TP N°3 : Test Statistiques

1 Test du χ^2 d'adéquation

1.1 Exemple du loto

Dans cet exercice, on cherche à vérifier si les numéros tirés lors de l'ancienne version du loto étaient bien tous tirés à probabilité égale. Dans cette version, à chaque tirage était tiré 7 boules parmi 49 numérotées de 1 à 49. On trouvera ici le vecteur V tel que $V(i)$ correspond au nombre de fois que la boule i a été tirée sur les 4858 tirages qui ont eu lieu entre 1976 et 2008. (source: fdj.fr)

(i) Télécharger les données à l'aide de la fonction *np.loadtxt*. Quelle est l'hypothèse $\mathcal{H}_0$ que l'on souhaite tester?

(ii) Calculer la statistique du χ^2 associée.

Même si les tirages ne sont pas fait de façon indépendante (on dit qu'ils sont 7-dépendant) les résultats du cours s'appliquent quand même dans ce cas-là.

(iii) Dans notre cas, la statistique du χ^2 à pour loi asymptotique sous $\mathcal{H}_0$ un χ^2 à combien de degrés de liberté? Donner le quantile $q_{0.95}$ de niveau 0.95 correspondant à l'aide de la fonction *chi2.ppf*.

(iv) Peut-on conclure que les tirages du loto sont bien équilibrés?

1.2 La loi de Benford

Une variable aléatoire réelle X suit la loi de Benford si elle vérifie:

$$\forall i \in \{1, \cdots, 9\}, \quad \mathbb{P}(\text{le premier chiffre significatif de } X \text{ est } i) = \log_{10}\left(1 + \frac{1}{i}\right)$$

Il se trouve que de nombreux jeux de données ont tendance à vérifier cette loi. Ce résultat a déjà été historiquement utilisé pour vérifier si un jeu de données était contrefait ou non. Notre objectif dans cette exercice est de vérifier si le nombre d'habitant dans les différentes villes françaises vérifient la loi de Benford. On trouvera ici un vecteur contenant le nombre d'habitant d'un échantillon de 10000 villes françaises. (source: sql.sh)

(i) Créer un vecteur U tel que $U(i)$ correspond à la proportion de villes françaises (parmi les 10000) dont le nombre d'habitant commence par le chiffre i. On pourra s'aider des fonctions *math.floor* et *math.log10*.

(ii) Créer le vecteur V des probabilités associées à la loi de Benford.

(iii) Appliquer le test du χ^2 avec niveau de confiance 95% et conclure.

1.3 La loi de Poisson

On cherche à modéliser la loi du nombre d'accidents de la route lors d'une période de temps donnée. Pour cela, on trouvera ici un vecteur comprenant le nombre d'accidents de la route au Royaume-Uni entre 13:10 et 13:20 pour chaque Lundi de l'année 2016. (source: data.gov.uk/dataset/road-accidents-safety-data)

(i) On suppose que les données suivent une loi de Poisson. Choisir un paramètre adapté pour cette loi.

Afin de vérifier notre hypothèse par un test du χ^2, on a besoin de découper l'espace des possibilités en plusieurs classes de sorte que l'on possède au moins 5 observations dans chaque classe.

(ii) Choisir une décomposition en classe adaptée et créer le vecteur U des fréquences d'observation de chaque classe et V le vecteur des probabilités sous $\mathcal{H}_0$ associées à chaque classe.

(iii) Appliquer le test du χ^2 avec niveau de confiance 95% et conclure.

2 Test du χ^2 d'indépendance

2.1 Cancer de l'œsophage

Une étude cherchant un lien entre la consommation d'alcool, de tabac et le cancer de l'œsophage à recueillit des données sur 975 patients dont les résultats sont donnés ci-dessous. (source: Breslow, N. E. and Day, N. E. (1980) Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies.)

Consommation d'alcool quotidienne (en g)	Nombre de patients avec cancer	Nombre de patients sans cancer	Total
0-39	29	386	415
40-79	75	280	355
80-119	51	87	138
120+	45	22	67
Total	200	775	975

Consommation de tabac quotidienne (en g)	Nombre de patients avec cancer	Nombre de patients sans cancer	Total
0-9	78	447	525
10-19	58	178	236
20-29	33	99	132
30+	31	51	82
Total	200	775	975

(i) On s'intéresse d'abord au premier tableau de données. Quelle est l'hypothèse $\mathcal{H}_0$ que l'on souhaite tester?

(ii) Calculer la statistique du χ^2 associée.

(iii) Cette statistique à pour loi asymptotique un χ^2 à combien de degrés de liberté? Donner le quantile $q_{0.95}$ de niveau 0.95 correspondant.

(iv) Quel conclusion peut-on tirer de ces résultats?

(v) Refaire les mêmes questions avec le second tableau de données.

(vi) Comparer les deux statistiques du χ^2 obtenues. A votre avis, quel dépendance est la plus forte?

2.2 Couleur des cheveux VS couleur des yeux

On cherche à établir s'il existe un lien entre le sexe, la couleur des cheveux et la couleur des yeux d'une personne. Pour cela, on se base sur les résultats d'un sondage établi sur 592 étudiants à l'Université du Delaware. Les résultats obtenus sont les suivants: (source: http://euclid.psych.yorku.ca/ftp/sas/vcd/catdata/haireye.sas)

Femme	Yeux verts	Yeux marrons	Yeux bleus	Total
Cheveux noirs	2	41	9	52
Cheveux marrons	14	95	34	143
Cheveux roux	7	23	7	37
Cheveux blonds	8	9	64	81
Total	31	168	114	313

Homme	Yeux verts	Yeux marrons	Yeux bleus	Total
Cheveux noirs	3	42	11	56
Cheveux marrons	15	78	50	143
Cheveux roux	7	17	10	34
Cheveux blonds	8	8	30	46
Total	33	145	101	279

(i) Écrire une fonction $Chi(M, \alpha)$ effectuant un test du χ^2 avec niveau de confiance $1 - \alpha$ sur les données de la matrice M. Plus précisément, la fonction retournera *True* ou *False* selon que M passe le test ou non.

(ii) Appliquer cette fonction sur les tableaux précédents avec $\alpha = 0.05$ et vérifier s'il y a indépendance entre le sexe et la couleur des yeux, entre le sexe et la couleur des cheveux ou entre la couleur des yeux et des cheveux.

3 Test de Kolmogorov-Smirnov

Vous pouvez consulter la doc en ligne https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html