Introduction au test du χ^2

Les parties marquées d'un ⛰, ne doivent être abordées que si vous avez de l'avance.

Objectifs, motivations et premier pas

Le but de ce TP est de présenter le test statistique du χ^2.

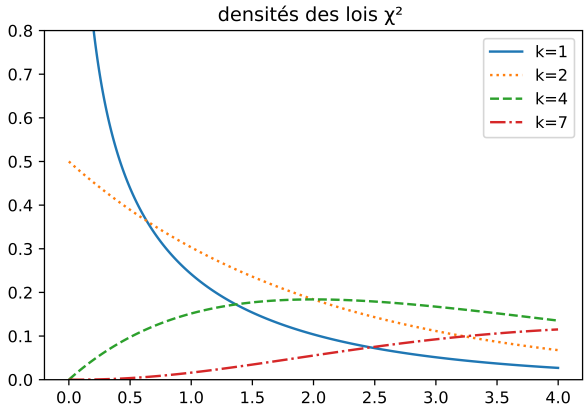
À programmer 1.

- Vous devez créer deux fichiers : un fichier `khi2.py`, celui-ci contiendra les fonctions que vous allez écrire, et un fichier `TP_nom.py` Dans ce deuxième fichier vous testerez les fonctions écrites sur les exemples proposés.
- Dans le fichier `khi2.py` importer le module `numpy.random`
- Dans le fichier `TP_nom.py` importer le fichier `khi2` et `matplotlib.pyplot`.

I Loi du χ^2

Définition 1 (Loi du χ^2 à k degrés de liberté)**.**
C'est une loi continue, suivie par une variable aléatoire X *qui est la somme des carrés de* k *variables aléatoires* Z_i *indépendantes qui suivent la loi normale centrée réduite.*

$$X = {Z_1}^2 + {Z_2}^2 + \cdots + {Z_k}^2$$



Remarque :

- Cette loi dépend d'un paramètre noté[1] k dans ce qui suit (et sur le graphique ci-dessus), appelé **nombre de degrés de liberté** : c'est un entier naturel.
- La table du χ^2 (www.delfaud.fr/Enac/Tables) permet, pour une probabilité α fixée à l'avance, de connaître la valeur x qui a la probabilité α d'être dépassée par la valeur de X, soit $\mathbf{P}(\mathbf{X} > \boldsymbol{x}) = \boldsymbol{\alpha}$.
  Attention : C'est la fonction d'antirépartition (ou de survie) qui est utilisée ici, et non pas la fonction de répartition utilisée dans le cours de probabilités. ⚠

Exemple : Si X suit une loi du χ^2 à 5 degrés de liberté, la valeur x qui aura une probabilité de 10% d'être dépassée est 9.24

Sur papier 1.
En utilisant la table donnée,

1. Si X suit une loi du χ^2 à 9 degrés de liberté, trouver la valeur x que X a 5% de «chances» de dépasser.
2. trouver la valeur x qu'une variable aléatoire suivant la loi du χ^2 à 11 degrés de liberté dépassera avec une probabilité de 1%.

II Simulation d'une loi du χ^2

À programmer 2.

- Écrire dans le fichier `khi2.py` une fonction `khi2(k,N)` qui renvoie une liste de N réalisations d'une variable aléatoire X suivant la loi du χ^2 à k degrés de liberté. Vous utiliserez la fonction `numpy.random.normal(0,1,k)` qui renvoie un tableau de k simulations suivant des lois normales centrées réduites indépendantes.
- Dans le programme principal, fixer la valeur de k puis construire un échantillon S de N valeurs de la loi du χ^2 à k degrés de liberté. (N = 10000)
- Tracer un histogramme de S et comparer l'allure obtenue avec les fonctions de densité ci-dessus, pour différentes valeurs du nombre k : on utilisera `plt.hist(S, bins =100, density = True )`

Remarque : On rappelle que l'option `density = True` permet d'ajuster les valeurs de façon à ce que l'aire totale de l'histogramme soit 1.

1. Il peut aussi être noté v

À programmer 3.

- Écrire une fonction `seuil_khi2(k, alpha)` dans le fichier `khi2.py` qui estime et renvoie, la valeur x telle que $P(X > x) = \alpha$.

  On commencera par obtenir un échantillon S de N (N très grand) réalisations d'une variable aléatoire X suivant la loi du χ^2 à k degrés de liberté),
  On peut ensuite trier S en ordre décroissant avec `S.sort(reverse = True)`.

  Cette probabilité sera estimée par une proportion : on renvoie la plus petite valeur x de l'échantillon vérifiant (nombre de valeurs $> x$) $/ N \leqslant \alpha$.
- Tester sur quelques valeurs de α et de k et comparer avec les données correspondantes de la table.

