Semaine 5 - Tests d'hypothèses, d'ajustement et d'independance

Exercice 1 – Test entre hypothèses simples

Parmi les personnes atteintes d'une certaine maladie, que l'on ne sait pas traiter, 36% guérissent spontanément, les 64% restant devenant des malades chroniques.
Un laboratoire pharmaceutique propose un remède très coûteux avec lequel, affirme-t-il, le pourcentage de guérison passe à 50%.
Un service hospitalier doute de l'efficacité de ce remède ; pour le tester, il l'administre à un échantillon de 100 patients atteints de la maladie ; les patients sont numérotés de $k = 1$ à $k = 100$.

Q 1.1 – Mise en place du test d'hypothèse :

Q 1.1.1 Quelles sont les hypothèses simples en présence ? (on appellera θ le paramètre). Laquelle doit-on prendre comme hypothèse H_0 ?

Q 1.1.2 Au patient k est associée la variable X_k qui prend la valeur 1 si ce patient guérit et la valeur 0 sinon. Quel est le type de loi suivie par X_k dans les deux hypothèses H_0 et H_1 ?
Vérifier que les probabilités élémentaires des deux lois peuvent se mettre sous la forme :

$$P_\theta(X_k = x_k) = \theta^{x_k}(1-\theta)^{1-x_k},\ x_k \in \{0,1\}$$

avec $\theta = \theta_0$ pour l'une, $\theta = \theta_1$ pour l'autre.

Q 1.1.3 En déduire l'expression de la vraisemblance [= la probabilité d'obtenir l'échantillon $(x_1, \ldots, x_n)$ conditionnellement à θ],

$$L(\mathbf{x}, \theta) = L(x_1, \ldots, x_k, \ldots, x_n, \theta) = \prod_{k=1}^{n} P_\theta(X_k = x_k).$$

Montrer qu'elle s'exprime comme une fonction de la moyenne empirique $\bar{x}$ et de θ.

Q 1.1.4 En déduire que, pour tout test du rapport de vraisemblance $L(\mathbf{x}, \theta_0)/L(\mathbf{x}, \theta_1)$, il existera un nombre positif λ tel que :

$$\bar{x} < \lambda \ \Rightarrow \text{accepter } H_0$$
$$\bar{x} > \lambda \ \Rightarrow \text{rejeter } H_0.$$

Q 1.2 – Réalisation effective du test :

Q 1.2.1 Que représente la variable $Y = n\bar{X}$? Quelle est la loi de Y dans l'hypothèse H_0 ?

Q 1.2.2 La table ci-dessous donne les probabilités exactes d'observer k guérisons au moins parmi les 100 malades sous l'hypothèse H_0 (de $k = 42$ à $k = 50$) ; les valeurs obtenues par l'approximation normale sont données au-dessous.

k	42	43	44	45	46	47	48	49	50
$P_{\theta_0}(Y \geq k)$	0.126	0.089	0.060	0.040	0.025	0.015	0.009	0.0052	0.0029
val. appr.	0.125	0.089	0.059	0.038	0.023	0.014	0.008	0.0047	0.0025

Au niveau de signification $\alpha = 0.05$, quelles sont les valeurs de k pour lesquelles on doit : accepter l'hypothèse H_0 ? rejeter H_0 ? rejeter H_0 avec une certaine probabilité ? (que l'on précisera).

Q 1.2.3 Sous l'hypothèse H_1,

$$P_{\theta_1}(Y \leq 43) = 0.0967 \text{ et } P_{\theta_1}(Y = 44) = 0.039$$
$$P_{\theta_1}(Y \leq 46) = 0.242 \text{ et } P_{\theta_1}(Y = 47) = 0.066$$

En déduire la puissance du test de niveau de signification $\alpha = 0.05$.

Q 1.2.4 Mêmes questions que précédemment mais cette fois au niveau de signification $\alpha = 0.01$.

Q 1.2.5 L'approximation normale est-elle bonne ici ? Quel est le théorème de convergence qui laisse prévoir ce fait ?

