
TP N° 2 : Test d'adéquation, test d'indépendance

Objectifs du TP : savoir lancer `jupyter notebook`, lancer / sauvegarder un notebook, utiliser les premières commandes `Python` standard, importer une librairie.

Quelques conseils de base : Pour chaque fiche de TP, nous utiliserons un script `jupyter notebook`. Il est conseillé de créer un répertoire HLMA408, puis un sous-répertoire pour chaque TP (*e.g.*, TP1, TP2, TP3). Dans un répertoire TP2, vous stockerez donc :

- le sujet de TP au format PDF,
- les fichiers de données (qui seront téléchargés automatiquement),
- le fichier `TP2-squelette.ipynb` qui contient les commandes (pratiquement) pré-remplies correspondant au TP, et que vous pouvez télécharger sur le site du cours.

Consignes :

Pour ce travail vous devez déposer un **unique** fichier sur le moodle du cours HLMA408. Vous rendrez un unique fichier d'extension `ext=ipynb`. Attention, vous devrez veiller à ce que le format soit correct : si le fichier ne peut pas être ouvert par le correcteur avec `jupyter-notebook`, la note sera de zéro.

Pour faciliter la correction, le nom du fichier devra respecter le format suivant :

`nom_fichier = tp_note2_hlma408_gr_nbgr_prenom_nom.ipynb`

le tout en minuscule et sans accent ni espace. Vous remplirez votre nom, prénom, le numéro de groupe qui vous concerne (remplacer `nbgr` par C, D ou E) de manière adéquate¹. Un point de malus sera appliqué pour les fichiers dont le format est erroné.

Vous devez charger votre fichier sur Moodle, avant le mardi 10/02/2020, 23h55. La note totale est sur 20 points, répartis comme suit :

- qualité des réponses aux questions : 14 pts,
- qualité de rédaction et d'orthographe : 1 pt,
- qualité des graphiques (légendes, couleurs) : 2 pts
- qualité d'écriture du code (noms de variable clairs, commentaires, code synthétique, etc.) : 1 pt
- Rendu reproductible et absence de bug : le code doit s'exécuter sur la machine du correcteur sans manipulation de sa part (par exemple le correcteur n'est pas supposé aller chercher les fichiers sur internet, les enregistrer, etc.). On veillera donc à ce que le chargement des bases de données soit automatisé : 2 pts

Les personnes qui n'auront pas soumis leur devoir sur Moodle avant la limite obtiendront **zéro**.

1 Test d'adéquation du χ^2 à une loi donnée

La fonction `chisquare` du package `scipy.stats` permet de mettre en œuvre un test du χ^2 avec `Python`, voir par exemple <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html> pour son fonctionnement.

- 1) Importer la fonction `chisquare` mentionnée.

1.1 Loi discrète

Un couple de cobayes à pelage gris et lisse a donné naissance à 64 descendants. Leurs pelages se répartissent ainsi :


- 33 gris et lisses,

1. Exemple : Joseph Salmon est dans le groupe C, son TP s'appelle `tp_note2_hlma408_gr_C_joseph_salmon.ipynb`

- 13 blancs et lisses,
- 15 gris et rudes,
- 3 blancs et rudes.

Le modèle de Mendel donne les probabilités données dans le tableau suivant :

type	effectif observé	probabilité théorique	effectif théorique
gris et lisse	33	9/16	36
blanc et lisse	13	3/16	12
gris et rude	15	3/16	12
blanc et rude	3	1/16	4

- 2) Saisir les données dans un vecteur ( : *array*) ainsi

```
import numpy as np
cobaye = np.array([33, 13, 15, 3])
```

après avoir importé `numpy` sous le nom de `np`.