III Utilisation : test d'une hypothèse d'indépendance (test d'homogénéité)

III.1 Un exemple pas à pas

Dans une enquête effectuée auprès de 213 clients d'un magasin, 113 ont indiqué qu'ils se sont rendus dans le magasin à cause d'une publicité lue dans la presse. Parmi les 213 clients interrogés, 95 ont acheté un article, dont 55 qui avaient vu la publicité. À partir de cette enquête, peut-on considérer que l'achat d'un article est indépendant de la réalisation d'une campagne de publicité ?

Étape 1 : Énoncé de l'hypothèse nulle (H_0)**.** C'est une hypothèse d'égalité de comportement des deux populations : les différences constatées seraient dues simplement à la fluctuation d'échantillonnage, c'est ce que l'on appelle l'hypothèse nulle.

Dans notre exemple, H_0 : « il n'y a pas de différence significative entre les deux populations dans la réalisation d'un achat. »

Étape 2 : Choix du seuil de risque (α) Ici par exemple, nous fixons le seuil à 1% soit $\alpha = 0,01$. C'est la probabilité (le risque) de rejeter l'hypothèse alors qu'elle était vraie.

Étape 3 : Tableau de contingence Grâce aux résultats de l'enquête nous disposons des résultats observés (O_i) suivants :

Sur papier 2.
À l'aide de la description des résultats observés compléter le tableau de contingence suivant :

	Achat	Pas d'achat	Totaux
Publicité vue	$O_1 = 55$	$O_2 = \ldots\ldots$	113
Publicité non vue	$O_3 = \ldots\ldots$	$O_4 = \ldots\ldots$	$\ldots\ldots$
Totaux	95	$\ldots\ldots$	213

Sur papier 3.
Parmi les 213 clients interrogés, 95 ont acheté un article, soit 44,6%. En supposant l'hypothèse (H_0) vraie, les deux populations réagissent de la même manière ; dans ce cas

- sur les 113 clients ayant vu la publicité, $113 \times 0,446$ soit 50,4 clients auraient dû réaliser un achat ;
- sur les 100 clients n'ayant pas vu la publicité, $100 \times 0,446$ soit 44,6 clients auraient dû réaliser un achat.

Ce sont les *effectifs théoriques* (T_i). Compléter le tableau suivant :

	Achat		Pas d'achat		Totaux
Publicité vue	$O_1 = 55$	$T_1 = 50.4$	$O_2 = \ldots\ldots$	$T_2 = \ldots\ldots$	113
Publicité non vue	$O_3 = \ldots\ldots$	$T_3 = \ldots\ldots$	$O_4 = \ldots\ldots$	$T_4 = \ldots\ldots$	$\ldots\ldots$
Totaux	95		$\ldots\ldots$		213

Étape 4 : Mesure statistique Soit la variable aléatoire X nommée **distance du** χ^2 et définie par

$$X = \sum_{i \in \text{données}} \frac{(O_i - T_i)^2}{T_i}$$

Dans notre exemple $\mathbf{X} = \frac{(55-50,4)^2}{50,4} + \frac{(58-62,6)^2}{62,6} + \frac{(40-44,6)^2}{44,6} + \frac{(60-55,4)^2}{55,4} = 1,614$.

On montre que si H_0 est vraie, $\mathbf{X}$ suit une loi du χ^2 à k degrés de liberté, où $k = (\ell - 1)(c - 1)$ avec ℓ le nombre de lignes et c le nombre de colonnes du tableau de contingence.

Dans notre exemple on a donc $k = 1$.

Étape 5 : Détermination de la valeur critique x **et décision** La commande `seuil_khi2(1,0.01)` donne 6,630, Pour pouvoir rejeter l'hypothèse nulle avec une probabilité au plus de 1% de se tromper il faut que notre mesure statistique dépasse ce seuil. Ce n'est pas le cas, on ne peut pas rejeter l'hypothèse nulle : il est possible que la publicité n'ait pas eu d'influence sur la réalisation d'un achat.

III.2 Programmation

À programmer 4 (⛰ Calcul du tableau de contingence théorique)**.**
Écrire (dans `khi2.py`) une fonction `calcul_table_H0(O)` qui reçoit un tableau de contingence **observé** et qui renvoie le tableau **théorique** si H_0 est vraie.

À programmer 5.
Écrire une fonction `calcul_X` qui prend en paramètre le tableau des effectifs observés et celui des effectifs théoriques et qui renvoie la valeur de X, la «distance du χ^2 » entre effectifs observés et théoriques. Tester sur l'exemple ci-dessus, puis sur un des exemples ci-dessous.