Q 1.3 On suppose que le chef du service hospitalier est capable :
- d'attribuer des probabilités *a priori* aux deux hypothèses

$$\pi_0 = P(H_0)\,;\, \pi_1 = P(H_1) = 1 - \pi_0\,;$$

- et d'estimer le coût d'une erreur de $1^{\text{ère}}$ espèce, C_0 et de 2^{eme} espèce C_1, ce qui permet de se placer dans le cadre de la statistique bayésienne.

Soit un test T_W, caractérisable par sa région critique W, c'est-à-dire tel que :

$$H_0 \text{ rejeté} \Leftrightarrow x \in W$$
$$H_0 \text{ accepté} \Leftrightarrow x \notin W$$

Q 1.3.1 Montrer que l'espérance mathématique du coût de ce test est :

$$\pi_0 C_0 \sum_{x \in W} L(\mathbf{x},\, \theta_0) + \pi_1 C_1 \sum_{x \notin W} L(\mathbf{x},\, \theta_1).$$

Q 1.3.2 En déduire que, s'il existe $x \in W$ tel que :

$$L(\mathbf{x},\, \theta_0) > \frac{\pi_1 C_1}{\pi_0 C_0} L(\mathbf{x},\, \theta_1),$$

alors il existe un autre test d'espérance de coût strictement inférieure à celle du test T_W.

Q 1.3.3 Montrer de même que s'il existe $x \notin W$ tel que

$$L(\mathbf{x},\, \theta_0) < \frac{\pi_1 C_1}{\pi_0 C_0} L(\mathbf{x},\, \theta_1),$$

alors il existe un autre test d'espérance de coût strictement inférieure à celle du test T_W.

Q 1.3.4 En déduire qu'un test optimal bayésien est nécessairement un test du rapport de vraisemblance. Comment λ varie-t-il avec π_0 et π_1 d'une part et C_0 et C_1 d'autre part ?
Donner un test optimal bayésien lorsque $\pi_0 = 0.95$, $C_0 = 1$ et $C_1 = 10$ (unités monétaires).

Exercice 2 – Investissement à la bourse

Vous voulez investir à la bourse. Afin d'optimiser vos profits, vous relevez pendant 16 jours le cours du CAC40. Au début de ces deux semaines, celui-ci vaut 5715 points. Dans l'échantillon de 16 jours, le CAC40 vaut en moyenne 5726,025 points, avec un écart-type de 6 points. Vous ne voulez investir que si le CAC40 est à la hausse.

Q 2.1 Sachant que la variable $X =$ « valeur du CAC40 » suit une loi normale de variance 36, effectuez un test d'hypothèse de niveau de confiance 99% pour savoir si le CAC40 a augmenté. Vous préciserez bien les hypothèses H_0 et H_1.

Q 2.2 D'après le test précédent, peut-on conclure que le CAC40 a augmenté ?

Q 2.3 Calculez la puissance du test pour $\mu = 5726,025$. Pour vous aider, vous pourrez supposer que si une variable $Y \sim \mathcal{N}(0,1)$, alors :

$$P(Y > -1) = 0,8413 \qquad P(Y > -2) = 0,9772 \qquad P(Y > -3) = 0,9986 \qquad P(Y > -4) \approx 1.$$

Exercice 3 – Test d'ajustement du χ^2

Dans un supermarché, on maintient 8 caisses de plus de 10 articles en opération durant les nocturnes du jeudi. Normalement, la clientèle devrait se répartir uniformément entre les caisses. Afin de vérifier cela, on a recensé le nombre de clients passés à chacune des caisses un jeudi soir. Les résultats observés ont été les suivants :

Numéro de la caisse	Nombre de clients
1	72
2	70
3	71
4	52
5	45
6	59
7	67
8	48
Total	484

Hypothèse H_0 à tester : « la clientèle se répartit uniformément entre les 8 caisses ».

