1	44.9	45.2	45.1	48.6

Quelle est la conclusion du test ?
e) Déterminer à l'aide des tables fournies un encadrement de la puissance du test en $m_1 = 46$. Commenter le résultat.
f) Combien de mesures devrait-on effectuer si l'on voulait pouvoir construire un test uniformément plus puissant parmi les tests de niveau 5%, dont la puissance en $m_1 = 46$ serait au moins égale à 90% ?

2. On ne suppose plus la variance σ^2 connue.
a) Décrire le modèle statistique considéré.

b) Construire de façon intuitive un test de niveau $\alpha \in]0,1[$ de l'hypothèse nulle $(H_0) : m \leq 45$ contre l'alternative $(H_1) : m > 45$.
c) Quelles propriétés ce nouveau test vérifie-t-il ?
d) Sur la base des mesures relevées dans la question 1. d), donner une valeur approchée de la $p-$ valeur de ce nouveau test, et en déduire la conclusion du test pour un niveau $\alpha = 5\%$.

Partie II : Influence de l'âge des enfants sur le choix de vaccination

On se demande ici si le choix parental de vacciner ses enfants est lié ou non à l'âge des enfants. Pour cela, on relève pour 200 enfants choisis au hasard dans la population, la tranche d'âge des enfants ainsi que le nombre d'enfants vaccinés par tranche d'âge. On obtient les résultats suivants :

Tranche d'âge	moins de 9 ans	plus de 9 ans (inclus)
Vaccinés	27	7
Non vaccinés	99	67

1. A l'aide d'un test du Khi-Deux d'indépendance de niveau asymptotique 5% que l'on décrira et justifiera avec précision, déterminer si le choix parental de vaccination des enfants est lié ou non à l'âge des enfants.
2. Répondre à la même question à l'aide d'un test paramétrique de comparaison de probabilités de niveau asymptotique 5% que l'on décrira et justifiera également avec précision.
3. Comparer les deux tests ci-dessus (avantages, inconvénients respectifs ?).

Partie III : Impact de la campagne publicitaire pour la vaccination

On souhaite étudier finalement l'impact de la campagne publicitaire de promotion de la vaccination. On a ainsi noté pour 20 personnes vaccinées le temps écoulé (en jours) entre le début de la diffusion de cette campagne et la date de vaccination, et on a obtenu les résultats suivants, notés $x = (x_1, \ldots, x_{20})$:

0.9	28.1	25.3	12.3	17.8	1.3	7.5	65.2	25.7	9.1
2.8	18.9	47.4	48.9	14.9	16.1	66.3	74.9	15.1	18.7

1. On se demande si la loi P du temps écoulé entre le début de la diffusion de la campagne et la date de vaccination d'une personne vaccinée est une loi exponentielle de paramètre 0.1, ou une loi uniforme sur $[0, 100]$.
a) Montrer qu'un test de Neyman-Pearson de niveau $\alpha \in]0,1[$ de l'hypothèse $(H_0) : P$ est une loi exponentielle de paramètre 0.1 contre $(H_1) : P$ est une loi uniforme sur $[0, 100]$ a une région critique de la forme $\{x, \sum_{i=1}^{20} x_i > k\} \cap [0, 100]^{20}$.
b) On rappelle que si $(X_1, \ldots, X_n)$ est un n échantillon de la loi exponentielle de paramètre $\theta > 0, 2\theta \sum_{i=1}^{n} X_i$ suit une loi du Khi-Deux à $2n$ degrés de liberté. En déduire un test uniformément plus puissant parmi les tests de niveau 5% de (H_0) contre (H_1) et sa conclusion.
2. On veut maintenant déterminer si la loi P du temps écoulé entre le début de la diffusion de la campagne et la date de vaccination d'une personne vaccinée est une loi exponentielle

de paramètre 0.1 à l'aide d'un test de Kolmogorov-Smirnov d'adéquation basé sur les 10 premières observations seulement.
a) Décrire le test de Kolmogorov-Smirnov à considérer en donnant les arguments intuitifs et théoriques qui le justifient.
b) Donner la conclusion du test pour un niveau 5%.
c) Quel autre test non paramétrique peut-on utiliser pour résoudre la même question ? Décrire de façon succinte la démarche à adopter pour mettre en œuvre cet autre test.

Extraits des tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi gaussienne : on donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0,1)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(N \leq q)$	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(N \leq q)$	0.90	0.92	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	0.99

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{15,\alpha}$	1.341	1.753	2.131
$t_{16,\alpha}$	1.337	1.746	2.12

Table de la loi de Student : on donne pour différentes valeurs de q la valeur de $P(T \leq q)$ lorsque $T \sim \mathcal{T}(15)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(T \leq q)$	0.54	0.58	0.62	0.65	0.69	0.72	0.75	0.78	0.81	0.83	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(T \leq q)$	0.89	0.91	0.92	0.93	0.95	0.95	0.96	0.97	0.97	0.98	0.98	0.99

Table de la loi de Student : on donne pour différentes valeurs de q la valeur de $P(T \leq q)$ lorsque $T \sim \mathcal{T}(16)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(T \leq q)$	0.54	0.58	0.62	0.65	0.69	0.72	0.75	0.78	0.81	0.83	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(T \leq q)$	0.89	0.91	0.92	0.94	0.95	0.95	0.96	0.97	0.97	0.98	0.98	0.99

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{15,15,\alpha}$	0.35	0.42	2.40	2.86
$f_{16,16,\alpha}$	0.36	0.43	2.33	2.76

Table de la loi du Khi-Deux : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que $P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{1,\alpha}$	0.001	0.004	3.841	5.024
$k_{2,\alpha}$	0.051	0.103	5.991	7.378
$k_{3,\alpha}$	0.216	0.352	7.815	9.348
$k_{4,\alpha}$	0.484	0.711	9.488	11.143
$k_{5,\alpha}$	0.831	1.145	11.070	12.833
$k_{6,\alpha}$	1.237	1.635	12.592	14.449
$k_{10,\alpha}$	3.247	3.940	18.307	20.483
$k_{20,\alpha}$	9.591	10.851	31.41	34.17
$k_{40,\alpha}$	24.433	26.509	55.758	59.342

Table de Kolmogorov-Smirnov : pour différentes valeurs de n, on donne $q_{0.95}$ tel que $P(\sup_{t\in[0,1]} |F_{U,n}(t) - t| \leq q_{0.95}) = 0.95$ (sans la racine carrée $\sqrt{n}$), lorsque $F_{U,n}$ est la fonction de répartition empirique associée à un n-échantillon de la loi uniforme sur $[0,1]$.

n	10	15	30	$n > 100$
$q_{0.95}$	0.409	0.338	0.242	$1.358/\sqrt{n}$

Correction

Partie I : Etude de la teneur en thiomersal d'un vaccin grippal pandémique

1. a) Soit $X = (X_1, \dots, X_{16})$ un 16-échantillon de la loi $\mathcal{N}(m, 4)$ modélisant la teneur (en μg) en thiomersal de 16 doses de vaccin prises au hasard. On note $x = (x_1, \dots, x_{16})$ l'observation obtenue par les mesures effectuées.
Le modèle statistique considéré est représenté par $(\mathcal{X}, \mathcal{A}, (P_m)_{m\in\mathbb{R}_+})$, avec $\mathcal{X} = \mathbb{R}^{16}$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^{16})$, et pour $m \in \mathbb{R}_+$, P_m est la loi de X c'est-à-dire $P_m = \mathcal{N}(m, 4)^{\otimes 16}$.
b) Les hypothèses nulle et alternative à considérer sont les suivantes si l'on souhaite se prémunir en priorité du risque :

— lié à la contamination bactérienne du vaccin : $(H_0) : m \leq 45$ contre $(H_1) : m > 45$,
— lié à la toxicité du thiomersal : $(H_0) : m \geq 45$ contre $(H_1) : m < 45$.

Contrôler en priorité le risque lié à la contamination bactérienne du vaccin revient ici à contrôler en priorité le risque de déclarer que le vaccin contient assez de thiomersal alors qu'il n'en contient pas assez. On choisit donc d'après le principe de Neyman et Pearson comme hypothèse privilégiée $(H_0) : m \leq 45$.
c) Le modèle considéré est dominé par la mesure de Lebesgue sur $\mathbb{R}^{16}$, et sa vraisemblance est donnée par $L(x, m) = \frac{1}{\sqrt{8\pi}^{16}} e^{-\frac{1}{8}\sum_{i=1}^{16}(x_i-m)^2}$. En prenant $0 \leq m_1 < m_2$, le rapport $L(x, m_2)/L(x, m_1)$ s'écrit :

$$\frac{L(x, m_2)}{L(x, m_1)} = e^{2(m_1^2-m_2^2)} e^{\frac{1}{4}(m_2-m_1)\sum_{i=1}^{16} x_i} = e^{2(m_1^2-m_2^2)} e^{4(m_2-m_1)\bar{x}},$$

où $\bar{x} = \sum_{i=1}^{16} x_i/16$. Le modèle considéré est donc à rapport de vraisemblance strictement croissant en la statistique $\varphi(x) = \bar{x}$. D'après le théorème de Lehmann, il existe un test de (H_0) contre (H_1) de niveau 5%, uniformément plus puissant parmi les tests de niveau 5% de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \varphi(x) > k, \\ c & \text{si} \quad \varphi(x) = k, \\ 0 & \text{si} \quad \varphi(x) < k, \end{cases}$$

De plus, $\sup_{m\leq 45} P_m(\{x, \phi(x) = 1\}) = P_{45}(\{x, \phi(x) = 1\}) = 0.05$.
Cette équation du niveau est équivalente à $P_{45}(\{x, \bar{x} > k\}) + cP_{45}(\{x, \bar{x} = k\}) = 0.05$. Si $X \sim P_{45}$, alors la loi de $\bar{X}$ est absolument continue par rapport à la mesure de Lebesgue, donc $P_{45}(\{x, \bar{x} = k\}) = 0$. Par conséquent, on peut prendre pour c n'importe quelle valeur de $[0, 1]$, par exemple $c = 1$, de façon à obtenir un test non randomisé. L'équation devient alors $P_{45}(\{x, \bar{x} \geq k\}) = 0.05$ ou encore $P_{45}(\{x, 2(\bar{x} - 45) \geq 2(k - 45)\}) = 0.05$. Sachant que si $X \sim P_{45}$, $2(\bar{X} - 45) \sim \mathcal{N}(0, 1)$, on obtient $k = 45 + 1.645/2 = 45.8225$. La fonction de test correspondante s'écrit alors $\phi(x) = \mathbb{1}_{\bar{x}\geq 45.8225}$.
d) On a ici $\bar{x} = 45.1875$ donc on ne rejette pas l'hypothèse (H_0) au profit de (H_1) pour un niveau de 5%.
e) La puissance du test en 46 vaut $\gamma(46) = 1 - F(2(45.8225 - 46)) = F(0.355) \in [0.62, 0.66]$.
f) La taille n de l'échantillon observé devra vérifier $P_{45}(\{x, \sum_{i=1}^n x_i/n \geq k\}) = 0.05$, et $P_{46}(\{x, \sum_{i=1}^n x_i/n \geq k\}) \geq 0.9$, c'est-à-dire $k = 45 + 2\frac{1.645}{\sqrt{n}}$ et $k \leq 46 - 2\frac{1.282}{\sqrt{n}}$, d'où $n \geq 4(1.645 + 1.282)^2$ ou encore $n \geq 35$.
2. a) Modèle statistique : Soit $X = (X_1, \dots, X_{16})$ un 16-échantillon de la loi $\mathcal{N}(m, \sigma^2)$, avec $(m, \sigma^2) \in \Theta = \mathbb{R}_+ \times \mathbb{R}_+^*$, modélisant la teneur en thiomersal de 16 doses de vaccin prises au

hasard. $x = (x_1, \dots, x_{16})$ désigne l'observation de cet échantillon. Soit $\mathcal{X} = \mathbb{R}^{16}$ $\mathcal{A} = \mathcal{B}(\mathbb{R}^{16})$, et pour $\theta = (m, \sigma^2) \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{N}(m, \sigma^2)^{\bigotimes 16}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.
b) Statistique de test et fonction de test : on prend comme statistique de test $T(x) = 4\frac{\bar{x}-45}{S(x)}$, où $S^2(x) = \frac{1}{15}\sum_{i=1}^{16}(x_i - \bar{x})^2$. $\bar{X}$ étant l'estimateur empirique de m, on choisit de rejeter (H_0) lorsque $\bar{x}$ prend de grandes valeurs, ou lorsque $T(x)$ prend de grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{T(x)\geq s}$.
Calcul de la constante s : pour que le test soit de niveau α, on doit avoir
$\sup_{\theta\in[0,45]\times\mathbb{R}_+^*} P_\theta(\phi = 1) = \alpha$, ce qui équivaut à $\sup_{\sigma^2\in\mathbb{R}_+^*} P_{(45,\sigma^2)}(\{x, T(x) \geq s\}) = \alpha$. Lorsque $X \sim P_{(45,\sigma^2)}$, la loi de $T(X)$ ne dépend pas de σ^2 : c'est une loi de Student à 15 degrés de liberté, donc s est le quantile de niveau $(1-\alpha)$ de la loi de Student à 15 degrés de liberté.
c) Par extension du théorème de Lehmann au cas de tests avec paramètres de nuisance, ce nouveau test est sans biais, uniformément plus puissant parmi les tests de niveau α.
d) On a ici $S^2(x) = 3.885$ donc $T(x) = 0.38$, et par conséquent, la $p-$ valeur du test vaut $p = P(Y \geq 0.38)$ lorsque Y suit la loi de Student à 15 degrés de liberté i.e. $p \simeq 0.35$. Pour un niveau $5\% < p$, on ne rejette pas (H_0).

Partie II : Influence de l'âge des enfants sur le choix de vaccination

1. Soit (X, Y) un couple de v.a.r. de loi P modélisant la tranche d',ge et le statut de vaccination d'un enfant pris au hasard dans la population. On dispose de l'observation $z = ((x_1, y_1), \dots, (x_{200}, y_{200}))$ d'un 200-échantillon $Z = ((X_1, Y_1), \dots, (X_{200}, Y_{200}))$ de la loi de ce couple.
On veut tester (H_0) X et Y sont indépendantes contre (H_1) X et Y ne sont pas indépendantes, sur la base de l'observation z. On fait pour cela un test du Khi-Deux d'indépendance.
Statistique de test : $T(z) = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(\frac{N_{i,*}N_{*,j}}{200} - N_{i,j})^2}{\frac{N_{i,*}N_{*,j}}{200}}$, où

- $N_{i,j}$ est le nombre d'enfants dans la tranche d'âge i, et de statut j,
- $N_{i,*}$ est le nombre d'enfants dans la tranche d'âge i,
- $N_{*,j}$ est le nombre d'enfants de statut j.

