TP 10 - Test du χ^2

1 Niveau de test et p-valeur

Le niveau d'un test α est défini comme le seuil de probabilité fixé *a priori* tel que le test rejette à tort $\mathcal{H}_0$ avec une probabilité α :

$$\mathbb{P}_{\mathcal{H}_0}\left[\text{Rejet de } \mathcal{H}_0\right] = \alpha$$

La région de rejet de $\mathcal{H}_0$ est alors définie à partir de cette égalité et d'une statistique de test T dont on connaît la loi sous l'hypothèse $\mathcal{H}_0$. Au lieu de donner directement l'acceptation ou le rejet de l'hypothèse nulle, les statisticiens préfèrent souvent renvoyer la statistique du test T en donnant précisément le seuil limite auquel $\mathcal{H}_0$ aurait été rejetée compte tenu de l'observation. Le seuil est d'autant plus bas qu'on est « loin »de $\mathcal{H}_0$ à la vue de la réalisation de T.

Question 1 Donner la région de rejet pour $\mathcal{H}_0 = \{T \sim \mathcal{N}(0;1)\}$ au seuil $\alpha = 0.05$. Supposons alors mesurée la valeur $T = 2.72$, donner le seuil limite de rejet.

Definition 1.1 (P-valeur) *Soit $\mathcal{H}_0$ l'hypothèse nulle, T la statistique de test et F_0 sa fonction de répartition sous $\mathcal{H}_0$ supposée continue. On appelle p-valeur d'une valeur t prise par T la quantité*

$$p(t) = \min\left\{F_0(t); 1 - F_0(t)\right\}$$

lorsque le test est bilatère (région d'acceptation du type $t \in [a;b]$). Pour un test unilatère à droite (rejet des valeurs trop grandes, la p-valeur est donnée par

$$p(t) = 1 - F_0(t)$$

Dans le cas inverse (rejet des valeurs trop petites), la p-valeur vaut alors

$$p(t) = F_0(t)$$

La connaissance de la p-valeur rend inutile le calcul de la région de rejet : si $p(t)$ est la p-valeur d'une observation t sous $\mathcal{H}_0$, on obtient un test de seuil α par la règle :

$$\textit{Rejet de } \mathcal{H}_0 \iff p(T) < \alpha$$

Ainsi, on rejette $\mathcal{H}_0$ avec d'autant moins de chance de se tromper que $p(t)$ est petite.

Remarque 1.2 *Il existe d'autres formules pour les p-valeurs selon que le test est en fait bilatère, unilatéral à droite ou à gauche, mais l'esprit de ces p-valeurs est le même.*

Question 2 On observe sur un 1000-échantillon la statistique de Kolmogorov Smirnov qui vaut $D_{KS}(F_0; \hat{F}) = 0.047$. Donner la valeur de la statistique de KS et la p-valeur correspondante. Accepte-t-on l'hypothèse nulle ?

2 Test du χ^2 d'ajustement

Soit $X_1, \dots X_n$ un n-échantillon de loi μ inconnue. On souhaite tester $\mathcal{H}_0 = \acute{} \ \mu = \nu \ \grave{}$ contre $\mathcal{H}_1 = \acute{} \ \mu \neq \nu \ \grave{}$ où ν est connue. On considère une partition en k classes $(I_1, \dots I_k)$ du support de ν et $\nu_i = \int_{I_i} \nu$. On note l'effectif n_i associé à la partition I_i pour la loi ν : $n_i = \nu_i n$, la la statistique de test du χ^2 est alors donnée par

$$D_n = \sum_{i=1}^{k} \frac{(d_i - n_i)^2}{d_i}$$

Question 3 Donner sous $\mathcal{H}_0$ le comportement asymptotique de D_n.

Question 4 L'exemple classique du test est l'expérience de Mendel. Chez les pois, le caractère couleur est codé par un gène présentant deux formes allèles C et c, correspondant aux couleurs jaune et vert. Le jaune est dominant, le vert récessif. La forme, rond ou ridé, est portée par un autre gène à deux allèles R (dominant) et r (récessif). Si on croise deux individus dont le génotype est CcRr, on peut obtenir 16 génotypes équiprobables. Les descendants seront jaunes et ronds dans 9 cas sur 16, jaunes et ridés dans 3 cas sur 16, verts et ronds dans 3 cas sur 16, verts et ridés dans 1 cas sur 16. Dans ses expériences, Mendel a obtenu les résultats suivants.