Q 3.1 Sous l'hypothèse H_0, quels sont les effectifs théoriques ν_i dans chaque classe ?

Q 3.2 Calculer la statistique d'ajustement :

$$A = \sum_{i=1}^{I} \frac{(n_i - \nu_i)^2}{\nu_i}.$$

Q 3.3 Quel est le nombre de degrés de liberté ? Au niveau de signification $\alpha = 0.05$, doit-on accepter H_0 ?

Exercice 4 – Boules de couleur

Soit une urne contenant des boules de 5 couleurs différentes : (R)ouges, (B)leues, (V)ertes, (J)aunes, (N)oires. On suspecte que la distribution de probabilité sur les couleurs des boules de l'urne est la suivante :

$$P(R) = 0,2 \quad P(B) = 0,4 \quad P(V) = 0,1 \quad P(J) = 0,2 \quad P(N) = 0,1.$$

Par ailleurs, on a tiré un échantillon i.i.d. de 20 boules et on a noté le nombre de boules de chaque couleur :

Couleur	R	B	V	J	N
Nb boules	2	9	4	5	0

Faites un test d'ajustement avec un niveau de confiance $1 - \alpha = 90\%$ pour déterminer si, oui ou non, la distribution de probabilité sur les couleurs des boules est celle indiquée ci-dessus.

Exercice 5 – Test d'indépendance du χ^2

Un échantillon de 200 contribuables est prélevé afin de vérifier si le revenu brut annuel d'un individu est un caractère dépendant du niveau de scolarité de l'individu. Les observations recueuillies sont données dans le tableau suivant :

scolarité (années) $\rightarrow$ revenu (kF) $\downarrow$	$[0;7[$	$[7;12[$	$[12;14[$	$[14;\rightarrow[$	*total*
$[0;75[$	17	14	9	5	45
$[75;120[$	12	37	11	5	65
$[120;200[$	7	20	20	8	55
$[200;\rightarrow[$	4	9	10	12	35
total	40	80	50	30	200

Q 5.1 On admet que les fréquences relatives déduites des marges du tableau donnent les vraies lois de probabilité, $p_{r.}$ et $p_{.s}$ des variables $R(evenu)$ et $S(colarité)$. Donner le tableau des fréquences théoriques, $200 \times p_{rs}$, correspondantes en cas d'indépendance des deux variables.

Q 5.2 Calculer le χ^2. Expliquez pourquoi il y a 9 degrés de liberté. Doit-on rejeter l'hypothèse d'indépendance au risque $\alpha = 5\%$?

Exercice 6 – Notation en MAPSI

On sait, par expérience, que les notes de partiel de MAPSI suivent une loi normale $\mathcal{N}(\mu; 6^2)$. On considère l'échantillon de notes i.i.d. suivant :

10	8	13	20	12	14	9	7	15

.

Q 6.1 Par expérience, les années précédentes, la moyenne au partiel de MAPSI était égale à 14. Dressez un test d'hypothèse de niveau de confiance $1 - \alpha = 95\%$ pour confronter les hypothèses $H_0 =$ « la moyenne est égale à 14 » et $H_1 =$ « la moyenne a baissé, i.e., elle est inférieure à 14 ».

Q 6.2 Calculez la puissance du test pour une moyenne de 12 (H_1 : la moyenne est égale à 12).