Fonction de test : $\phi(z) = \mathbb{1}_{\{T(z)\geq s\}}$.
Calcul de la constante s : sous l'hypothèse (H_0), $T(Z)$ suit asymptotiquement la loi $\chi^2(1)$. Pour un niveau asymptotique 5%, on prend $s = 3.841$.
Ici, $T(z) = 4.73$ donc on rejette l'hypothèse d'indépendance au niveau asymptotique 5%.
2. Modèle statistique : Soit $Y = (Y_1, \dots, Y_{126})$ un 126-échantillon de la loi $\mathcal{B}(p_1)$, modélisant le statut de vaccination d'un enfant de moins de 9 ans (Y_i vaut 1 si l'enfant i est vacciné, 0 sinon), et $Z = (Z_1, \dots, Z_{74})$ un 74-échantillon de la loi $\mathcal{B}(p_2)$, modélisant le statut de vaccination d'un enfant de plus de 9 ans (Z_i vaut 1 si l'enfant i est vacciné, 0 sinon), avec $(p_1, p_2) \in \Theta = [0,1]^2$ inconnu. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$. Soit $x = (y, z)$ avec $y = (y_1, \dots, y_{126})$, $z = (z_1, \dots, z_{74})$, l'observation de ces deux échantillons indépendants. Soit $\mathcal{X} = \{0,1\}^{200}$, et $\mathcal{A}$ l'ensemble des parties de $\mathcal{X}$. Pour $\theta = (p_1, p_2) \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{B}(p_1)^{\bigotimes 126} \bigotimes \mathcal{B}(p_2)^{\bigotimes 74}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.
hypothèses : $(H_0) : p_1 = p_2$ contre $(H_1) : p_1 \neq p_2$.
Statistique de test :

$$T(x) = \frac{\bar{y} - \bar{z}}{\sqrt{\frac{126\bar{y}+74\bar{z}}{200}\left(1 - \frac{126\bar{y}+74\bar{z}}{200}\right)\left(\frac{1}{126} + \frac{1}{74}\right)}}.$$

Fonction de test : puisque $\bar{Y}$ et $\bar{Z}$ sont les estimateurs empiriques de p_1 et p_2, on rejette (H_0) pour de grandes valeurs de $|T(x)|$. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{|T(x)|\geq s}$. Ce test est asymptotique puisque la loi de $T(X)$ n'est connue qu'asymptotiquement lorsque $p_1 = p_2$.
Calcul de s pour un niveau asymptotique 5% : si $p_1 = p_2 = p$, i.e. si $X \sim P_{(p,p)}$, $T(X)$ converge en loi vers une loi $\mathcal{N}(0,1)$, donc on peut choisir $s = 1.96$. La fonction de test asymptotique s'écrit finalement $\phi(x) = \mathbb{1}_{|T(x)|\geq 1.96}$.
Conclusion : on a ici $T(x) = 2.176$, donc on rejette l'hypothèse d'égalité des probabilités p_1 et p_2 pour un niveau asymptotique 5%.
3. Le principal avantage du test paramétrique par rapport au test du Khi-Deux est qu'il peut s'étendre facilement au test unilatère de $(H_0) : p_1 \geq p_2$ contre $(H_1) : p_1 < p_2$ par exemple, qui permet de savoir dans quel sens la tranche d'âge influence le statut de vaccination. Il est aussi en général plus puissant que le test du Khi-Deux.

Partie III : Impact de la campagne publicitaire pour la vaccination

1. Le modèle statistique considéré est défini par $(\mathbb{R}_+^n, \mathcal{B}(\mathbb{R}_+^n), (P_\theta)_{\theta\in\{1,2\}})$ $(n = 20)$, P_1 étant la loi d'un n-échantillon de la loi exponentielle de paramètre 0.1, P_2 la loi d'un n-échantillon de la loi uniforme sur $[0, 100]$. Il s'agit d'un modèle non paramétrique. Il est dominé par la mesure de Lebesgue sur $\mathbb{R}_+^n$. Sa vraisemblance est donnée par $L(x, 1) = 0.1^n e^{-0.1\sum_{i=1}^n x_i}$, et $L(x, 2) = \frac{1}{100^n}\mathbb{1}_{min(x_i)\geq 0}\mathbb{1}_{max(x_i)\leq 100}$.
a) Un test de Neyman-Pearson de niveau α est de la forme

$$\phi(x) = \begin{cases} 1 & \text{si} \quad L(x,2) > kL(x,1) \\ c(x) & \text{si} \quad L(x,2) = kL(x,1) \\ 0 & \text{si} \quad L(x,2) < kL(x,1), \end{cases}$$

qui est équivalente, puisque $L(x, 1) > 0$ à

$$\phi(x) = \begin{cases} 1 & \text{si} \quad L(x,2) > kL(x,1) \text{ et } x_i \in [0,100]\ \forall i \\ c(x) & \text{si} \quad L(x,2) = kL(x,1) \text{ et } x_i \in [0,100]\ \forall i \\ 0 & \text{si} \quad L(x,2) < kL(x,1). \end{cases}$$

Lorsque pour tout i, $x_i \in [0, 100]$, $L(x, 2) > kL(x, 1)$ équivaut à $\sum_{i=1}^n x_i > k'$ donc ϕ a bien une région critique de la forme $\{x, \sum_{i=1}^{20} x_i > k'\} \cap [0, 100]^{20}$.
b) D'après le lemme fondamental de Neyman-Pearson, un test uniformément plus puissant parmi les tests de niveau 5% est un test de Neyman-Pearson de taille 5% donc de la forme ci-dessus.
Par ailleurs, $P_1(\{x, L(x, 2) = kL(x, 1)\}) = 0$ (loi continue), donc on peut choisir $c(x) = 1$ par exemple.
Enfin, $L(x, 2) < kL(x, 1)$ équivaut à $\sum_{i=1}^n x_i < k'$ ou il existe $x_i \notin [0, 100]$. ϕ s'exprime alors sous la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^n x_i \geq k' \text{ et } x_i \in [0,100]\ \forall i \\ 0 & \text{si} \quad \sum_{i=1}^n x_i < k' \text{ ou s'il existe } i,\ x_i \notin [0,100]. \end{cases}$$

Lorsque $X = (X_1, \ldots, X_n) \sim P_1$, $0.2\sum_{i=1}^n X_i \sim \chi^2(2n)$ et $P_1([0, 100]^{20}) \simeq 1$ donc $P_1(\{x, \phi(x) = 1\}) \simeq P_1(\{x, \sum_{i=1}^n x_i \geq k'\}) = 5\%$ et on peut choisir lorsque $n = 20$, $k' = 55.758/0.2 = 278.79$.
Ici, $\sum_{i=1}^n x_i = 517.2$ donc on rejette (H_0) au profit de (H_1) pour un niveau 5%.

2. Soit $X = (X_1, \ldots, X_n)$ un n échantillon ($n = 10$) d'une loi P inconnue (absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$) modélisant le temps écoulé entre le début de la diffusion de la campagne et la vaccination de 10 personnes vaccinées prises au hasard. On veut tester sur la base d'une observation x de cet échantillon l'hypothèse (H_0) "P est la loi exponentielle de paramètre 0.1" contre (H_1) "P n'est pas la loi exponentielle de paramètre 0.1". On utilise pour cela un test de Kolmogorov-Smirnov d'adéquation.
a) Statistique de test : $D_n^0(X) = \sqrt{n} \sup_{t\in\mathbb{R}_+} |F_n(t) - F^0(t)|$, où F_n est la fonction de répartition associée à X et F^0 est la fonction de répartition de la loi exponentielle de paramètre 0.1. Cette statistique s'exprime également sous la forme :

$$D_n^0(X) = \max_{1\le i\le n} \max \left\{ \left| F^0(X_{(i)}) - \frac{i}{n} \right|, \left| F^0(X_{(i)}) - \frac{i-1}{n} \right| \right\},$$

où $(X_{(1)}, \ldots, X_{(n)})$ est la statistique d'ordre associée à X.
$D_n^0(X)$ estime une distance entre la loi P et la loi exponentielle de paramètre 0.1. De plus, on sait que lorsque P n'est pas la loi exponentielle de paramètre 0.1, $D_n^0(X) \to +\infty$ p.s. et que sous (H_0), pour tout n, $D_n^0(X)$ suit la même loi que $U_n = \sqrt{n} \sup_{t\in[0,1]} |F_{U,n}(t) - t|$, où $F_{U,n}$ est la fonction de répartition associée à un n-échantillon de la loi uniforme sur $[0, 1]$. Cette loi est tabulée pour toute valeur de n (tables fournies à la fin du sujet). La fonction de test est donc de la forme $\phi(x) = \mathbb{1}_{D_n^0(x)\ge 0.409\sqrt{10}}$ pour $n = 10$ et un niveau 5%.
b) Conclusion : Pour calculer la valeur de $D_n^0(x)$, on utilise le tableau suivant.

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	0.9	1.3	7.5	9.1	12.3	17.8	25.3	25.7	28.1	65.2
$F^0(x_{(i)})$	0.086	0.122	0.528	0.597	0.708	0.831	0.920	0.923	0.940	0.999
$\lvert F^0(x_{(i)}) - i/10\rvert$	0.014	0.078	0.228	0.197	0.208	0.231	0.220	0.123	0.040	0.001
$\lvert F^0(x_{(i)}) - (i-1)/10\rvert$	0.086	0.022	0.328	0.297	0.308	0.331	0.320	0.223	0.140	0.099

On obtient $D_n^0(x) = 0.331\sqrt{10}$, donc on ne rejette pas l'hypothèse (H_0) au profit de (H_1) pour un niveau 5%.
c) On pourrait utiliser également un test du Khi-Deux d'adéquation.

7.4 Finale de Roland Garros

Sujet d'examen, année universitaire 2008-2009, durée 2h30

On veillera pour chaque test à préciser le modèle statistique et à poser les hypothèses de façon claire.

Problème I

Lors de la finale de Roland Garros en 2009, les vitesses de premier service de Robin Söderling ont été en moyenne plus élevées que celles de Roger Federer. On cherche à déterminer si elles l'ont été de façon significative d'un point de vue statistique.

Tests de comparaison en modèle gaussien

On suppose que la vitesse de premier service de Robin Söderling suit une loi gaussienne $\mathcal{N}(m_1, \sigma_1^2)$ et que celle de Roger Federer suit une loi gaussienne $\mathcal{N}(m_2, \sigma_2^2)$. On considère par ailleurs que les vitesses de premier service des deux joueurs sont indépendantes.
On relève maintenant les vitesses de 30 premiers services de Robin Söderling puis de Roger Federer. On obtient les résultats suivants (en km/h).

Pour R. Söderling :

183	209	204	219	221	189	183	206	216	205	188	181	185	209	178
194	168	194	203	214	199	199	196	198	167	181	212	207	185	217

Pour R. Federer :

193	202	184	198	178	204	195	203	215	199	222	172	170	188	172
194	190	185	199	165	192	183	187	180	165	176	198	187	187	217

1. Montrer à l'aide d'un test statistique de niveau 5% que l'on peut supposer que $\sigma_1^2 = \sigma_2^2$.
2. En déduire un test de l'hypothèse (H_0) "$m_1 \leq m_2$" contre (H_1) "$m_1 > m_2$" de niveau 5%.
3. Que signifie ce choix d'hypothèses et quelle est la conclusion du test ?

Justification du modèle paramétrique gaussien

On souhaite maintenant valider le modèle gaussien posé dans la partie précédente à l'aide de tests non paramétriques du Khi-Deux.
1. A l'aide d'un test du Khi-Deux de niveau asymptotique 5% sur les classes $]-\infty, 170[$, $[170, 180[$, $[180, 190[$, $[190, 200[$, $[200, 210[$, $[210, 220[$, $[220, +\infty[$, dire si les vitesses de premier service de Robin Söderling peuvent effectivement être considérées comme la réalisation d'un 30-échantillon de loi gaussienne $\mathcal{N}(m_1, \sigma_1^2)$ (avec les paramètres m_1 et σ_1^2 inconnus, pour lesquels on prendra soin de montrer qu'ils peuvent être estimés par $\widehat{m_1} = 197$ et $\widehat{\sigma_1^2} = 221.38$).
2. *Question facultative :* Même question pour les vitesses de premier service de Roger Federer.
3. Quel test pourrait-on faire si l'on souhaitait maintenant s'assurer de l'indépendance entre les vitesses de premier service des deux joueurs ? Expliquer en quelques lignes la démarche à suivre pour réaliser ce test.

Problème II

On considère un 15-échantillon $X = (X_1, \dots, X_{15})$ d'une loi inconnue P, dont on dispose de l'observation $x = (x_1, \dots, x_{15})$ suivante :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1.69	1.33	0.05	0.17	2.41	1	14.41	0.36	6.79	0.50	0.15	0.25	1.64	2.28	7.64

Sur la base de cette observation, on souhaite savoir si la loi P inconnue peut être considérée comme une loi log-normale de paramètres $(0, 1)$.
On rappelle ici que la densité de la loi log-normale de paramètres $(\theta, 1)$, pour $\theta \in \mathbb{R}$, est définie par

$$f_\theta(x) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\ln x - \theta)^2}{2}} \mathbb{1}_{]0,+\infty[}(x).$$

Construction d'un test paramétrique uniformément plus puissant

On suppose dans cette première partie que l'échantillon $(X_1, \dots, X_{15})$ est un échantillon de la loi log-normale de paramètres $(\theta, 1)$, pour $\theta \in \mathbb{R}$ inconnu.
1. Montrer que pour tout $i \in \{1, \dots, 15\}$, $\ln X_i$ suit une loi gaussienne $\mathcal{N}(\theta, 1)$. En déduire la loi de la variable $\sum_{i=1}^{15} \ln X_i$.
2. Montrer que le modèle statistique considéré est à rapport de vraisemblance monotone en $\varphi(x) = \sum_{i=1}^{15} \ln x_i$.
3. En déduire un test ϕ_1 uniformément plus puissant parmi les tests de niveau 2.5% de l'hypothèse nulle "$\theta = 0$" contre l'alternative "$\theta < 0$".
4. Soit ϕ_2 un test uniformément plus puissant parmi les tests de niveau 2.5% de l'hypothèse nulle "$\theta = 0$" contre l'alternative "$\theta > 0$". Le test $\phi_1 + \phi_2$ est-il un test uniformément plus puissant parmi les tests de niveau 5% de l'hypothèse nulle "$\theta = 0$" contre l'alternative "$\theta \neq 0$" ?
5. Existe-t-il un test uniformément plus puissant parmi les tests de niveau 5% de l'hypothèse nulle "$\theta \neq 0$" contre l'alternative "$\theta = 0$" ?