III.3 Exemple 2

Dans un élevage bovin, comportant des bêtes de quatre races différentes, on a étudié 142 gestations et on en a déduit la distribution des effectifs donnée ci-dessous. On se demande s'il existe une relation entre les races et l'issue de la gestation.

	Vêlage	Avortement
Race 1	36	16
Race 2	18	10
Race 3	22	8
Race 4	24	8

TABLE 1 – Issue de la gestation

Sur papier 4.

1. Énoncer l'hypothèse nulle, on choisit $\alpha = 5\%$.
2. Donner le tableau de contingence théorique.
3. En utilisant les fonctions précédentes calculer la distance statistique X.
4. Donner la valeur du seuil critique x pour cette loi et pour $\alpha = 5\%$. Peut on rejeter l'hypothèse nulle?

III.4 Prise d'aspirine pour la prévention d'AVC

On trouve les données ci-dessous pour la survenue d'un AVC ischémique chez les femmes prenant préventivement de l'aspirine ou non, on veut savoir si la prise d'aspirine influence la survenue d'un AVC.

	Aspirine	Placebo
AVC ischémique	176	230
Pas d'AVC	21035	21018

TABLE 2 – Prise d'aspirine et AVC

Sur papier 5.
Énoncer l'hypothèse nulle. Peut on rejeter l'hypothèse nulle au seuil de 5%?

III.5 ⛰ Population d'oiseaux

Des scientifiques ont recensé les populations d'oiseaux sur des rives de rivières de Californie. Ces oiseaux sont séparés selon leur espèce et le type de végétation de la zone riparienne : dégradée ou restaurée. On veut savoir si la végétation a une influence sur la composition de la population aviaire. On a obtenu les résultats suivants

	Dégradée	Restaurée
Roitelet à couronne rubis	677	198
Bruant à couronne blanche	408	260
Bruant de Lincoln	270	187
Bruant à couronne dorée	300	89
Orite	198	91
Bruant chanteur	150	50
Tohi tacheté	137	32
Troglodyte de Bewick	106	48
Grive solitaire	119	24
Junco ardoisé	34	39
Chardonneret mineur	57	15
Autre	457	125

TABLE 3 – Population d'oiseaux en zone riparienne

Sur papier 6.
Énoncer l'hypothèse nulle ; on fixe $\alpha = 0.05$. Peut on rejeter l'hypothèse nulle au seuil de 5% ?

IV Test du χ^2 d'adéquation à une loi théorique

On étudie une population que l'on peut classer en n catégories, des considérations théoriques permettent de prévoir une distribution théorique . Dans ce qui suit on utilise un test du χ^2 pour déterminer si des résultats observés confirment la loi théorique.

Étape 1 Énoncer l'hypothèse 0 et en déduire les proportions attendues

Étape 2 Calculer les effectifs théoriques de chaque catégories

Étape 3 Calculer la mesure statistique de la même façon que précédemment

Étape 4 Cette mesure doit suivre une loi de χ^2.

Étape 3 On peut alors calculer la probabilité de faire une erreur en rejetant l'hypothèse nulle.

Attention : la détermination du nombre de degrés de liberté est différent, il est égal à $n-1$ où n est le nombre de catégories

IV.1 Premier exemple

On étudie deux caractères chez les cobayes

- la couleur gris (G, dominant) ou blanc (b, récessif).
- La longueur des poils lisse (L, dominant) ou rude (r, récessif)

Le croisement d'un cobaye ($G//G$, $L//L$) avec un cobaye ($b//b$, $r//r$) a donné à la deuxième génération naissance à 128 descendants dont les pelages se répartissent de la manière suivante :

- 78 au pelage gris et lisse $[G, L]$
- 19 au pelage blanc et rude $[b, r]$
- 26 au pelage blanc et lisse $[b, L]$
- 5 au pelage gris et rude $[G, r]$

Si on suppose que les deux caractères sont libres et si l'on admet que la transmission de ces deux caractères suit les lois de Mendel, les fréquences théoriques d'apparition $f_{GL}, f_{bL}, f_{br}, f_{Gr}$ des différentes catégories doivent être proportionnelles à $9,3,3$ et 1 .

À programmer 6.
1. Énoncer l'hypothèse nulle
2. Calculer les effectifs théoriques de chaque catégorie si cette hypothèse est vérifiée.
3. Calculer la mesure statistique $x = \sum_i \frac{(O_i - T_i)^2}{T_1}$
4. Cette mesure doit suivre la loi du χ^2 à 4-1 degré de liberté. Quelle est la probabilité que $[X \geqslant x]$? Que peut on en conclure.