	Jaune	Jaune	Vert	Vert
	Rond	Ridé	Rond	Ridé
Effectif	315	101	108	32
$\hat{P}(c_h)$				
$P(c_h)$				

Donner la valeur de la statistique du test du χ^2. Donner la région de rejet au seuil $\alpha = 0.05$. Calculer la p-valeur de notre statistique. Conclusion ?

Question 5 L'exemple suivant concerne 10000 familles de 4 enfants pour lesquelles on connaît le nombre de garçons, entre 0 et 4. Le modèle le plus simple qu'on puisse proposer est que les naissances sont indépendantes, les deux sexes étant équiprobables.

– Identifier l'hypothèse nulle.
– Les fréquences observées et théoriques sont les suivantes :

Garçons	0	1	2	3	4
Effectifs	572	2329	3758	2632	709
$P_0(c_h)$					

Donner la valeur prise par la statistique du test du χ^2. Quelle est la région de rejet au seuil $\alpha = 0.05$? Donner la p-valeur de la statistique de test.

Question 6 Le test du chi-deux est souvent utilisé pour tester l'ajustement à une famille particulière dépendant d'un paramètre. Dans ce cas, on est amené à estimer le paramètre à partir des données. Le résultat de la question 3 n'est alors plus tout à fait valable. Si on a estimé h paramètres par la méthode du maximum de vraisemblance, à partir des fréquences des différentes classes, on doit remplacer la loi $\chi^2(r-1)$ par la loi $\chi^2(r-1-h)$.

Reprendre alors l'exemple de la question 5 en décrivant précisément combien de paramètres h sont à estimer en réalité. Renouveller les calculs précédents sur les statistiques de test et p-valeur. Conclusion ?

Question 7 On va tracer dans cette question la courbe des p-valeurs en fonction de la distance d'une loi donnée à une loi variable. Plus précisément :

– Générer aléatoirement une distribution de probabilités sur 4 modalités.
– Simuler pour chacune des lois des échantillons de taille 50.

– Calculer pour chacun de ces 50 échantillons la statistique du χ^2 d'ajustement à la loi uniforme ainsi que la p-valeur.
– Tracer alors le graphe des p-valeur en fonction de la distance $d(p;\mathcal{U})$.

Question 8 Tester la normalité du générateur randn de Matlab. On regroupera les réalisations des variables aléatoires simulées en 20 classes régulièrement espacées entre -2 et 2.

Question 9 Sur le 10ème jeu de données de la Stixbox obtenu *via* getdata(10), tester la normalité de la colonne correspondant au coût de la vie.

3 Test du χ^2 d'indépendance

L'observation est constituée d'un n-échantillon d'un couple (X, Y) à valeurs dans $E \times F$. On veut test $\mathcal{H}_0 =$ "$\mu_{(X,Y)} = \mu_X \otimes \mu_Y$" contre $\mathcal{H}_1 =$ "$\mu_{(X,Y)} \neq \mu_X \otimes \mu_Y$". Pour cela, on choisit des partitions de E et F et on désigne par N_{ij} le nombre de points (X_i, Y_j) dans $E_i \times F_j$.

Question 10 Si l'on désigne par p_i et q_j les marginales associées à ces 2 partitions de E et F, adapter le test du χ^2 à notre situation.

Question 11 Quel est le nombre de degré de libertés de la statistique précédemment donnée ? Quel est alors le comportement asymptotique sous $\mathcal{H}_0$?

Question 12 Tester l'indépendance pour 2 échantillons de loi uniforme sur $[0; 1]$ obtenus par le générateur rand de Matlab.

Question 13 Voici un exemple de deux caractères binaires, concernant des malades, pour lesquels on a observé s'il ont ou non une tendance suicidaire (caractère X). Leurs maladies ont été classées en "psychoses" et "névroses" (caractère Y). On souhaite savoir s'il y a une dépendance entre les tendances suicidaires et le classement des malades. La table de contingence observée est :

	Tendance	Sans tendance	Total
Psychoses	20	180	200
Névroses	60	140	200
Total	80	320	400

Y a-t-il indépendance de ces deux variables ?

Question 14 Sur le onzième jeu de données de la Stixbox, tester l'indépendance entre le pourcentage de gens faisant des études (colonne 6) et le produit national brut (colonne 7).[1]

[1] toto