Test non paramétrique de Kolmogorov-Smirnov

On souhaite maintenant construire un test de Kolmogorov-Smirnov de l'hypothèse (H_0) "P est la loi log-normale de paramètres $(0, 1)$" contre (H_1) "P n'est pas la loi log-normale de paramètres $(0, 1)$".
1. Montrer que la fonction de répartition F de la loi log-normale de paramètres $(0, 1)$ vérifie

$$F(t) = \Phi(\ln t)\mathbb{1}_{]0,+\infty[}(t),$$

où Φ est la fonction de répartition de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.
2. Donner la statistique de test de Kolmogorov-Smirnov pour le problème de test considéré, et donner une expression de cette statistique en fonction de Φ et de la statistique d'ordre associée à $(x_1, \dots, x_{15})$.
3. Déterminer, en la justifiant précisément, la fonction de test de Kolmogorov-Smirnov pour un niveau 5%.
4. Quelle est la conclusion de ce test ?

Extraits des tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi gaussienne : on donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0,1)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(N \leq q)$	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(N \leq q)$	0.90	0.92	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	0.99

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{29,\alpha}$	1.311	1.699	2.045
$t_{30,\alpha}$	1.310	1.697	2.042
$t_{58,\alpha}$	1.296	1.672	2.002
$t_{59,\alpha}$	1.296	1.671	2.001
$t_{60,\alpha}$	1.296	1.671	2

Table de la loi du Khi-Deux : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que $P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{1,\alpha}$	0.001	0.004	3.841	5.024
$k_{2,\alpha}$	0.051	0.103	5.991	7.378
$k_{3,\alpha}$	0.216	0.352	7.815	9.348
$k_{4,\alpha}$	0.484	0.711	9.488	11.143
$k_{5,\alpha}$	0.831	1.145	11.070	12.833
$k_{6,\alpha}$	1.237	1.635	12.592	14.449
$k_{29,\alpha}$	16.047	17.708	42.557	45.722
$k_{30,\alpha}$	16.791	18.493	43.773	46.979

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{29,29,\alpha}$	0.476	0.537	1.861	2.101
$f_{30,30,\alpha}$	0.482	0.543	1.841	2.074

Table de Kolmogorov-Smirnov : pour différentes valeurs de n, on donne $q_{0.95}$ tel que $P(\sup_{t\in[0,1]} |F_{U,n}(t) - t| \leq q_{0.95}) = 0.95$ (sans la racine carrée $\sqrt{n}$), lorsque $F_{U,n}$ est la fonction de répartition empirique associée à un n-échantillon de la loi uniforme sur $[0,1]$.

n	10	15	30	$n > 100$
$q_{0.95}$	0.409	0.338	0.242	$1.358/\sqrt{n}$

Correction

Problème I

Tests de comparaison en modèle gaussien

Soit $Y = (Y_1, \dots, Y_{30})$ un 30-échantillon de la loi $\mathcal{N}(m_1, \sigma_1^2)$, modélisant les vitesses de premier service de R. Söderling, et $Z = (Z_1, \dots, Z_{30})$ un 30-échantillon de la loi $\mathcal{N}(m_2, \sigma_2^2)$, modélisant les vitesses de premier service de R. Federer, avec $(m_1, \sigma_1^2, m_2, \sigma_2^2) \in \Theta = (\mathbb{R}_+^*)^4$ inconnu. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$. Soit $x = (y, z)$ avec $y = (y_1, \dots, y_{30})$, $z = (z_1, \dots, z_{30})$, l'observation de ces deux échantillons indépendants. Soit $\mathcal{X} = \mathbb{R}^{60}$, et $\mathcal{A} = \mathcal{B}(\mathbb{R}^{60})$. Pour $\theta = (m_1, \sigma_1^2, m_2, \sigma_2^2) \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{N}(m_1, \sigma_1^2)^{\otimes 30} \bigotimes \mathcal{N}(m_2, \sigma_2^2)^{\otimes 30}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$.

1. On va d'abord tester l'égalité des variances i.e. (H_0) : "$\sigma_1^2 = \sigma_2^2$" contre (H_1) : "$\sigma_1^2 \neq \sigma_2^2$".
Statistique de test : $F(x) = \frac{S^2(y)}{S^2(z)}$, où $S^2(y) = \frac{1}{29}\sum_{i=1}^{30}(y_i - \bar{y})^2$ et $S^2(z) = \frac{1}{29}\sum_{i=1}^{30}(z_i - \bar{z})^2$.
Fonction de test : puisque $S^2(y)$ et $S^2(z)$ sont les estimateurs empiriques sans biais de σ_1^2 et σ_2^2, on choisit de rejeter (H_0) lorsque $F(x)$ prend de petites ou de grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{F(x) \leq s_1} + \mathbb{1}_{F(x) \geq s_2}$, avec $s_2 > s_1$.

Calcul de s_1 et s_2 : lorsque $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e. lorsque $X \sim P_{(m_1, \sigma^2, m_2, \sigma^2)}$, on sait que $F(X) \sim \mathcal{F}(29, 29)$, donc pour un niveau de test 5%, on prend $s_1 = 0.476$, $s_2 = 2.101$. La fonction de test d'égalité des variances s'écrit finalement $\phi(x) = \mathbb{1}_{F(x) \leq 0.476} + \mathbb{1}_{F(x) \geq 2.101}$.
Conclusion : ici, on a $F(x) = 221.38/210.69 = 1.05$, donc on ne rejette pas l'hypothèse d'égalité des variances. On peut mettre en œuvre le test d'égalité des moyennes, en supposant que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
2. On suppose donc que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
Statistique de test : $T(x) = \frac{\bar{y} - \bar{z}}{S(y,z)\sqrt{\frac{1}{15}}}$, où $S^2(y, z) = \frac{S^2(y) + S^2(z)}{2}$.
Fonction de test : puisque $\bar{y}$ et $\bar{z}$ sont les estimateurs empiriques de m_1 et m_2, on choisit de rejeter (H_0) lorsque $T(x)$ prend de grandes valeurs. La fonction de test correspondante s'écrit $\phi'(x) = \mathbb{1}_{T(x) \geq s}$.
Calcul de la constante s : lorsque $m_1 = m_2$, $T(X)$ suit la loi de Student à 58 degrés de liberté. Donc pour un niveau de test 5%, $s = 1.672$. La fonction de test d'égalité des moyennes s'écrit finalement $\phi'(x) = \mathbb{1}_{T(x) \geq 1.672}$.
Conclusion : ici, on a $\bar{y} = 197$ et $\bar{z} = 190$, d'où $T(x) = 1.845$, donc on rejette l'hypothèse (H_0) au profit de (H_1). Pour un niveau de test de 5%, on peut conclure que la moyenne des vitesses de premier service de R. Söderling est significativement plus élevée que celle de R. Federer.

Justification du modèle paramétrique gaussien

1. Modèle statistique : on ne suppose plus ici gaussienne la loi de la vitesse de premier service de R. Söderling (cadre non paramétrique). Soit $Y = (Y_1, \dots, Y_{30})$ un 30-échantillon d'une v.a. de loi P (totalement inconnue) modélisant les vitesses de 30 premiers services de R. Söderling. Soit y l'observation de cet échantillon. On veut tester (H_0) P est une loi normale contre (H_1) P n'est pas une loi normale. On utilise pour cela un test du χ^2 d'adéquation.

Supposant que P est la loi d'un échantillon de loi gaussienne, on commence par estimer les paramètres (m_1, σ_1^2) de la loi. On estime m_1 par $\bar{Y}$ et σ_1^2 par $S^2(Y)$. L'estimateur $(\bar{Y}, S^2(Y))$ est consistant et asymptotiquement normal (propriétés des EMV - en renormalisant $S^2(X)$). Ici, $\bar{y} = 197$, et $S^2(y) = 221.38$.
Statistique de test : les effectifs dans les classes proposées sont les suivants : 2, 1, 8, 6, 7, 5 et 1. On regroupe donc les trois premières classes et les deux dernières de façon à avoir des effectifs supérieurs à 5. On obtient ainsi 4 classes qui sont les suivantes : $]-\infty, 190[$ (classe 1), $[190, 200[$ (classe 2), $[200, 210[$ (classe 3), $[210, +\infty[$ (classe 4). La statistique de test est alors :

$$T(y) = \sum_{i=1}^{4} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i},$$

où N_i = le nombre de vitesses dans la classe i, $\hat{p}_i$ = la probabilité que Y_i appartienne à la classe i lorsque P est la loi normale de paramètres $(197, 221.38)$. On a si N est une v.a. de loi gaussienne centrée réduite : $\hat{p}_1 = P(N < (190 - 197)/\sqrt{221.38}) = P(N < -0.5) = 0.31$, $\hat{p}_2 = P(N < (200 - 197)/\sqrt{221.38}) - P(N < (190 - 197)/\sqrt{221.38}) = P(N < 0.2) - P(N < -0.5) = 0.58 - 0.31 = 0.27$, $\hat{p}_3 = P(N < (210 - 197)/\sqrt{221.38}) - P(N < (200 - 197)/\sqrt{221.38}) = P(N < 0.9) - P(N < 0.2) = 0.82 - 0.58 = 0.24$ et $\hat{p}_4 = 1 - P(N < (210 - 197)/\sqrt{221.38}) = 1 - P(N < 0.82) = 0.18$.
On a bien $30\hat{p}_i \geq 5$ pour tout i donc on conserve les classes choisies.
Fonction de test : $\phi(y) = \mathbb{1}_{T(y) \geq s}$.
Calcul de s : Lorsque P est une loi normale, $T(Y)$ suit asymptotiquement une loi du χ^2 à $4 - 1 - 2 = 1$ degré de liberté, donc on choisit pour un niveau asymptotique 5% $s = 3.841$.
Conclusion : Ici, $T(y) = 0.93$, donc on ne rejette pas (H_0) au niveau asymptotique 5%.
2. *Question facultative :* modèle statistique : soit $Z = (Z_1, \ldots, Z_{30})$ un 30-échantillon d'une v.a. de loi P (totalement inconnue) modélisant les vitesses de 30 premiers services de R. Federer. Soit z l'observation de cet échantillon. On veut tester (H_0) P est une loi normale contre (H_1) P n'est pas une loi normale à l'aide d'un test du χ^2 d'adéquation.
Supposant que P est la loi d'un échantillon de loi gaussienne, on commence par estimer les paramètres (m_2, σ_2^2) de la loi. On estime m_2 par $\bar{Z}$ et σ_2^2 par $S^2(Z)$. Ici, $\bar{z} = 190$, et $S^2(z) = 210.69$.
Statistique de test : les effectifs dans les classes proposées sont les suivants : 2, 5, 8, 9, 3, 2 et 1. On regroupe donc les deux premières classes et les trois dernières de façon à avoir des effectifs supérieurs à 5. On obtient ainsi 4 classes qui sont les suivantes : $]-\infty, 180[$ (classe 1), $[180, 190[$ (classe 2), $[190, 200[$ (classe 3), $[200, +\infty[$ (classe 4). La statistique de test est alors :

$$T(z) = \sum_{i=1}^{4} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i},$$

où N_i = le nombre de vitesses dans la classe i, $\hat{p}_i$ = la probabilité que Z_i appartienne à la classe i lorsque P est la loi normale de paramètres $(190, 210.69)$. On a : $\hat{p}_1 = P(N < -0.7) = 0.24$, $\hat{p}_2 = P(N < 0) - P(N < -0.7) = 0.5 - 0.24 = 0.26$, $\hat{p}_3 = P(N < 0.7) - P(N < 0) = 0.26$ et $\hat{p}_4 = 1 - P(N < 0.7) = 0.24$.
On a bien $30\hat{p}_i \geq 5$ pour tout i donc on conserve les classes choisies.
Fonction de test : $\phi(z) = \mathbb{1}_{T(z) \geq s}$.
Calcul de s : Lorsque P est une loi normale, $T(Z)$ suit asymptotiquement une loi du χ^2 à $4 - 1 - 2 = 1$ degré de liberté, donc on choisit pour un niveau asymptotique 5% $s = 3.841$.
Conclusion : Ici, $T(z) = 0.395$, donc on ne rejette pas (H_0) au niveau asymptotique 5%.

3. On réaliserait ici un test du Khi-Deux d'indépendance. Pour la démarche à suivre, on pourra consulter le cours ou les corrigés des exercices de TD correspondants.

Problème II

Construction d'un test paramétrique uniformément plus puissant

Soit $\mathcal{X} =]0,+\infty[^{15}$ et $\mathcal{A} = \mathcal{B}(]0,+\infty[^{15})$. Pour $\theta \in \Theta = \mathbb{R}$, on note P_θ la loi de l'échantillon $(X_1,\dots,X_{15})$ de la loi log-normale de paramètres $(\theta,1)$. Le modèle statistique considéré est défini par $(\mathcal{X},\mathcal{A},(P_\theta)_{\theta\in\Theta})$.

1. Soit h une fonction borélienne bornée. $E[h(X_i)] = \int_0^{+\infty} h(\ln x) f_\theta(x)dx = \int_{-\infty}^{+\infty} h(y)\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-\theta)^2}{2}}dy$ (changement de variables). On a donc bien $\ln X_i \sim \mathcal{N}(\theta,1)$, d'où par indépendance des X_i, $\sum_{i=1}^{15}\ln X_i \sim \mathcal{N}(15\theta,15)$.

2. Le modèle statistique considéré est dominé par la mesure de Lebesgue sur $]0,+\infty[^{15}$, et sa vraisemblance est donnée par $L(x_1,\dots,x_n,\theta) = \frac{1}{\prod_{i=1}^{15} x_i\sqrt{2\pi}^{15}}e^{-\frac{\sum_{i=1}^{15}(\ln x_i-\theta)^2}{2}}$. Soit $\theta_1 < \theta_2$. Alors

$$\frac{L(x_1,\dots,x_n,\theta_2)}{L(x_1,\dots,x_n,\theta_1)} = e^{\frac{15(\theta_1^2-\theta_2^2)}{2}}e^{(\theta_2-\theta_1)\sum_{i=1}^{15}\ln x_i}.$$

Le modèle est donc à rapport de vraisemblance strictement croissant en la statistique $\varphi(x) = \sum_{i=1}^{15}\ln x_i$.