- 3) Saisir les probabilités théoriques :

```
mendel = np.array([9/16, 3/16, 3/16, 1/16])
```

- 4) Utiliser la commande `chisquare` pour conclure si l'on accepte ou non l'hypothèse que le modèle de Mendel est valide ou non. (1 pt)

1.2 Loi continue et regroupement en classes

Au préalable, nous devons d'abord comprendre deux fonctions de `pandas` : `cut` et `value_counts`. La fonction `cut` permet de faire du regroupement par classes d'intervalles. Voyons un exemple sur un échantillon simulé de 50 observations provenant de 50 tirages aléatoires de lois gaussiennes centrées réduites. Les classes sont $] -\infty; -2]$, $] -2; -1]$, \dots , $]1; 2]$ et $]2; +\infty]$:

```
import pandas as pd # import de pandas
echantillon_gaussien = np.random.randn(50) # génération de données synthétiques
bins = [-np.inf, -2, -1, 0, 1, 2, np.inf]
echantillon_regroupe = pd.cut(echantillon_gaussien, bins=bins)
```

- 5) Affichez `echantillon_regroupe` pour comprendre ce que contient l'objet créé. (0.5 pt)

La fonction `value_counts` calcule les effectifs (le nombre d'occurrences) de chaque valeur possible dans un échantillon. Si on reprend l'échantillon regroupé calculé ci-dessus, et que l'on applique la fonction `value_counts`, on obtient les effectifs de chacune des classes :

```
effectifs_classes = pd.value_counts(echantillon_regroupe)
```

ou de manière équivalente :

```
effectifs_classes = echantillon_regroupe.value_counts()
```

On souhaite maintenant mettre en place un test du χ^2 pour vérifier que ces données simulées proviennent bien d'une loi normale centrée réduite.

La fonction `norm.cdf` de `scipy.stats` permet de calculer les probabilités de la forme $\mathbb{P}(Z \leq x)$ lorsque Z suit une loi normale centrée réduite et x est un nombre réel. Par exemple `norm.cdf(0)` donne 0.5 et `norm.cdf(1.96)` donne environ 0.975.

- 6) Que vaut la probabilité de chacune des classes $] -\infty; -2]$, $] -2; -1]$, \dots , $]1; 2]$ et $]2; +\infty]$ sous la loi normale centrée réduite? Vous ferez ce calcul avec la fonction `np.diff`. (1 pt)

- 7) Faire un test du χ^2 pour regarder si l'on peut considérer que les effectifs observés des différentes classes proviennent d'un regroupement par classe d'une loi normale centrée réduite. (1pt)
- 8) Refaire ce qui précède avec un nouvel échantillon simulé de taille 100 et les classes suivantes : $]-\infty ; -1.5]$, $]-1.5 ; -0.5]$, $]-0.5 ; 0.5]$, $]0.5 ; 1.5]$ et $]1.5 ; +\infty[$, le tout dans une unique cellule. Relancez la cellule plusieurs fois ? Observez vous des différences ? Si oui d'où viennent-elles ? (1 pt)

1.3 Un autre test de normalité (test de Shapiro-Wilk)

Le test de Shapiro-Wilk² permet de décider entre

$$\mathcal{H}_0 : "X \text{ suit une loi gaussienne}" \quad \text{vs.} \quad \mathcal{H}_1 : "X \text{ ne suit pas une loi gaussienne}"$$

Lire la page d'aide de la commande `shapiro`³ pour comprendre comment celle-ci fonctionne

- 9) Exécuter ce test sur l'échantillon de taille 100 que l'on avait généré et conclure. (0.5pt)
- 10) Exécuter ce test sur un échantillon de taille 100, d'observations tirées selon une loi exponentielle (d'espérance 1) et conclure. (0.5pt)

2 Test d'adéquation du χ^2 à une famille de lois

Python ne dispose pas de fonction toute faite pour répondre à ce problème. Pour cela, nous allons reprendre l'exemple du cours, concernant le test d'adéquation à la famille des lois de Poisson et mettre en place un test comparant

$$\mathcal{H}_0 : "X \text{ suit une loi de Poisson}" \quad \text{vs.} \quad \mathcal{H}_1 : "X \text{ ne suit pas une loi de Poisson}".$$

On rappelle que la variable X représente le nombre de palindromes dans une unité de longueur de 4000 paires de bases.