Exercice 7 – Il faut assurer

La loi oblige tout automobiliste à contracter une assurance. La prime exigée annuellement d'un assuré dépend de plusieurs facteurs : la zone habitée, le type de véhicule, l'utilisation à des fins commerciales ou non, la distance estimée que parcourra l'assuré... Il est presque impossible d'estimer la distance parcourue par un automobiliste pour une année donnée. Voilà pourquoi tous les assurés d'un véhicule non utilisé à des fins commerciales se voient imposer le même montant sur ce point. Celui-ci est fonction de la distance moyenne parcourue annuellement par les automobilistes de cette catégorie. Des études ont montré par le passé que celle-ci était de 18000 km avec un écart-type de 5000 km. Le montant que l'on prévoit d'exiger est de 2 centimes du km, autrement dit $18000 \times 0,02 = 360$€. Le montant de cette prime a continuellement augmenté ces dernières années, de telle sorte que l'opinion publique commence à être très mécontente et à exercer de fortes pressions sur les compagnies d'assurances pour qu'elles baissent leurs tarifs.
C'est ainsi que la MAIF est priée de réévaluer tous les facteurs considérés dans le calcul de la prime. Le plus vulnérable de ces facteurs est précisément la distance parcourue annuellement. Un statisticien est donc chargé de réexaminer le bien-fondé de l'estimation à 18000 km de la moyenne contestée. La démarche qu'il compte suivre est de prélever rapidement un échantillon de 400 individus afin de tester si la moyenne a effectivement diminué. Si tel est le cas, une étude plus exhaustive, menée sur un grand nombre d'assurés, sera entreprise afin d'estimer très précisément la valeur de la moyenne. Sinon, ce facteur ne sera pas révisé.
La variable qu'étudie le statisticien est X : la distance parcourue en 2013 (dernière année complète sur laquelle on peut fonder l'étude), par un véhicule utilisé à des fins non commerciales. Il décide pour l'instant de ne pas remettre en cause l'estimation de l'écart-type σ de X (5000 km). En revanche, il veut réestimer la moyenne μ et vous demande donc de l'aider :

Q 7.1 Dans un test d'hypothèses, quelles hypothèses H_0, H_1 formuleriez-vous pour tester s'il faut revoir les tarifs de l'assurance à la baisse ? Quelle serait la forme de la région critique ?

Q 7.2 Le statisticien s'interroge sur les conséquences qu'auraient le fait de rejeter H_0 alors que celle-ci est vraie. Cela entrainerait la réalisation de l'étude exhaustive pour rien, donc une dépense inutile, et porterait atteinte à la réputation du statisticien. Il décide donc de ne pas prendre de risque et de fixer la probabilité de commettre une telle erreur à $\alpha = 0,01$. Exprimez α en fonction de la région critique.

Q 7.3 Quelle est la loi suivie par $\overline{X}$ sous H_0 ?

Q 7.4 À partir de quelle valeur le test nous indique-t-il de rejeter H_0 ?

Q 7.5 Notre statisticien veut maintenant examiner la puissance de son test afin de voir si sa règle de décision est « solide ». Il réfléchit alors sur les conséquences de rejeter H_1 alors que H_1 est vraie. Si une telle erreur se produisait, les automobilistes n'obtiendraient pas une réduction du prix de la prime alors qu'il y auraient

droit. Si la diminution à laquelle ils avaient droit se chiffrait à 20 euros ou moins, on ne pourrait pas parler de conséquences sérieuses. à quel nombre moyen k de kilomètres parcourus correspond une baisse de 20 euros de l'assurance ?

Q 7.6 Calculez la puissance du test pour k. Cette puissance nous indique la probabilité que la règle de décision soit « fiable » pour une moyenne de k kilomètres, c'est-à-dire lorsque les assurés devraient commencer à percevoir une différence au niveau du prix de leur assurance. Est-ce que le statisticien peut procéder au recueil des données auprès des 400 personnes ou bien son test d'hypothèses n'est-il pas sûr ?

Exercice 8 – Sécurité sociale

On souhaite comparer l'efficacité de deux médicaments censés combattre la même maladie. Le premier médicament est générique et son prix est réduit, le deuxième est un médicament de marque de prix beaucoup plus élevé. La Sécurité Sociale a effectué une enquête sur les guérisons obtenues grâce à chacun de ces médicaments. Le nombre de guérisons et de non guérisons (sur les 250 personnes testées) sont consignés dans le tableau ci-dessous :

	générique	marque
guérisons	44	156
non guérisons	6	44

À un niveau de risque de 5%, peut-on estimer que le taux de guérison dépend du médicament (générique ou marque) ? Justifiez votre réponse mathématiquement.