3. D'après un corollaire du lemme de Neyman Pearson, il existe un test de l'hypothèse "$\theta = 0$" contre "$\theta < 0$" uniformément plus puissant parmi les tests de niveau 2.5% de la forme

$$\phi_1(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{15}\ln x_i < k \\ c & \text{si} \quad \sum_{i=1}^{15}\ln x_i = k \\ 0 & \text{si} \quad \sum_{i=1}^{15}\ln x_i > k, \end{cases}$$

avec $E_0[\phi_1] = P_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i < k) + cP_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i = k) = 2.5\%$. Puisque $P_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i = k) = 0$, on peut prendre $c = 1$ par exemple, d'où

$$\phi_1(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{15}\ln x_i \le k \\ 0 & \text{si} \quad \sum_{i=1}^{15}\ln x_i > k, \end{cases}$$

avec k tel que $P_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i \le k) = 2.5\%$. Lorsque $(X_1\dots,X_{15}) \sim P_0$, $\sum_{i=1}^{15}\ln X_i \sim \mathcal{N}(0,15)$ donc on choisit $k = -1.96\sqrt{15} = -7.59$.

Ici $\sum_{i=1}^{15}\ln x_i = -0.14$, on ne rejette pas (H_0) pour un niveau 2.5%.

4. Soit ϕ_2 un test uniformément plus puissant parmi les tests de niveau 2.5% de l'hypothèse "$\theta = 0$" contre "$\theta > 0$". D'après le corollaire du lemme de Neyman Pearson, ce test est de la forme

$$\phi_2(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{15}\ln x_i > k \\ c(x) & \text{si} \quad \sum_{i=1}^{15}\ln x_i = k \\ 0 & \text{si} \quad \sum_{i=1}^{15}\ln x_i < k, \end{cases}$$

avec $E_0[\phi_2] = P_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i > k) + cP_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i = k) = 2.5\%$. Puisque $P_0((x_1,\dots,x_{15}),\sum_{i=1}^{15}\ln x_i = k) = 0$, on prend $k = 1.96\sqrt{15} = 7.59$.

Le test $\phi_1 + \phi_2$ est un test bilatère de l'hypothèse "$\theta = 0$" contre "$\theta \neq 0$" de niveau 5%, mais il ne peut pas être uniformément plus puissant parmi les tests de niveau 5%, car un tel test n'existe pas !
5. Nous sommes dans un modèle exponentielle à rapport de vraisemblance croissant en $\varphi(x)$ donc il existe bien un test uniformément plus puissant parmi les tests de niveau 5% de l'hypothèse "$\theta \neq 0$" contre "$\theta = 0$". Ce test est de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{15} \ln x_i \in]-0.2428, 0.2428[\\ c_1(x) \text{ ou } c_2(x) & \text{si} \quad \sum_{i=1}^{15} \ln x_i = -0.2428 \text{ ou } 0.2428 \\ 0 & \text{si} \quad \sum_{i=1}^{15} \ln x_i \notin]-0.2428, 0.2428[. \end{cases}$$

Test non paramétrique de Kolmogorov-Smirnov

Soit $X = (X_1, \ldots, X_n)$ un n échantillon d'une loi P inconnue (absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$). On veut tester (H_0) "P est la loi log-normale de paramètres $(0, 1)$" contre (H_1) "P n'est pas la loi log-normale de paramètres $(0, 1)$". On utilise pour cela un test de Kolmogorov-Smirnov d'adéquation.
1. Si X_i suit une loi log-normale de paramètres $(0, 1)$, pour $t > 0$, $P(X_i \leq t) = P(\ln X_i \leq \ln t) = \Phi(\ln t)$, puisque $\ln X \sim \mathcal{N}(0, 1)$ (c.f. Question 1 de la partie ci-dessus).
2. Statistique de test : $D_n^0(X) = \sqrt{n} \sup_{t \in \mathbb{R}_+} |F_n(t) - F^0(t)|$, où F_n est la fonction de répartition associée à l'échantillon X et F^0 est la fonction de répartition de la loi log-normale de paramètres $(0, 1)$. Cette statistique s'exprime également sous la forme :

$$D_n^0(X) = \sqrt{n} \max_{1 \leq i \leq n} \max \left\{ \left| \Phi(\ln X_{(i)}) - \frac{i}{n} \right|, \left| \Phi(\ln X_{(i)}) - \frac{i-1}{n} \right| \right\},$$

où $(X_{(1)}, \ldots, X_{(n)})$ est la statistique d'ordre associée à X.
3. On sait que sous (H_0), pour tout n, $D_n^0(X)$ suit la même loi que $U_n = \sqrt{n} \sup_{t \in [0,1]} |F_{U,n}(t) - t|$, où $F_{U,n}$ est la fonction de répartition associée à un n-échantillon de la loi uniforme sur $[0, 1]$. Cette loi est tabulée pour toute valeur de n (tables fournies à la fin du sujet).Par ailleurs, on sait que lorsque P n'est pas la loi log-normale de paramètres $(0, 1)$, $D_n^0(X) \to +\infty$ p.s. donc la fonction de test est de la forme $\phi(x) = \mathbb{1}_{D_n^0(x) \geq 0.338 \sqrt{15}}$ pour un niveau de test 5%.

4. Conclusion : Pour calculer la valeur de $D_n^0(x)$, on utilise le tableau suivant.

i	1	2	3	4	5	6	7	8
$\ln x_{(i)}$	-3	-1.9	-1.8	-1.4	-1	-0.7	0	0.3
$\Phi(\ln x_{(i)})$	0	0.03	0.04	0.08	0.16	0.24	0.5	0.62
$\lvert\Phi(\ln x_{(i)}) - i/15\rvert$	0.067	0.103	0.16	0.187	0.173	0.16	0.033	0.087
$\lvert\Phi(\ln x_{(i)}) - (i-1)/15$	0.67	0.703	0.76	0.787	0.773	0.427	0.233	0.18
i	9	10	11	12	13	14	15	
$\ln x_{(i)}$	0.5	0.5	0.8	0.9	1.9	2	2.7	
$\Phi(\ln x_{(i)})$	0.69	0.69	0.79	0.82	0.97	0.98	1	
$\lvert\Phi(\ln x_{(i)}) - i/15\rvert$	0.09	0.023	0.057	0.02	0.103	0.047	0	
$\lvert\Phi(\ln x_{(i)}) - (i-1)/15$	0.177	0.243	0.123	0.087	0.17	0.113	0.067	

On obtient $D_n^0(x) = 0.787\sqrt{15}$, donc on rejette l'hypothèse (H_0) au profit de (H_1) au niveau 5%.

7.5 Solidarités

Sujet d'examen, année universitaire 2008-2009, durée : 1h30

Les trois parties du problème suivant peuvent être traitées de façon indépendante. Des extraits de tables statistiques sont donnés à la fin du sujet.

Des associations comme Solidarités mènent depuis de nombreuses années des campagnes publicitaires visant à sensibiliser le grand public au problème de la mortalité due à la consommation d'eau non potable dans le monde.

Partie I : Modélisation par une loi normale

En dehors de toute campagne publicitaire, on admet que le nombre annuel de décès (en millions) imputables à l'insalubrité de l'eau suit une loi normale $\mathcal{N}(8,1)$. après la mise en place des campagnes publicitaires, des experts étudient la mortalité considérée sur 15 ans et obtiennent les chiffres suivants :

$$5.6,\ 7.6,\ 7.5,\ 8.5,\ 7.3,\ 7.6,\ 9.2,\ 8.9,\ 7.4,\ 8.8,\ 8,\ 7.2,\ 7.3,\ 8.6,\ 8.6.$$

1. On admet que la mortalité annuelle après la mise en place des campagnes publicitaires suit toujours une loi normale, mais d'espérance m et de variance σ^2 inconnues.
a) Décrire le modèle statistique considéré.
b) Vérifier à l'aide d'un test (intuitif) de niveau 5% que l'on peut raisonnablement supposer que la variance σ^2 est toujours égale à 1.

2. On suppose maintenant que $\sigma^2 = 1$, et on souhaite tester l'hypothèse

$$(H_0)\ m = 8 \text{ contre } (H_1)\ m < 8.$$

a) Décrire le modèle statistique considéré, et montrer que ce modèle est à rapport de vraisemblance monotone.
b) Justifier le choix des hypothèses (H_0) et (H_1) en se basant sur le principe de Neyman et Pearson.
c) Construire un test uniformément plus puissant parmi les tests de niveau 5% de (H_0) contre (H_1).
d) Quelle est la conclusion du test ?
e) Ce test est-il en accord avec l'intuition ? Expliquer.
f) Calculer la puissance du test en $m = 7$. Commenter.
g) Pour quel niveau de test aurait-on une conclusion différente ? Quelle valeur a-t-on calculé ici ? Expliquer et commenter le résultat obtenu.
h) Si l'on souhaitait construire un test uniformément plus puissant de l'hypothèse

$$(H_0)\ m \geq 8 \text{ contre } (H_1)\ m < 8,$$

quelle serait la conclusion du test ?

Partie II : Modélisation par une loi géométrique

En dehors de toute campagne publicitaire, on admet que la probabilité que le nombre mensuel de décès imputables à l'insalubrité de l'eau dépasse 7.2 centaines de milliers de personnes est de 0.3. après la mise en place des campagnes publicitaires, les associations portent à la connaisance des experts le nombre de mois se passant avant chaque nouveau dépassement de la valeur de 7.2 centaines de milliers de décès. Les valeurs obtenues, notées $x_1, \ldots, x_{18}$, sont les suivantes :

$$5,\ 5,\ 4,\ 1,\ 1,\ 3,\ 4,\ 5,\ 4,\ 8,\ 2,\ 7,\ 1,\ 7,\ 5,\ 5,\ 2,\ 1.$$

On souhaite savoir si la probabilité θ que le nombre mensuel de décès imputables à l'insalubrité de l'eau dépasse 7.2 centaines de milliers est désormais plutôt égale à 0.2.
On admet que $x = (x_1, \ldots, x_{18})$ est l'observation d'un 18-échantillon $(X_1, \ldots, X_{18})$ d'une loi géométrique de paramètre θ.
Rappel : Si Y suit une loi géométrique de paramètre θ, $P(Y = y) = \theta(1-\theta)^{y-1}$ pour $y \in \mathbb{N}^*$.

1. Décrire le modèle statistique considéré et montrer que ce modèle statistique est à rapport de vraisemblance décroissant en la statistique $\varphi(x) = \sum_{i=1}^{18} x_i$.
2. Expliquer en quelques mots pourquoi $\varphi(X)$ suit une loi binomiale négative (dite aussi loi de Pascal) de paramètres $(18, \theta)$.
3. Construire un test uniformément plus puissant parmi les tests de niveau 5% de l'hypothèse

$$(H_0)\ \theta = 0.2 \text{ contre } (H_1)\ \theta = 0.3.$$

4. Quelle est la conclusion du test ici ?

Partie III : Test d'adéquation du Khi-Deux

On souhaite maintenant valider la modélisation précédente. Si X est une variable aléatoire modélisant le nombre de mois se passant avant chaque nouveau dépassement de la valeur de 7.2 centaines de milliers de décès, on souhaite vérifier que X suit bien une loi géométrique. On observe pour 80 dépassements les valeurs suivantes :

Nombre de mois avant dépassement	1	2	3	4	5	6	7	8	9	10	11	13	19
Nombre de fois où ce nombre est observé	18	15	14	5	9	6	4	2	2	2	1	1	1

1. Si la loi de X est une loi géométrique de paramètre θ, expliquer pourquoi on peut estimer la valeur de θ par 0.25.
2. Vérifier à l'aide d'un test du Khi-Deux d'adéquation de niveau asymptotique 5% que l'on peut raisonnablement supposer que la loi de X est une loi géométrique.

Extraits des tables statistiques

On donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

On donne pour différentes valeurs de q la valeur de $P(N \leq q)$ lorsque $N \sim \mathcal{N}(0,1)$.

q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
$P(N \leq q)$	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88
q	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
$P(N \leq q)$	0.90	0.92	0.93	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	0.99

On donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{14,\alpha}$	1.345	1.761	2.145
$t_{15,\alpha}$	1.34	1.753	2.131

On donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que $P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{5,\alpha}$	0.831	1.145	11.07	12.833
$k_{6,\alpha}$	1.237	1.635	12.592	14.449
$k_{7,\alpha}$	1.69	2.167	14.067	16.013
$k_{8,\alpha}$	2.18	2.733	15.507	17.535
$k_{9,\alpha}$	2.7	3.325	16.919	19.023
$k_{10,\alpha}$	3.247	3.94	18.307	20.483
$k_{11,\alpha}$	3.816	4.575	19.675	21.92
$k_{12,\alpha}$	4.404	5.226	21.026	23.337
$k_{13,\alpha}$	5.009	5.892	22.362	24.736
$k_{14,\alpha}$	5.629	6.571	23.685	26.119
$k_{15,\alpha}$	6.262	7.261	24.996	27.488

Pour différentes valeurs de k, on donne $P(X = k)$ lorsque X suit une loi géométrique de paramètre 0.25.

k	1	2	3	4	5	6	7	8	9	10
$P(X = k)$	0.250	0.188	0.141	0.105	0.079	0.059	0.044	0.033	0.025	0.019
k	11	12	13	14	15	16	17	18	19	20
$P(X = k)$	0.014	0.011	0.008	0.006	0.004	0.003	0.003	0.002	0.001	0.001

Pour différentes valeurs de $p \in [0, 1]$ et k, on donne $P(X \leq k)$ lorsque X suit une loi binomiale négative (ou de Pascal) de paramètres $(18, p)$.

k	$p = 0.2$	$p = 0.3$
40	0.0003	0.0320
41	0.0004	0.0414
42	0.0006	0.0526
43	0.0009	0.0658
44	0.0012	0.0812
45	0.0017	0.0986
60	0.0427	0.5486
61	0.0496	0.5813
62	0.0573	0.6131
63	0.0657	0.6438
64	0.0750	0.6732
65	0.0850	0.7012
80	0.3292	0.9469
81	0.3501	0.9538
82	0.3713	0.9599
83	0.3926	0.9652
84	0.4141	0.9699
85	0.4356	0.9741
120	0.9353	1.0000
121	0.9405	1.0000
122	0.9453	1.0000
123	0.9498	1.0000
124	0.9540	1.0000
125	0.9578	1.0000

