GABAS Clément

Rapport de stage d'application en statistiques de 2A

Structure d'accueil : Observatoire Midi-Pyrénées

Thème du stage : blablabla

Maître de stage : Rémi Cabanac

Référent Pédagogique : Rémi Pépin

Introduction

Éléments de physique et de cosmologies nécessaires à la compréhension des enjeux du sujet

Comment sommes-nous passés d'un univers sans matière à un univers structuré, via la seule gravitation, en seulement 15 milliard d'années ? Pour bien comprendre les enjeux de cette question, intéressons nous d'abord à la structure de l'univers.

La force de gravité, ou gravitation, est la force à l'origine du regroupement des particules de gaz qui forment les étoiles, du regroupement des étoiles qui forment les galaxies, et du regroupement des galaxies et de diverses matières qui forment des structures de tailles bien larges. Ces groupements de galaxies forment de longs filaments, parsemés de vide, ressemblant à une toile d'araignée. C'est ainsi qu'on se réfère couramment à la *Toile Cosmique*. En langage scientifique, nous préférerons l'appellation de **Large Scale Structure** (LSS) de l'univers, qui fait donc référence à cet ensemble de structures de galaxies, sur des échelles bien plus grandes que le simple amas de galaxies.
Lorsque l'on observe la LSS de l'univers, nous observons les étoiles dont la lumière nous parvient. Cependant, la luminosité des étoiles, ainsi que leurs distances à la Terre, jouent un rôle dans notre capacité à les observer. Ainsi, pour savoir à quelle distance un objet céleste se trouve de la Terre, on ne peut pas se contenter d'observer sa luminosité.
Du fait de l'expansion de l'univers et de la dilatation de l'espace-temps induite, un décalage vers les grandes longueurs d'ondes (et donc vers le rouge pour la lumière visible) du spectre des objets lointains est observé. On parle alors de **décalage vers le rouge** (ou ***redshift***). Pour une galaxie lointaine, la mesure du redshift de l'objet permet d'avoir une idée de sa distance. Pour les galaxies les plus proches, leur mouvement propre est non négligeable devant leur mouvement induit par l'expansion de l'univers et il faut donc utiliser une autre méthode de calcul des distances.
Lorsque l'on observe les différentes valeurs de redshift des objets célestes, on se rend compte que la vitesse d'expansion de l'univers est non linéaire, et plus particulièrement que cette expansion accélère. Ainsi, pour observer les objets les plus anciens, on regarde ceux qui ont le plus grand redshift. En particulier, lorsque l'on observe les objets ayant un redshift de 300 , nous observons en réalité 380 000 ans après le big bang. L'espace est alors a une température de 3000K et est assez froid pour permettre aux premiers atomes de se créer. C'est le **rayonnement fossile**, ou **fond diffus cosmologique**.
Ce rayonnement correspond à la plus ancienne lumière observable depuis la Terre. Son observation nous renseigne sur l'état d'homogénéité et d'isotropie spatiale de l'univers primordial. Nous apprenons ainsi que l'univers primordial est très homogène mais présente de légère variations qui, au fil de son expansion, ont résultées en la création de la Toile Cosmique. Une meilleure connaissance de l'univers initial nous permettrait de réaliser de meilleures modélisations de son évolution au fil du temps, et donc résulterait en une meilleure connaissance de la répartition des objets célestes dans l'espace actuel.

Contexte de l'étude

Ici, expliquer en quoi la connaissance de la répartition des lentilles gravitationnelles est reliée à la connaissance de l'homogénéité de l'univers primitif.

Présentation des données CFHTLS-T0007

Pour ce travail, nous disposons des données issues de la $7^{\text{ème}}$ et dernière version du sondage spatial *Canada-France-Hawaii Telescope Legacy Survey*. Plus précisément, nous exploiterons les données du CFTHLS-T0007 Wide, constitué de 171 pointeurs profonds dits MegaCam, chacun fournissant des mesures sur 1×1 deg^2 [1]. Cet ensemble de MegaCam produit une cartographie des objets célestes d'une taille approximative de 155 deg^2 répartie sur 4 parcelles distinctes (cf. W1-4 sur Figure 1).

Les objets observés sont des galaxies occupant un cône de lumière projeté sur le ciel. Au premier ordre, ces galaxies sont divisibles en deux populations : une population d'avant plan et une population d'arrière-plan. Nous considérons les objets d'avant-plan comme ceux ayant un redshift inférieur à 0.3, et ceux d'arrière plan ayant un redshift supérieur à ? ?.