- 11) Importer et mettre en forme les données du fichier `hcmv.data` dans un dataframe nommé `df_hcmv`. (0.5pt)
- 12) Ce test suppose que l'on commence par estimer le paramètre λ de la loi de Poisson. Donner un estimateur naturel dans ce contexte. Calculer la valeur numérique sur le jeu de données et l'enregistrer dans `lambda_hat`; on pourra utiliser qu'il y a `n_basis = 229354` et compter le nombre de palindromes. (0.5pt)

Les commandes suivantes permettent de compter le nombre de palindromes dans chacune des régions de longueur 4000 paires de bases.

```
new_bins = np.concatenate([[ -np.inf], np.arange(1, n_basis, step=4000), [np.inf]])
hcmv_regroupe = pd.cut(df_hcmv['location'], bins=new_bins)
counts_palindrome = hcmv_regroupe.value_counts(sort=False)
```

- 13) Regrouper maintenant les palindromes par fréquence d'apparition dans les boîtes de longueur 4000 bp, en utilisant les bornes $-\infty, 2, 3, 4, 5, 6, 7, 8, +\infty$. On notera `count_regroupe` cette variable; on pourra utiliser la variable `counts_bin = [-np.inf, 2, 3, 4, 5, 6, 7, 8, np.inf]` et la fonction `cut` pour cela. (0.5pt)
- 14) Que fait la commande ci-dessous ? (0.5pt)

```
from scipy.stats import poisson
bins_poisson = [-np.inf, 2, 3, 4, 5, 6, 7, 8, np.inf]
np.diff(poisson.cdf(bins_poisson, lambda_hat * 4000))
```

- 15) On veut utiliser la fonction `chisquare` ici avec

2. pour les lecteurs curieux il est décrit ici : https://fr.wikipedia.org/wiki/Test_de_Shapiro-Wilk; on notera que la statistique ≈ 1 sous \mathcal{H}_0 .

3. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html>

```
chisquare(count_regroupe, count_theoric, ddof=1)
```

où `count_theoric` est le vecteur des effectifs "théoriques" de chaque classe. Que signifie `ddof = 1` dans l'aide de la fonction `chisquare`? (0.5pt)

- 16) Conclure sur le test initial en calculant les quantiles d'un χ^2 avec les fonctions `chi2.cdf`. (0.5pt)

3 Test d'indépendance du χ^2

Il existe un test d'indépendance du χ^2 que nous n'avons pas vu en cours ni en TD. Son but est de discriminer entre les deux hypothèses suivantes :

$$\mathcal{H}_0 : "X \text{ et } Y \text{ sont indépendantes}" \quad \text{vs.} \quad \mathcal{H}_1 : "X \text{ et } Y \text{ sont liées}"$$

lorsque X et Y sont deux variables aléatoires discrètes. Rappelons que deux variables sont dites indépendantes si connaître la valeur de l'une ne donne aucune information sur la valeur de l'autre. Par exemple, si on lance deux dés simultanément, l'un rouge, l'autre blanc et que l'on modélise par X le résultat du dé blanc et Y le résultat du dé rouge, alors X et Y sont indépendantes.

Pour cela, on utilise la fonction `chi2_contingency` (du module `scipy.stats`) sur un tableau de contingence croisé entre deux variables. Il s'agit d'un tableau dont chaque ligne correspond à une valeur possible de X et chaque colonne à une valeur possible de Y . Dans la case sur la ligne x et la colonne y , on compte le nombre d'observations où l'on voit simultanément $X = x$ et $Y = y$. De tels tableaux de contingence se calculent avec la fonction `pd.crosstab`. Par exemple, sur les données de la table `babies`⁴, chargées de la manière suivante :

```
df_babies = pd.read_csv("babies23.data", skiprows=38, sep='\s+')
```

il suffit d'exécuter :

```
pd.crosstab(df_babies['ed'], df_babies['smoke'])
```

pour obtenir un tableau de contingence croisant le niveau d'éducation des mères avec leur statut tabagique.