Correction

Partie I

1. a) Modèle statistique : Soit $X = (X_1, \dots, X_{15})$ un 15-échantillon d'une loi $\mathcal{N}(m, \sigma^2)$, avec $(m, \sigma^2) \in \Theta = \mathbb{R}_+ \times \mathbb{R}_+^*$, modélisant les nombres annuels de décès (en millions) sur 15 ans après la mise en place des campagnes publicitaires, et $x = (x_1, \dots, x_{15})$ l'observation de cet échantillon. Pour $(m, \sigma^2) \in \Theta$, on note $P_{(m,\sigma^2)}$ la loi de X : $P_{(m,\sigma^2)} = \mathcal{N}(m, \sigma^2)^{\bigotimes 15}$. Le modèle statistique considéré est défini par $(\mathbb{R}^{15}, \mathcal{B}(\mathbb{R}^{15}), (P_{(m,\sigma^2)})_{(m,\sigma^2)\in\Theta})$.
b) On teste $(H_0) : \sigma^2 = 1$ contre $(H_1) : \sigma^2 \neq 1$.
Statistique de test et fonction de test : on prend comme statistique de test l'estimateur empirique sans biais de σ^2 défini par $S^2(x) = \frac{1}{14}\sum_{i=1}^{15}(x_i - \bar{x})^2$, où $\bar{x} = \sum_{i=1}^{15} x_i/15$. On choisit de rejeter (H_0) lorsque $S^2(x)$ prend de grandes ou de petites valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{S^2(x)\leq s_1} + \mathbb{1}_{S^2(x)\geq s_2}$.
Calcul des constantes s_1 et s_2 pour un niveau 5% : pour que le test soit de niveau 5%, on choisit par exemple de prendre s_1 et s_2 telles que $\sup_{m\geq 0} P_{(m,1)}(\{x, S^2(x) \leq s_1) = 0.025$ et $\sup_{m\geq 0} P_{(m,1)}(\{x, S^2(x) \geq s_2) = 0.025$. Lorsque $X \sim P_{(m,1)}$, quelle que soit la valeur de m, $14S^2(X) \sim \chi^2(14)$ donc $s_1 = 0.402$ et $s_2 = 1.866$.
La fonction de test s'écrit finalement $\mathbb{1}_{S^2(x)\leq 0.402} + \mathbb{1}_{S^2(x)\geq 1.866}$.
Conclusion : on a ici $S^2(x) = 0.852$ donc pour un niveau 5%, on ne rejette pas (H_0) au profit de (H_1).
2. a) Modèle statistique : Soit $X = (X_1, \dots, X_{15})$ un 15-échantillon de la loi $\mathcal{N}(m, 1)$, avec $m \in \Theta = \mathbb{R}_+$, modélisant les nombres annuels de décès (en millions) sur 15 ans après la mise en place des campagnes publicitaires, et $x = (x_1, \dots, x_{15})$ l'observation de cet échantillon. Pour $m \in \Theta$, on note P_m la loi de X : $P_m = \mathcal{N}(m, 1)^{\bigotimes 15}$. Le modèle statistique considéré est défini par $(\mathbb{R}^{15}, \mathcal{B}(\mathbb{R}^{15}), (P_m)_{m\in\Theta})$. Ce modèle est dominé par la mesure de Lebesgue sur $\mathbb{R}^{15}$, et sa vraisemblance est donnée par $L(x, m) = \frac{1}{\sqrt{2\pi}^{15}} e^{-\frac{1}{2}\sum_{i=1}^{15}(x_i - m)^2}$. En prenant $0 \leq m_1 < m_2$, le rapport $L(x, m_2)/L(x, m_1)$ s'écrit :

$$\frac{L(x, m_2)}{L(x, m_1)} = e^{\frac{15}{2}(m_1^2 - m_2^2)} e^{(m_2 - m_1)\sum_{i=1}^{15} x_i}.$$

Le modèle considéré est donc à rapport de vraisemblance strictement croissant en la statistique $\varphi(x) = \sum_{i=1}^{15} x_i$.
b) En faisant ce choix d'hypothèses, on privilégie l'hypothèse que les campagnes publicitaires mises en place ne sont pas efficaces, dans le sens où l'on préfére se dire qu'elles ne le sont pas tant que les chiffres de mortalité n'ont pas prouvé qu'elles le sont. On se prémunit ainsi du risque de déclarer que les campagnes de publicité sont efficaces à tort. Ce risque aura pour probabilité maximale le niveau du test par définition.
c) D'après un corollaire du lemme fondamental de Neyman-Pearson, il existe un test uniformément plus puissant parmi les tests de niveau 5% de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \varphi(x) < k, \\ c & \text{si} \quad \varphi(x) = k, \\ 0 & \text{si} \quad \varphi(x) > k, \end{cases}$$

de taille exactement 5%.
Calcul des constantes c et k : le test étant de taille 5%, on doit trouver c et k telles que

$P_8(\{x, \varphi(x) < k\}) + cP_8(\{x, \varphi(x) = k\}) = 0.05$. Si $X \sim P_8$, alors la loi de $\varphi(X)$ est absolument continue par rapport à la mesure de Lebesgue, donc $P_8(\{x, \varphi(x) = k\}) = 0$. Par conséquent, on peut prendre pour c n'importe quelle valeur de $[0,1]$, par exemple $c = 1$, de façon à obtenir un test non randomisé. L'équation précédente devient alors $P_8(\{x, \varphi(x) \leq k\}) = 0.05$ ou encore $P_8\left(\left\{x, \frac{\varphi(x)-120}{\sqrt{15}} \leq \frac{k-120}{\sqrt{15}}\right\}\right)$. Sachant que si $X \sim P_8$, $\frac{\varphi(X)-120}{\sqrt{15}} \sim \mathcal{N}(0,1)$, on obtient $k = 120 - 1.645\sqrt{15} = 113.629$. La fonction de test correspondante s'écrit alors $\phi(x) = \mathbb{1}_{\varphi(x)\leq 113.629}$.
d) On a ici $\varphi(x) = \sum_{i=1}^{15} x_i = 118.1$, donc pour un niveau de 5%, on ne rejette pas (H_0) au profit de (H_1).
e) En construisant un test de façon intuitive on aurait trouvé un test qui rejette (H_0) au profit de (H_1) lorsque l'estimateur empirique $\bar{x} = \varphi(x)/15$ est inférieur à la valeur 7.575, et qui donc serait égal au test précédent.
f) La puissance du test pour $m = 7$ est égale à $P_7(\{x, \varphi(x) \leq 113.629\}) = P_7\left(\left\{x, \frac{\varphi(x)-105}{\sqrt{15}} \leq \frac{113.629-105}{\sqrt{15}}\right\}\right) = F(2.228) = 0.987$, où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$. Cette puissance est assez importante malgré la petite taille d'échantillon considéré car les deux hypothèses $m = 8$ et $m = 7$ sont très éloignées l'une de l'autre, donc facile à distinguer.
g) On aurait obtenu une conclusion différente pour un niveau de test égal à $p(x) = P_8(\{x, \varphi(x) \leq 118.1\}) = F(-0.49) = 0.312$. On a calculé la p-valeur du test. Ici $5\% \leq 0.312$ donc on ne rejette pas (H_0) au profit de (H_1) au niveau 5%.
h) D'après le théorème de Lehmann, le test construit précédemment est aussi uniformément plus puissant parmi les tests de niveau 5% pour ce nouveau problème de test, donc la conclusion serait identique.

Partie II

1. Modèle statistique : Soit $X = (X_1, \ldots, X_{18})$ un 18-échantillon de la loi $\mathcal{G}(\theta)$, avec $\theta \in \Theta = [0,1]$, modélisant les nombres de mois s'écoulant entre chaque nouveau dépassement de la valeur de 7.2 centaines de milliers de décès après la mise en place des campagnes publicitaires, et $x = (x_1, \ldots, x_{18})$ l'observation de cet échantillon. Pour $\theta \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{G}(\theta)^{\bigotimes 18}$. Le modèle statistique considéré est défini par $(\mathbb{N}^{*18}, \mathcal{P}(\mathbb{N}^{*18}), (P_\theta)_{\theta\in\Theta})$. Ce modèle est dominé par la mesure de comptage sur $\mathbb{N}^{*18}$, et sa vraisemblance est donnée par $L(x,\theta) = \theta^{18}(1-\theta)^{\sum_{i=1}^{18} x_i - 18}$. En prenant $0 \leq \theta_1 < \theta_2 \leq 1$, le rapport $L(x,\theta_2)/L(x,\theta_1)$ s'écrit :

$$\frac{L(x,\theta_2)}{L(x,\theta_1)} = \left(\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}\right)^{18} \left(\frac{1-\theta_2}{1-\theta_1}\right)^{\sum_{i=1}^{18} x_i}.$$

Le modèle considéré est donc à rapport de vraisemblance strictement décroissant en la statistique $\varphi(x) = \sum_{i=1}^{18} x_i$.
2. Les X_i sont indépendantes entre elles, de même loi géométrique de paramètre θ, donc $\varphi(X) = \sum_{i=1}^{18} X_i$ suit bien une loi binomiale négative de paramètres $(18, \theta)$.
3. D'après le lemme fondamental de Neyman-Pearson, un test uniformément plus puissant parmi les tests de niveau 5% est de la forme :

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \varphi(x) < k, \\ c & \text{si} \quad \varphi(x) = k, \\ 0 & \text{si} \quad \varphi(x) > k, \end{cases}$$

et de taille égale à 5%.
Calcul des constantes c et k : le test étant de taille 5%, on doit trouver c et k telles que $P_{0.2}(\{x, \varphi(x) < k\}) + cP_{0.2}(\{x, \varphi(x) = k\}) = 0.05$. Si $X \sim P_{0.2}$, alors la loi de $\varphi(X)$ est la loi binomiale négative de paramètres $(18, 0.2)$, donc on peut prendre $k = 62$, et ensuite $c = (0.05 - 0.0496)/(0.0573 - 0.0496) = 0.052$. La fonction de test correspondante s'écrit alors $\phi(x) = \mathbb{1}_{\varphi(x)<62} + 0.052\mathbb{1}_{\varphi(x)=62}$.
4. Ici, $\varphi(X) = \sum_{i=1}^{18} x_i = 70$, donc on ne rejette pas (H_0) au profit de (H_1) (mais on n'a pas adopté le même point de vue cette fois).

Partie III

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon d'une v.a. de loi P modélisant le nombre de mois s'écoulant avant un dépassement de la valeur de 7.2 centaines de milliers de décès. Soit x l'observation de cet échantillon pour $n = 80$. On veut tester (H_0) P est une loi géométrique contre (H_1) P n'est pas une loi géométrique.
1. Supposant que P est une loi géométrique de paramètre θ, on sait que l'espérance de la loi P est égale à $1/\theta$. Par conséquent, un estimateur des moments de θ est donné par $1/\bar{X}$. Or ici $1/\bar{x} = 0.25$.
2. La statistique du test du Khi-Deux d'adéquation est égale à

$$T(x) = \sum_{i=1}^{k} \frac{(N_i - 80\hat{p}_i)^2}{80\hat{p}_i},$$

où N_i = le nombre de fois où le nombre de mois avant dépassement est dans la classe i, $\hat{p}_i$ = la probabilité que Y appartienne à la classe i lorsque Y suit la loi géométrique de paramètre 0.25. Le nombre de classes k est ensuite choisi de telle sorte que $80\hat{p}_i \geq 5$ et $N_i \geq 5$ pour tout i. On considère les classes suivantes

Nombre de mois avant dépassement	1	2	3	4	5	$\{6, 7\}$	≥ 8
N_i	18	15	14	5	9	10	9
$\hat{p}_i$ (table)	0.25	0.188	0.141	0.105	0.079	0.103	0.134
$80\hat{p}_i$	20	15.04	11.28	8.40	6.32	8.24	10.72

La fonction de test est de la forme : $\phi(x) = \mathbb{1}_{T(x)\geq s}$. Lorsque P est une loi géométrique, $T(X)$ suit asymptotiquement une loi du χ^2 à 5 degrés de liberté (il y a 7 classes donc dans le cas classique on aurait $7 - 1 = 6$ degrés de liberté, ici on perd un degré de liberté supplémentaire parce qu'on a dû estimer le paramètre), donc $s = 11.07$.
Puisqu'on a $T(x) = 4.02$, pour un niveau asymptotique 5%, on ne rejette pas (H_0) au profit de (H_1).

7.6 Réussite et insertion professionnelle

Sujet d'examen, année universitaire 2007-2008, durée 2h

Les trois problèmes suivants peuvent être traités de façon indépendante. Des extraits de tables statistiques sont donnés à la fin du sujet.

On veillera pour chaque test à préciser le modèle statistique et à poser les hypothèses de façon claire.

Problème I

On considère $(X_1, \dots, X_n)$ un n-échantillon de la loi d'une variable aléatoire X de densité définie par

$$f_\theta(x) = \begin{cases} \theta^2 \frac{\ln x}{x^{\theta+1}} & \text{si} \quad x > 1 \\ 0 & \text{si} \quad x \leq 1, \end{cases}$$

où θ est un paramètre réel inconnu strictement positif. Soit $(x_1, \dots, x_n)$ une observation de cet échantillon. On rappelle que la densité de la loi du χ^2 à 4 degrés de liberté, ou de la loi gamma de paramètres $(2, 1/2)$ est donnée par $g(y) = \frac{1}{4} y e^{-\frac{1}{2}y} \mathbb{1}_{[0,+\infty[}(y)$.

On souhaite tester sur la base de l'observation $(x_1, \dots, x_n)$ l'hypothèse nulle (H_0) "$\theta \leq 1$" contre l'alternative (H_1) "$\theta > 1$".

1. Montrer que la variable aléatoire $Y = 2\theta \ln X$ suit une loi du χ^2 à 4 degrés de liberté. En déduire que $2\theta \sum_{i=1}^n \ln X_i$ suit une loi du χ^2 à 4n degrés de liberté.
2. Montrer que le modèle statistique considéré est à rapport de vraisemblance strictement décroissant en la statistique $\varphi(x_1, \dots, x_n) = \sum_{i=1}^n \ln x_i$.
3. Construire un test de (H_0) contre (H_1) uniformément plus puissant parmi les tests de niveau α.
4. Quelle est la conclusion de ce test lorsque $n = 30$, $\alpha = 5\%$ et $\sum_{i=1}^n \ln x_i = 26.66$?

Questions facultatives :

5. Déterminer l'estimateur du maximum de vraisemblance du paramètre θ.
6. Quelle est la forme du test du rapport de vraisemblance maximal de niveau α pour les hypothèses (H_0) et (H_1) ? Comparer avec le test construit à la question 3.

Problème II

On souhaite déterminer si la réussite d'un étudiant à un examen de Tests Statistiques dépend de son sexe (masculin ou féminin).

Tests paramétriques

On considère dans un premier temps que la note d'une fille à l'examen de Tests Statistiques suit une loi normale $\mathcal{N}(m_1, \sigma_1^2)$ et que celle d'un garçon suit une loi normale $\mathcal{N}(m_2, \sigma_2^2)$. On observe les notes de 22 étudiants dont 12 filles et 10 garçons, et on obtient les résultats suivants.

Filles	8	7	4.5	13	13	8.5	10	12.5	11	11	15	13.5
Garçons	13.5	10	8	11.5	18.5	8.5	12.5	5.5	9.5	10.5		

1. Montrer que l'on ne rejette pas l'hypothèse d'égalité des variances σ_1^2 et σ_2^2 au niveau 5%.
2. En déduire un test de l'hypothèse (H_0) "$m_1 = m_2$" contre (H_1) "$m_1 \neq m_2$" de niveau 5%. Quelle est la conclusion de ce test ?
3. Ce test est-il uniformément plus puissant parmi les tests de niveau 5% ?

Test non paramétrique

On considère dans un deuxième temps les notes des 83 étudiants dont 23 filles et 60 garçons, et on obtient pour chaque classe de notes les effectifs suivants.

Notes	$[0, 8]$	$]8, 11]$	$]11, 13.5]$	$]13.5, 20]$
Filles	5	8	5	5
Garçons	14	18	18	10

1. Construire un test du χ^2 d'indépendance au niveau asymptotique 5%.
2. Quelle est la conclusion de ce test ?
3. Quelle critique peut-on opposer à cette conclusion ?

Problème III

On souhaite étudier le temps X (en mois) mis par un étudiant (fille ou garçon, sans distinction) diplômé de Master pour obtenir un emploi à durée indéterminée. On relève ce temps pour n jeunes diplomés.

On souhaite tester l'hypothèse (H_0) "X suit la loi exponentielle de paramètre 1/6" sur la base de ces observations.

1. Donner la statistique de test de Kolmogorov-Smirnov pour ce problème de test, ainsi qu'une expression de cette statistique facile à utiliser en pratique.
2. Expliquer pourquoi la loi de cette statistique sous l'hypothèse (H_0) peut être tabulée pour toute valeur de n.
3. Quelle est la loi asymptotique de cette statistique sous l'hypothèse (H_0) ?
4. Pour $n = 10$, on a relevé les résultats suivants.

Temps d'insertion	5.5	10.2	30.2	16	7.5	4.1	10.3	15.3	4.2	7

Quelle est la conclusion du test de Kolmogorov-Smirnov de niveau 5% ?
5. Si l'on souhaite seulement tester l'exponentialité du temps d'insertion, quel(s) test(s) non paramétrique(s) peut-on faire ? Expliquer.

Extraits des tables statistiques

Pour différentes valeurs de n, on donne $q_{n,0.95}$ et $q_{n,0.05}$ tels que $P(X \leq q_{n,0.95}) = 0.95$ et $P(X \leq q_{n,0.05}) = 0.05$ lorsque $X \sim \chi^2(n)$.

n	3	4	5	6	120
$q_{n,0.95}$	7.81	9.49	11.07	12.59	146.57
$q_{n,0.05}$	0.35	0.71	1.15	1.64	95.7

Pour $F \sim \mathcal{F}(12, 10)$, on donne pour α, q_α tel que $P(F \leq q_\alpha) = \alpha$.

α	0.025	0.05	0.95	0.975
q_α	0.3	0.36	2.91	3.62

Pour $F \sim \mathcal{F}(11, 9)$, on donne pour α, q_α tel que $P(F \leq q_\alpha) = \alpha$.

α	0.025	0.05	0.95	0.975
q_α	0.28	0.35	3.10	3.91

Pour $T \sim \mathcal{T}(20)$, on donne pour α, $q_{1-\alpha}$ tel que $P(T \leq q_{1-\alpha}) = 1 - \alpha$ (on rappelle que la loi de Student est symétrique).

α	0.01	0.025	0.05	0.1
$q_{1-\alpha}$	2.53	2.09	1.72	1.33

Pour $T \sim \mathcal{T}(22)$, on donne pour α, $q_{1-\alpha}$ tel que $P(T \leq q_{1-\alpha}) = 1 - \alpha$ (on rappelle que la loi de Student est symétrique).

α	0.01	0.025	0.05	0.1
$q_{1-\alpha}$	2.51	2.07	1.72	1.32

Table de Kolmogorov-Smirnov : pour différentes valeurs de n, on donne $q_{0.95}$ tel que $P(\sup_{t\in[0,1]} |F_{U,n}(t) - t| \leq q_{0.95}) = 0.95$ (sans la racine carrée $\sqrt{n}$), lorsque $F_{U,n}$ est la fonction de répartition empirique associée à un n-échantillon de la loi uniforme sur $[0, 1]$.

n	10	20	30	> 100
$q_{0.95}$	0.409	0.294	0.242	$1.358/\sqrt{n}$

Correction

Problème I

Soit $\mathcal{X} =]1, +\infty[^n$ et $\mathcal{A} = \mathcal{B}(]1, +\infty[^n)$. Pour $\theta \in \Theta = \mathbb{R}_+^*$, on note P_θ la loi de $(X_1, \dots, X_n)$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$.

1. Soit h une fonction borélienne bornée. $E[h(Y)] = \int_1^{+\infty} h(2\theta \ln x)\theta^2 \frac{\ln x}{x^{\theta+1}} dx = \int_0^{+\infty} h(y)\frac{y}{4}e^{-\frac{1}{2}y} dy$ (changement de variables). On a donc bien $Y \sim \chi^2(4)$, d'où $2\theta \ln X_i \sim \chi^2(4)$ et comme les X_i sont indépendantes, $2\theta \sum_{i=1}^n \ln X_i \sim \chi^2(4n)$.

2. Le modèle statistique considéré est dominé par la mesure de Lebesgue sur $]1, +\infty[^n$, et sa vraisemblance est donnée par $L(x_1, \dots, x_n, \theta) = \theta^{2n} \prod_{i=1}^n \ln x_i \left(\prod_{i=1}^n x_i\right)^{-\theta-1}$. Soit $\theta_1 < \theta_2$. Alors

$$\frac{L(x_1, \dots, x_n, \theta_2)}{L(x_1, \dots, x_n, \theta_1)} = \left(\frac{\theta_2}{\theta_1}\right)^{2n} \left(\prod_{i=1}^n x_i\right)^{\theta_1-\theta_2} = \left(\frac{\theta_2}{\theta_1}\right)^{2n} e^{(\theta_1-\theta_2)\sum_{i=1}^n \ln x_i}.$$

Le modèle est donc à rapport de vraisemblance strictement décroissant en la statistique $\varphi(x_1, \dots, x_n) = \sum_{i=1}^n \ln x_i$.

3. D'après le théorème de Lehmann, il existe un test uniformément plus puissant parmi les tests de niveau α de la forme

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^n \ln x_i < k \\ c & \text{si} \quad \sum_{i=1}^n \ln x_i = k \\ 0 & \text{si} \quad \sum_{i=1}^n \ln x_i > k, \end{cases}$$

avec $E_1[\phi] = P_1((x_1, \dots, x_n), \sum_{i=1}^n \ln x_i < k) + cP_1((x_1, \dots, x_n), \sum_{i=1}^n \ln x_i = k) = \alpha$. Puisque $P_1((x_1, \dots, x_n), \sum_{i=1}^n \ln x_i = k) = 0$, on peut prendre $c = 1$ par exemple, d'où

$$\phi(x) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^n \ln x_i \le k \\ 0 & \text{si} \quad \sum_{i=1}^n \ln x_i > k, \end{cases}$$

avec k tel que $P_1((x_1, \dots, x_n), 2\sum_{i=1}^n \ln x_i \le 2k) = \alpha$. Lorsque $(X_1 \dots, X_n) \sim P_1$, $2\sum_{i=1}^n \ln X_i \sim \chi^2(4n)$ donc on choisit $k = q_\alpha/2$ où q_α est le α quantile de la loi $\chi^2(4n)$.

4. Lorsque $n = 30$, $\alpha = 5\%$, $k = 95.7/2 = 47.85$. Comme ici $\sum_{i=1}^n \ln x_i = 26.66$, on rejette (H_0) au niveau 5%.

Questions facultatives :

5. L'estimateur du maximum de vraisemblance de θ est donné par $\hat\theta = 2n/\sum_{i=1}^n \ln x_i$.

6. Statistique de test :

$$\lambda(x_1, \dots, x_n) = \frac{\sup_{\theta\le 1} L(x_1, \dots, x_n, \theta)}{\sup_{\theta>0} L(x_1, \dots, x_n, \theta)},$$

d'où $\lambda(x_1, \dots, x_n) = 1$ si $\hat\theta \le 1$, et sinon

$$\lambda(x_1, \dots, x_n) = \frac{L(x_1, \dots, x_n, 1)}{L(x_1, \dots, x_n, \hat\theta)} = \frac{(\prod_{i=1}^n x_i)^{\frac{2n}{\sum_{i=1}^n \ln x_i}-1}(\sum_{i=1}^n \ln x_i)^{2n}}{(2n)^{2n}}.$$

On a $\ln\lambda(x_1, \dots, x_n) = 2n - 2n\ln(2n) + 2n\ln(\sum_{i=1}^n \ln x_i) - \sum_{i=1}^n \ln x_i = 2n - 2n\ln(2n) + h(\sum_{i=1}^n \ln x_i)$, pour $2n/\sum_{i=1}^n \ln x_i > 1$, avec $h(t) = 2n\ln t - t$ pour $t < 2n$. On montre alors que h est croissante sur $]2n, +\infty[$, donc que $\ln\lambda(x_1, \dots, x_n)$ est croissante en $\varphi(x_1, \dots, x_n)$ pour $\hat\theta > 1$. Le test du rapport de vraisemblance maximal revient donc à rejeter l'hypothèse (H_0) lorsque $\varphi(x_1, \dots, x_n) \le k$. On retombe sur le test construit à la question 3.

Problème II

Tests paramétriques

Soit $Y = (Y_1, \dots, Y_{12})$ un 12-échantillon de la loi $\mathcal{N}(m_1, \sigma_1^2)$, modélisant les notes des filles, et $Z = (Z_1, \dots, Z_{10})$ un 10-échantillon de la loi $\mathcal{N}(m_2, \sigma_2^2)$, modélisant les notes des garçons, avec $(m_1, \sigma_1^2, m_2, \sigma_2^2) \in \Theta = (\mathbb{R}_+^*)^4$ inconnu. On suppose que Y et Z sont indépendants, et on pose $X = (Y, Z)$. Soit $x = (y, z)$ avec $y = (y_1, \dots, y_{12})$, $z = (z_1, \dots, z_{10})$, l'observation de ces deux échantillons indépendants. Soit $\mathcal{X} = \mathbb{R}^{22}$, et $\mathcal{A} = \mathcal{B}(\mathbb{R}^{22})$. Pour $\theta = (m_1, \sigma_1^2, m_2, \sigma_2^2) \in \Theta$, on note P_θ la loi de X : $P_\theta = \mathcal{N}(m_1, \sigma_1^2)^{\otimes 12} \bigotimes \mathcal{N}(m_2, \sigma_2^2)^{\otimes 10}$. Le modèle statistique considéré est défini par $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$.

1. On va d'abord tester l'égalité des variances i.e. (H_0) : "$\sigma_1^2 = \sigma_2^2$" contre (H_1) : "$\sigma_1^2 \neq \sigma_2^2$".
Statistique de test : $F(x) = \frac{S^2(y)}{S^2(z)}$, où $S^2(y) = \frac{1}{11}\sum_{i=1}^{12}(y_i - \bar{y})^2$ et $S^2(z) = \frac{1}{9}\sum_{i=1}^{10}(z_i - \bar{z})^2$.
Fonction de test : puisque $S^2(y)$ et $S^2(z)$ sont les estimateurs empiriques sans biais de σ_1^2 et σ_2^2, on choisit de rejeter (H_0) lorsque $F(x)$ prend de petites ou de grandes valeurs. La fonction de test correspondante s'écrit $\phi(x) = \mathbb{1}_{F(x) \leq s_1} + \mathbb{1}_{F(x) \geq s_2}$, avec $s_2 > s_1$.

Calcul de s_1 et s_2 : lorsque $\sigma_1^2 = \sigma_2^2 = \sigma^2$, i.e. lorsque $X \sim P_{(m_1, \sigma^2, m_2, \sigma^2)}$, on sait que $F(X) \sim \mathcal{F}(11, 9)$, donc pour un niveau de test 5%, on prend $s_1 = 0.28$, $s_2 = 3.91$. La fonction de test d'égalité des variances s'écrit finalement $\phi(x) = \mathbb{1}_{F(x) \leq 0.28} + \mathbb{1}_{F(x) \geq 3.91}$.
Conclusion : ici, on a $F(x) = 0.756$, donc on ne rejette pas l'hypothèse d'égalité des variances. On peut mettre en œuvre le test d'égalité des moyennes, en supposant que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
2. On suppose donc que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
Statistique de test : $T(x) = \frac{\bar{y} - \bar{z}}{S(y,z)\sqrt{\frac{1}{12} + \frac{1}{10}}}$, où $S^2(y, z) = \frac{\sum_{i=1}^{12}(y_i - \bar{y})^2 + \sum_{i=1}^{10}(z_i - \bar{z})^2}{20}$.
Fonction de test : puisque $\bar{y}$ et $\bar{z}$ sont les estimateurs empiriques de m_1 et m_2, on choisit de rejeter (H_0) lorsque $|T(x)|$ prend de grandes valeurs. La fonction de test correspondante s'écrit $\phi'(x) = \mathbb{1}_{|T(x)| \geq s}$.
Calcul de la constante s : lorsque $m_1 = m_2$, $T(X)$ suit la loi de Student à $(12 + 10 - 2) = 20$ degrés de liberté. Donc pour un niveau de test 5%, $s = 2.09$. La fonction de test d'égalité des moyennes s'écrit finalement $\phi'(x) = \mathbb{1}_{|T(x)| \geq 2.09}$.
Conclusion : ici, on a $T(x) = -0.153$, donc on ne rejette pas l'hypothèse d'égalité des moyennes. On n'a pas assez d'éléments pour conclure que la réussite varie selon le sexe de l'étudiant.
3. Le test est seulement UPPSB au niveau 5% (test bilatère avec paramètres de nuisance).