- 17) Calculer le tableau de contingence entre les variables `ed` et `smoke`. Que pensez-vous du résultat ? Enregistrer ce tableau dans un dataframe que l'on va appeler `cont_table`. (0.5pt)
- 18) Lancer `chi2_contingency(cont_table)` et conclure. (0.5pt)
- 19) En parcourant l'aide de la fonction `chi2_contingency` proposer le même test en utilisant la fonction `chisquare`. (0.5pt)

4 Courbe ROC (4pts)

Dans cette question on s'intéresse à la courbe ROC, décrite par exemple ici https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

Pour cela, on considère un cadre avec deux lois gaussiennes $\mathcal{N}(\mu_0, \sigma_0^2)$ et $\mathcal{N}(\mu_1, \sigma_1^2)$, et un test qui consiste à valider l'hypothèse H_0 "le processus sous-jacent suit la loi la $\mathcal{N}(\mu_0, \sigma_0^2)$ " si une statistique de test observée est plus petite qu'une grandeur q (et inversement à choisir l'hypothèse H_1 "le processus sous-jacent suit la loi la $\mathcal{N}(\mu_1, \sigma_1^2)$ " si la statistique est plus grande que q).

On souhaite visualiser sur un graphique (avec `subplots` par exemple) les zones d'erreurs de première espèce (ou faux positifs, FP) et de seconde espèce (ou faux négatifs, FN) à l'aide d'un widget.

On créera donc un widget avec `interact` ayant 5 curseurs :

- un curseur pour μ_0 (variant de 1 à 3),
- un curseur pour σ_0^2 (variant de 0.2 à 2),
- un curseur pour μ_1 (variant de 1 à 5),

4. <http://josephsalmon.eu/enseignement/datasets/babies23.data>

- un curseur pour σ_1^2 (variant de 0.2 à 2),
- un curseur pour le seuil q (variant de 0 à 8).

Ces curseurs permettront d'afficher sur un graphique de type `subplot`, la représentation visuelle du test et des deux densités gaussiennes (à droite), et la courbe ROC associée ainsi que le point correspondant au seuil q (à gauche). On pourra s'inspirer de l'exemple fourni ci-dessous :

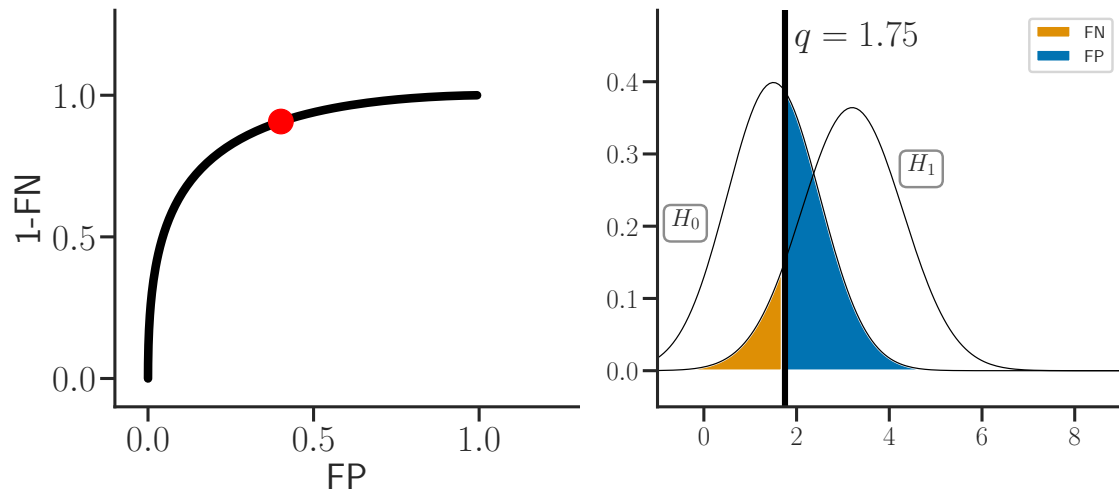


FIGURE 1 – Courbe ROC et représentation des gaussiennes et du seuil du test.