Test non paramétrique

1. Soit (X, Y) un couple de v.a.r. modélisant le sexe d'un étudiant et la classe (en terme de note) à laquelle il appartient. On dispose de l'observation $z = ((x_1, y_1), \dots, (x_n, y_n))$ $(n = 83)$ d'un n-échantillon $Z = ((X_1, Y_1), \dots, (X_n, Y_n))$ de ce couple.
On veut tester (H_0) X et Y sont indépendantes contre (H_1) X et Y ne sont pas indépendantes. On fait pour cela un test du chi deux d'indépendance.
Statistique de test : $T(z) = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(\frac{N_{i,*}N_{*,j}}{n} - N_{i,j})^2}{\frac{N_{i,*}N_{*,j}}{n}}$, où

— $N_{i,j}$ est le nombre d'étudiants de sexe i, et appartenant à la classe j,
— $N_{i,*}$ est le nombre d'étudiants de sexe i,

— $N_{*,j}$ est le nombre d'étudiants appartenant à la classe j.

Fonction de test : $\phi(z) = \mathbb{1}_{\{T(z)\geq s\}}$.

Calcul de la constante s : sous l'hypothèse (H_0), $T(Z)$ suit asymptotiquement la loi $\chi^2(3)$. Pour un niveau asymptotique 5%, on prend $s = 7.81$.

Ici, $T(z) = 0.786$, donc on ne rejette pas l'hypothèse d'indépendance au niveau asymptotique 5%.

3. La critique que l'on peut faire est que la taille de l'échantillon (des filles en particulier) est trop petite pour obtenir un test fiable. Les effectifs théoriques ne sont d'ailleurs pas tous supérieurs à 5 !

Problème III

Soit $Z = (X_1, \dots, X_n)$ un n-échantillon de la loi P de la v.a. X (absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$), modélisant le temps d'insertion d'un jeune diplomé de Master. Soit $z = (x_1, \dots, x_n)$ l'observation de cet échantillon. On veut tester (H_0) P est une loi $\mathcal{E}(1/6)$ contre (H_1) P n'est pas une loi $\mathcal{E}(1/6)$. On utilise pour cela un test de Kolmogorov-Smirnov d'adéquation.

1. Statistique de test : $D_n^0(Z) = \sqrt{n}\sup_{t\in\mathbb{R}_+} |F_n(t) - F^0(t)|$, où F_n est la fonction de répartition associée à l'échantillon Z et F^0 est la fonction de répartition de la loi $\mathcal{E}(1/6)$: $F^0(t) = 1 - e^{-t/6}$. Cette statistique s'exprime également sous la forme :

$$D_n^0(z) = \sqrt{n}\max_{1\leq i\leq n}\max\left\{\left|F^0(x_{(i)}) - \frac{i}{n}\right|, \left|F^0(x_{(i)}) - \frac{i-1}{n}\right|\right\}.$$

2. On sait que lorsque $P = \mathcal{E}(1/6)$, pour tout n, $D_n^0(Z)$ suit la même loi que $U_n = \sqrt{n}\sup_{t\in[0,1]} |F_{U,n}(t) - t|$, où $F_{U,n}$ est la fonction de répartition associée à un n-échantillon de la loi uniforme sur $[0, 1]$. Cette loi peut donc être tabulée pour toute valeur de n.

3. D'après le théorème de Kolmogorov, $D_n^0(Z) \to^{(\mathcal{L})}$ loi de fonction de répartiton H définie par $H(t) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2t^2}$. Cette loi est également tabulée.

4. On sait que lorsque P n'est pas la loi $\mathcal{E}(1/6)$, $D_n^0(Z) \to +\infty$ p.s. donc la fonction de test est de la forme $\phi(z) = \mathbb{1}_{D_n^0(z)\geq s}$.

Calcul de la constante s : on a vu que lorsque $P = \mathcal{E}(1/6)$, $D_n^0(Z)$ suit la même loi que $U_n = \sqrt{n}\sup_{t\in[0,1]} |F_{U,n}(t) - t|$ donc on choisit pour $n = 10$, $s = 0.409 * \sqrt{10} = 1.293$ pour un niveau 5%.

Conclusion : Pour calculer la valeur de $D_n^0(z)$, on utilise le tableau suivant.

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	4.1	4.2	5.5	7	7.5	10.2	10.3	15.3	16	30.2
$F^0(x_{(i)})$	0.495	0.503	0.6	0.689	0.713	0.817	0.82	0.922	0.931	0.993

On obtient $D_n^0(z) = 1.566$, donc on rejette l'hypothèse (H_0) au niveau 5%.

5. Pour tester l'exponentialité seule, on peut utiliser le test dérivé de Kolmogorov-Smirnov (c.f. Exercices), ou, si le nombre d'observations peut être augmenté, un test d'adéquation du χ^2 en regroupant les données dans des classes et en estimant le paramètre de la loi exponentielle par $1/\bar{z}$. La loi asymptotique de la statistique de test est alors une loi du χ^2 à $m - 2$ degrés de liberté si le nombre de classes est m.

Chapitre 8

Rappels utiles sur les lois usuelles dans $\mathbb{R}$ et dans $\mathbb{R}^n$

8.1 Lois usuelles dans $\mathbb{R}$

8.1.1 Lois discrètes

Loi	Paramètres	Fn de masse	Espérance	Variance	Modélisation	Observations
Dirac δ_a	a	$\pi(x) = \mathbb{1}_a(x), x \in \mathbb{R}$	a	0	X est une variable aléatoire prenant la valeur a quel que soit le résultat de l'expérience.	Fonction caractéristique (F.c.) $\varphi(t) = e^{ita}$.
Uniforme sur $\{1,\dots,n\}$		$\pi(x) = \frac{1}{n}$, $x \in \{1,\dots,n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	X est une variable aléatoire correspondant à un nombre entier compris entre 1 et n choisi au hasard, de façon équiprobable.	
Bernoulli $\mathcal{B}(p)$	$p \in [0,1]$	$\pi(x) = p^x(1-p)^{1-x}$, $x \in \{0,1\}$	p	$p(1-p)$	Dans une expérience à 2 issues, succès et échec, avec une probabilité de succès égale à p, X est la variable aléatoire qui vaut 1 si l'expérience conduit à un succès, 0 sinon.	Si $X_i \sim \mathcal{B}(p)$ pour $i = 1,\dots,n$, avec $\{X_i, i = 1,\dots,n\}$ indépendantes, alors $\sum_{i=1}^n X_i \sim \mathcal{B}(n,p)$. F.c. $\varphi(t) = pe^{it} + (1-p)$.
Binomiale $\mathcal{B}(n,p)$	$n \in \mathbb{N}^*$, $p \in [0,1]$	$\pi(x) = \binom{n}{x} p^x(1-p)^{n-x}$, $x \in \{0,1,\dots,n\}$	np	$np(1-p)$	On répète n fois l'expérience à 2 issues de la loi de Bernoulli de façon indépendante. X modélise le nombre total de succès obtenus.	Si $X_1 \sim \mathcal{B}(n_1,p)$, $X_2 \sim \mathcal{B}(n_2,p)$, X_1 et X_2 indépendantes, alors $X_1+X_2 \sim \mathcal{B}(n_1+n_2,p)$. Si $X \sim \mathcal{B}(n,p)$, $n-X \sim \mathcal{B}(n,1-p)$. F.c. $\varphi(t) = (pe^{it} + (1-p))^n$.
Poisson $\mathcal{P}(\lambda)$	$\lambda > 0$	$\pi(x) = \frac{\lambda^x}{x!}e^{-\lambda}$, $x \in \mathbb{N}$	λ	λ	Loi limite d'une binomiale de paramètres (n,p) lorsque $n \to +\infty$, $p \to 0$, $np \to \lambda$.	Si $X_1 \sim \mathcal{P}(\lambda_1)$, $X_2 \sim \mathcal{P}(\lambda_2)$, X_1 et X_2 indépendantes, alors $X_1+X_2 \sim \mathcal{P}(\lambda_1+\lambda_2)$. F.c. $\varphi(t) = e^{\lambda(e^{it}-1)}$
Géométrique $\mathcal{G}(p)$	$p \in [0,1]$	$\pi(x) = (1-p)^{x-1}p$, $x \in \{1,2,\dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	On répète plusieurs fois l'expérience à deux issues de la loi de Bernoulli, de façon indépendante. X est la variable aléatoire correspondant au nombre de fois où il a fallu répéter l'expérience pour obtenir un succès.	F.c. $\varphi(t) = pe^{it}/(1-(1-p)e^{it})$.

Loi	Paramètres	Fonction de masse	Espérance	Variance	Modélisation	Observations
Binomiale négative	$n \in \mathbb{N}^*$, $p \in [0,1]$	$\pi(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}$, $x \in \{n, n+1, \ldots\}$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$	On répète plusieurs fois l'expérience à deux issues de la loi de Bernoulli, de façon indépendante. X est la variable aléatoire correspondant au nombre de fois où il a fallu répéter l'expérience pour obtenir n succès.	Si $n = 1$, on retrouve la loi géométrique. Si X_1 suit une loi binomiale négative de paramètres (n_1, p), X_2 suit une loi binomiale négative de paramètres (n_2, p), X_1 et X_2 indépendantes, alors $X_1 + X_2$ suit une loi binomiale négative de paramètres $(n_1 + n_2, p)$.
Hypergéométrique	$N \in \mathbb{N}^*$, $N_1 = 1, \ldots, N$, $n = 1, \ldots, N$	$\pi(x) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$, $x \in \mathbb{N}$, $x \geq \max(0, n-N+N_1)$, $x \leq \min(n, N_1)$	$\frac{nN_1}{N}$	$\frac{nN_1(N-N_1)(N-n)}{N^2(N-1)}$	On tire n boules sans remise dans une urne contenant N_1 boules blanches, $N - N_1$ boules noires. X modélise le nombre total de boules blanches tirées.	On retrouve la loi binomiale en faisant tendre N_1 et N vers $+\infty$, $\frac{N_1}{N}$ vers p.

8.1.2 Lois absolument continues

Loi	Paramètres	Densité	Espérance	Variance	Modélisation	Observations
Uniforme $\mathcal{U}([a,b])$	$a,\ b \in \mathbb{R}$, $a < b$	$f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	X est la variable aléatoire modélisant un nombre choisi au hasard entre a et b.	Si $X \sim \mathcal{U}([0,1])$, si F^{-1} désigne l'inverse généralisée de la fonction de répartition d'une loi P, alors $F^{-1}(X) \sim P$. F.c. $\varphi(t) = \sin(at)/(at)$ pour $\mathcal{U}([-a,a])$.
Exponentielle $\mathcal{E}(\lambda)$	$\lambda > 0$	$f(x) = \lambda e^{-\lambda x}\mathbb{1}_{]0,+\infty[}(x)$	$1/\lambda$	$1/\lambda^2$		Si $X \sim \mathcal{E}(\lambda)$, $X \sim \gamma(1,\lambda)$. F.c. $\varphi(t) = \lambda/(\lambda - it)$
Gamma $\gamma(p,\lambda)$	$p > 0$, $\lambda > 0$	$f(x) = \frac{\lambda}{\Gamma(p)}e^{-\lambda x}(\lambda x)^{p-1}\mathbb{1}_{]0,+\infty[}(x)$	p/λ	p/λ^2		Si $X \sim \gamma(p_1,\lambda)$ et $Y \sim \gamma(p_2,\lambda)$ indépendantes, $X + Y \sim \gamma(p_1+p_2,\lambda)$. Pour $n \in \mathbb{N}^*$, $\Gamma(n) = (n-1)!$
Beta I $\beta_1(a,b)$	$a > 0$, $b > 0$	$f(x) = \frac{1}{\beta(a,b)}(1-x)^{b-1}x^{a-1}\mathbb{1}_{]0,1[}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$	Si $X \sim \gamma(a,1)$ et $Y \sim \gamma(b,1)$ indépendantes, $\frac{X}{X+Y} \sim \beta_1(a,b)$.	$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Les statistiques d'ordre d'une loi $\mathcal{U}([0,1])$ suivent des lois Beta I.
Beta II $\beta_2(a,b)$	$a > 0$, $b > 0$	$f(x) = \frac{1}{\beta(a,b)}\frac{x^{a-1}}{(1+x)^{a+b}}\mathbb{1}_{]0,+\infty[}(x)$	$\frac{a}{b-1}$ si $b > 1$	$\frac{a(a+b-1)}{(b-1)^2(b-2)}$ si $b > 2$	Si $X \sim \gamma(a,1)$ et $Y \sim \gamma(b,1)$ indépendantes, alors $\frac{X}{Y} \sim \beta_2(a,b)$.	Si $X \sim \beta_1(a,b)$, $\frac{X}{1-X} \sim \beta_2(a,b)$ et réciproquement.
Weibull $W(a,\lambda)$	$a > 1$, $\lambda > 0$	$f(x) = a\lambda x^{a-1}e^{-\lambda x^a}\mathbb{1}_{]0,+\infty[}(x)$	$\frac{\Gamma(1+\frac{1}{a})}{\lambda^{\frac{1}{a}}}$	$\frac{\Gamma(1+\frac{2}{a})-\Gamma^2(1+\frac{1}{a})}{\lambda^{\frac{2}{a}}}$	$X \sim W(a,\lambda)$ si $X^a \sim \mathcal{E}(\lambda)$	
Gaussienne (normale) $\mathcal{N}(m,\sigma^2)$	$m \in \mathbb{R}$, $\sigma^2 > 0$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}$	m	σ^2	Loi limite du TCL.	$X \sim \mathcal{N}(m,\sigma^2) \Rightarrow \frac{X-m}{\sigma} \sim \mathcal{N}(0,1)$. $X_1 \sim \mathcal{N}(m_1,\sigma_1^2)$, $X_2 \sim \mathcal{N}(m_2,\sigma_2^2)$, X_1 et X_2 indépendantes, alors $X_1 + X_2 \sim \mathcal{N}(m_1+m_2,\sigma_1^2+\sigma_2^2)$. F.c. $\varphi(t) = e^{imt}e^{-\sigma^2t^2/2}$.

Lois issues des lois gaussiennes

Loi	Paramètres	Densité	Espérance	Variance	Modélisation	Observations
Cauchy		$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$			Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{N}(0,1)$ indépendantes, si $U \sim \mathcal{U}(]-\pi/2, \pi/2[)$, X/Y et $\tan U$ suivent la loi de Cauchy	F.c. $\varphi(t) = e^{-\lvert t\rvert}$.
Log-normale	$m \in \mathbb{R}$, $\sigma^2 > 0$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - m)^2}{2\sigma^2}} \mathbb{1}_{]0,+\infty[}(x)$	$e^{m+\sigma^2/2}$	$e^{2m+\sigma^2}(e^{\sigma^2} - 1)$	X suit la loi log-normale de paramètres (m, σ^2) si $\ln X \sim \mathcal{N}(m, \sigma^2)$.	
Khi-Deux $\chi^2(n)$	$n \in \mathbb{N}^*$	$f(x) = \frac{1}{2\Gamma(n/2)} e^{-x/2} \left(\frac{x}{2}\right)^{n/2-1} \mathbb{1}_{]0,+\infty[}(x)$	n	$2n$	Si $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{N}(0,1)$, $X = X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$.	Si $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$ indépendantes, $X + Y \sim \chi^2(n_1 + n_2)$. Si $X \sim \chi^2(n)$, $X \sim \gamma(n/2, 1/2)$.
Student $\mathcal{T}(n)$	$n \in \mathbb{N}^*$	$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	0 si $n > 1$	$\frac{n}{n-2}$ si $n > 2$	Si $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(n)$, X et Y indépendantes, alors $\frac{X}{\sqrt{Y/n}} \sim \mathcal{T}(n)$.	Si $X \sim \mathcal{T}(n)$, $\frac{X^2}{n} \sim \beta_2\left(\frac{1}{2}, \frac{n}{2}\right)$.
Fisher-Snedecor $\mathcal{F}(m,n)$	$m \in \mathbb{N}^*$, $n \in \mathbb{N}^*$	$f(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{\beta\left(\frac{m}{2}, \frac{n}{2}\right)} x^{\frac{m}{2}-1} (n + mx)^{-\frac{n+m}{2}} \mathbb{1}_{]0,+\infty[}(x)$	$\frac{n}{n-2}$ si $n > 2$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ si $n > 4$	Si $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, X et Y indépendantes, $\frac{X/m}{Y/n} \sim \mathcal{F}(m,n)$.	Si $X \sim \mathcal{F}(m,n)$, alors $1/X \sim \mathcal{F}(n,m)$.

8.2 Lois usuelles dans $\mathbb{R}^n$

La loi multinomiale.

On répète n fois une expérience à k issues, de probabilités respectives $p_1, \dots, p_k$, de façon indépendante, et on considère pour tout $i = 1, \dots, k$, le nombre X_i de réalisations de l'issue i (parmi les n répétitions).

Définition 31. *On dit que le vecteur aléatoire $X = (X_1, \dots, X_k)$ suit une loi multinomiale de paramètres $(n, p_1, \dots, p_k)$, et on note $X \sim \mathcal{M}(n, p_1, \dots, p_k)$. On a*

$$P((X_1, \dots, X_k) = (x_1, \dots, x_k)) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}.$$

Propriétés.
Les variables marginales $X_1, \dots, X_k$ sont linéairement dépendantes par construction : $\sum_{i=1}^k X_i = n$, et on a bien sûr $\sum_{i=1}^k p_i = 1$.
Chaque variable marginale X_i suit une loi binomiale $\mathcal{B}(n, p_i)$.
On a donc $E[X] = (np_1, \dots, np_k)$, et la matrice de variances-covariances de X est égale à

$$\Sigma_X = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \dots & -np_1p_k \\ -np_1p_2 & np_2(1-p_2) & \dots & -np_2p_k \\ \vdots & \vdots & \vdots & \vdots \\ -np_1p_k & -np_2p_k & \dots & np_k(1-p_k) \end{pmatrix}.$$

Remarque : Cette matrice Σ_X n'est pas inversible.

La loi multinomiale a toute son importance en statistique. Elle est notamment à la base d'un test très utilisé en statistique : le test du khi-deux.

Vecteurs gaussiens - loi multinormale.

Retour sur la loi gaussienne dans $\mathbb{R}$: Densités, forme des densités, centrage et réduction d'une v.a.r. gaussienne, quantiles des lois gaussiennes, loi gaussienne dégénérée.

Définitions - Premières propriétés

Définition 32. *Un vecteur aléatoire de dimension k $X = (X_1, \dots, X_k)$ est dit* gaussien *si toute combinaison linéaire de ses variables marginales $a_1X_1 + \dots + a_kX_k$ est une v.a.r. gaussienne.*

Proposition 9. *La loi d'un vecteur aléatoire $X = (X_1, \dots, X_k)$ gaussien est entièrement déterminée par la donnée de son espérance $m_X = E[X]$ et de sa matrice de variances-covariances Σ_X. On note alors $X \sim \mathcal{N}_k(m_X, \Sigma_X)$.*
La fonction caractéristique de la loi $\mathcal{N}_k(m_X, \Sigma_X)$ est donnée par $\varphi_X(t) = e^{it'm_X} e^{-\frac{1}{2}t'\Sigma t}$.

Proposition 10. *Soit $X = (X_1, \dots, X_k)$ un vecteur gaussien. Par définition, les variables marginales $X_1, \dots, X_k$ sont des v.a.r. gaussiennes. Mais la réciproque est fausse.*

Preuve. On prend $X_1 \sim \mathcal{N}(0,1)$, ε une variable aléatoire de Rademacher, c'est-à-dire qui prend les valeurs 1 et -1 avec probabilité $1/2$, et on pose $X_2 = \varepsilon X_1$. On peut voir que X_1 et X_2 sont des v.a.r. gaussiennes centrées réduites, mais que $X_1 + X_2$ n'est pas une variable gaussienne (notamment, on a $P(X_1+X_2=0) \neq 0$!). Par conséquent (X_1, X_2) n'est pas un vecteur gaussien.

Proposition 11. *Si $X_1, \dots, X_k$ sont des v.a.r. gaussiennes* **indépendantes***, alors le vecteur aléatoire $X = (X_1, \dots, X_k)$ est un vecteur gaussien.*

Définition 33. *Un vecteur aléatoire $X = (X_1, \dots, X_k)$ est dit gaussien centré réduit ou standard si $X \sim \mathcal{N}_k(0, I_k)$.*

Proposition 12. *$X = (X_1, \dots, X_k) \sim \mathcal{N}_k(m_X, \Sigma_X)$ si et seulement si X peut s'écrire $X = AY + m_X$, avec $Y \sim \mathcal{N}_k(0, I_k)$, A étant de taille $k \times k$, satisfaisant $AA' = \Sigma_X$ et $rang(A) = rang(\Sigma_X)$.*
Si $X = (X_1, \dots, X_k) \sim \mathcal{N}_k(m_X, \Sigma_X)$, A une matrice de taille $l \times k$ et $B \in \mathbb{R}^l$, alors $AX + B \sim \mathcal{N}_l(Am_X + B, A\Sigma_X A')$. En particulier, le vecteur $\Sigma_X^{-1/2}(X - m_X)$ est gaussien standard.

Indépendance

Proposition 13. *Soit $X = (X_1, \dots, X_k) \sim \mathcal{N}_k(m_X, \Sigma_X)$. Les variables marginales $X_1, \dots, X_k$ sont indépendantes si et seulement si Σ_X est diagonale, autrement dit si et seulement si elles sont non corrélées.*
Soit $Z = (X_1, \dots, X_p, Y_1, \dots, Y_q)$ un vecteur gaussien. Les vecteurs $(X_1, \dots, X_p)$ et $(Y_1, \dots, Y_q)$ sont indépendants si et seulement s'ils sont non corrélés.

Preuve. Utilisation de la fonction caractéristique.

Attention : on rappelle que ces propriétés sont fausses dans le cadre général des vecteurs aléatoires ! Si on reprend l'exemple ci-dessus de $X_1 \sim \mathcal{N}(0,1)$, et $X_2 = \varepsilon X_1$, alors les variables X_1 et X_2 sont non corrélées, mais dépendantes.

Densité

Proposition 14. *Soit $X = (X_1, \dots, X_k) \sim \mathcal{N}_k(m_X, \Sigma_X)$. X admet une densité f si et seulement si $\det \Sigma_X \neq 0$, et on a*

$$f(x_1, \dots, x_k) = \frac{1}{(\sqrt{2\pi})^k \sqrt{det\ \Sigma_X}} e^{-\frac{1}{2}(x-m_X)'\Sigma_X^{-1}(x-m_X)},$$

avec $x = (x_1, \dots, x_k)'$.

Preuve. On a $Y = \Sigma_X^{-1/2}(X - m_X) \sim \mathcal{N}_k(0, I_k)$, donc ses variables marginales $Y_1, \dots, Y_k$ sont des v.a.r. gaussiennes centrées réduites indépendantes. Y a donc pour densité

$$g(y) = \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\sum_{i=1}^k y_i^2}.$$

Il suffit alors d'appliquer la formule du changement de variables. Le jacobien vaut $1/(\det \Sigma_X^{1/2}) = (\det \Sigma_X)^{-1/2}$, ce qui donne le résultat.

Exemple. Cas du vecteur gaussien de dimension 2 et lien avec le coefficient de corrélation linéaire.

Définition 34. *Un vecteur gaussien X dont la matrice de variances-covariances a un déterminant nul est dit* dégénéré.

Vecteurs gaussiens et loi du Khi-Deux

Rappels sur la loi du Khi-Deux.

Proposition 15. *Si $X = (X_1, \dots, X_k) \sim \mathcal{N}_k(m_X, \Sigma_X)$, avec Σ_X inversible, alors $(X-m_X)'\Sigma_X^{-1}(X-m_X)$ suit une loi du $\chi^2(k)$.*

Preuve. Il suffit ici aussi de voir que $\Sigma_X^{-1/2}(X - m_X) \sim \mathcal{N}_k(0, I_k)$.

Proposition 16. *Soit $X = (X_1, \dots, X_k) \sim \mathcal{N}_k(0, I_k)$.*
$X'AX$ suit une loi du χ^2 si et seulement si A est une matrice de projection orthogonale, i.e. $A^2 = A = A'$. Le nombre de degrés de liberté du χ^2 est alors égal au rang de A.
Soit A, B deux matrices de projection orthogonale. $X'AX$ et $X'BX$ sont indépendantes si et seulement si $AB = 0$.

Thèorème 16 (Théorème de Cochran)**.** *Soit $X = (X_1, \dots, X_k) \sim \mathcal{N}_k(0, I_k)$. Soit $A_1, \dots, A_p$ p matrices symétriques de taille $k \times k$ telles que $\sum_{l=1}^{p} X'A_lX = X'X$ (décomposition de la norme de X au carré). Alors les trois conditions suivantes sont équivalentes :*
- $\sum_{l=1}^{p} rang(A_l) = k$.
- Pour tout $l = 1, \dots, p$, $X'A_lX \sim \chi^2(rang\ A_l)$.
- Les $X'A_lX$ sont indépendantes.

Application aux projections sur des sous-espaces vectoriels de $\mathbb{R}^k$.

Bibliographie

[1] Anderson, T.W. (1962) *On the Distribution of the Two-Sample Cramer–von Mises Criterion*, The Annals of Mathematical Statistics, Vol. 33 (3), pp. 1148-1159.

[2] Anderson, T.W., Darling, D.A. (1954) *A Test of Goodness-of-Fit*, Journal of the American Statistical Association, Vol. 49. pp. 765-769.

[3] Casella, G., Berger, R. L. (2002) *Statistical inference*, Thomson Learning.

[4] Cramer, H. (1928) *On the composition of elementary errors*, Skand. Aktuarietidskr, Vol. 11, pp. 13-74 and pp. 171-180.

[5] Darling, D. A. (1955) *The Cramer-Smirnov test in the parametric case*, Annals of Mathematical Statistics, Vol. 26, pp. 1-20.

[6] Fourdrinier, D. (2002) *Statistiques inférentielles*, Dunod.

[7] Gauvrit, N. (2007) *Statistiques : méfiez-vous !*, Ellipses Marketing.

[8] Klatzmann, J. (1996) *Attention, statistiques ! : Comment en déjouer les pièges*, La Découverte.

[9] Kolmogorov, A. (1931) *Uber die analytischen Methoden in der Wahrscheinlichkeitsrechnung*, Math. Ann., Vol. 104, pp. 415-458.

[10] Kolmogorov, A. (1933a) *Sulla determinazione empirica di una legge di duistributione*, Giornale dell' Istituto Ialiano delgli Attuar, Vol. 4, pp. 1-11.

[11] Kolmogorov, A. (1933b) : *Uber die Grenzwertsatze der Wahrscheinlichkeitsrechnung*, Bulletin (Izvestija) Académie des Sciences URSS, pp. 363-372.

[12] Kolmogorov, A. (1941) *Confidence limits for an unknown distribution function*, Annals of Mathematical Statistics, Vol. 12, pp. 461-463.

[13] Lehmann, E. L. (1993) *The Fisher, Neyman-Pearson theories of testing hypotheses : one theory or two ?*, Journal of the American Statistical Association, Vol. 88, pp.1242–1249.

[14] Lehmann, E. L. (1998) *Nonparametrics - Statistical methods based on ranks*, Springer.

[15] Lehmann, E. L., Romano, J. P. (2005) *Testing Statistical Hypotheses*, Springer.

[16] Lilliefors, H. (1967) *On the Kolmogorov-Smirnov test for normality with mean and variance unknown*, Journal of the American Statistical Association, Vol. 62. pp. 399-402.

[17] Mann, H. B.,Whitney, D. R. (1947) *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other*, Annals of Mathematical Statistics, Vol. 18 (1), pp. 50–60.

[18] Monfort, A. (1997) *Cours de Statistique*, Economica.

[19] Neyman, J., Pearson, E. S. (1933) *On the problem of the most efficient tests of statistical hypotheses*, Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences, Vol. 231 (694–706), pp. 289–337.

[20] Pearson, K. (1900) *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, Philosophical Magazine Series 5, Vol. 50 (302), pp. 157–175.

[21] Pupion, G., Pupion, P.-C. (1998) *Tests non paramétriques*, Economica.

[22] Shapiro, S. S., Wilk, M. B. (1965) *An analysis of variance test for normality (complete samples)*, Biometrika, Vol. 52 (3/4), pp. 591-611.

[23] Smirnov, N. (1939a) *Sur les écarts de la courbe de distribution empirique (Ob uklonenijah empiriceskoi krivoi raspredelenija)*, Recueil Mathematique (Matematiceskii Sbornik), N. S. 6 (48), pp. 13-26.

[24] Smirnov, N. (1939b) *On the estimation of the discrepancy between empirical curves of distribution for two independent samples*, Bulletin Mathémtatique de l'Université de Moscou, Vol. 2 (2), pp. 3-14.

[25] Tassi, P. (2004) *Méthodes statistiques*, Economica.

[26] Von Mises, R. (1931) *Vorlesungen aus dem Gebiete der Angewandten Mathematik, 1, Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theoretischen Physik*, Leipzig and Wien.

[27] Wilcoxon, F. (1945) *Individual comparisons by ranking methods*, Biometrics Bulletin, Vol. 1 (6), pp. 80–83.