1. Chaque MegaCam observe sur un champs 1 deg × 1 deg avec 0.186"/pixel (19354 x 19354 pixels)

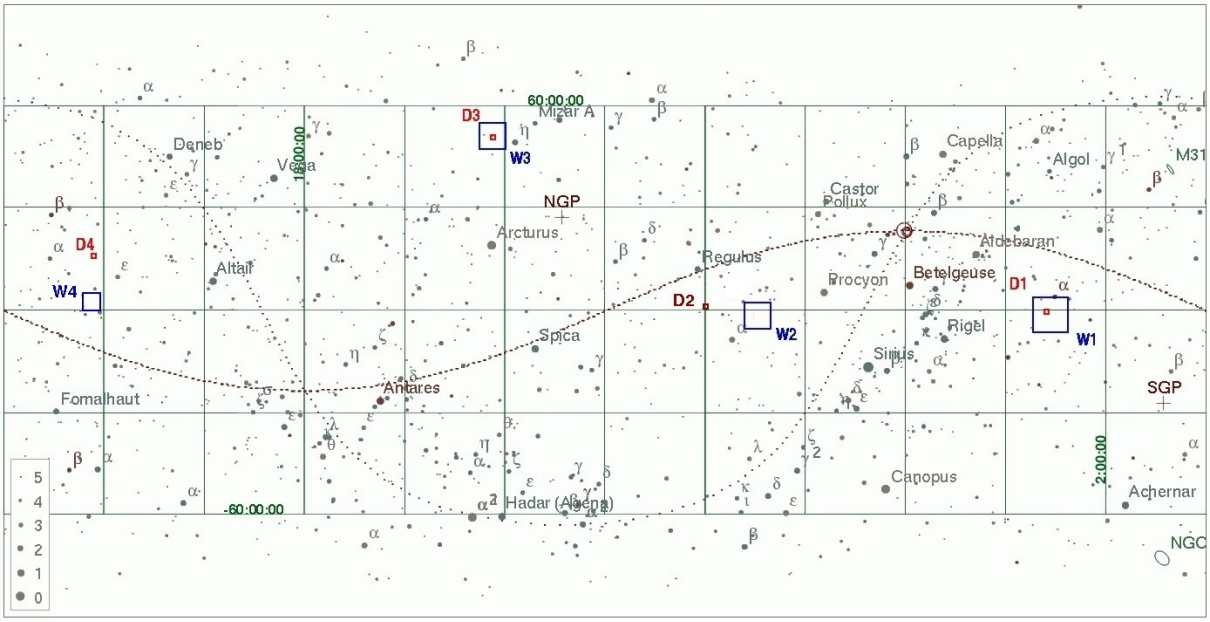

Fig. 1 – Disposition des 4 champs larges W1-4 dans le ciel

Source: [Hud+12]

En de rares occasions, certaines galaxies d'arrière-plan sont très exactement alignées géométriquement avec des galaxies d'avant-plan dans notre ligne de visée et créent ce qu'on appelle des mirages gravitationnels. L'ensemble formé de la galaxie d'avant-plan et du mirage gravitationnel de la source d'arrière-plan est appelé **lentille gravitationnelle**. Nous disposons d'environ 130 lentilles gravitationnelles réparties sur 4 champs du ciel couvrant 140 deg^2 parmi les millions d'objets. Nous considérons que notre échantillon observé est complet et pur, nous avons observé l'ensemble des lentilles gravitationnelles présentes dans le plan.
La première mission de ce stage est d'étudier la répartition spatiale des lentilles gravitationnelles. En effet

Table des matières

1 Étude du biais d'observation

On commence par étudier un potentiel biais d'observation des lentilles gravitationelles. En effet, il est possible que l'observation de lentilles soit favorisée par certaines conditions d'observation. Ainsi, il nous faut vérifier que les conditions d'observations des différents champs du ciel et que les valeurs d'observation des lentilles sont concordantes. Pour ce faire, on étudie la répartition des lentilles en fonction des valeurs de seeing et de temps d'exposition, que l'on compare à la répartition du seeing et de l'exposition des différents champs observés.

1.1 Étude du biais selon le seeing

Ici, bien que les deux histogrammes semblent se chevaucher de façon plus ou moins homogène, on remarque une sur-densité aux alentours d'un seeing de 0.5, ainsi qu'une autre aux alentours de 0.7, ainsi qu'une sous-densité pour un seeing de 0.6. Notons également qu'il y a une coupure nette à 0.466 de seeing et une autre à 0.84.

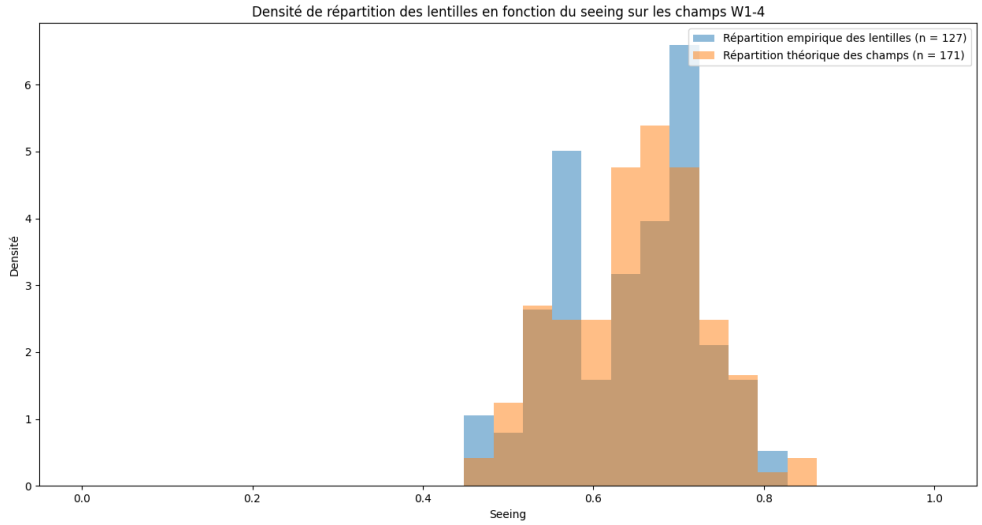

Fig. 2 – Histogramme des densités de répartitions en fonction du seeing pour les lentilles et les champs observés

Source: Données CFHTLS [Hud+12]

L'objectif est de voir si la répartition des lentilles en fonction du seeing suit la même loi que la répartition théorique du seeing des mesures. Pour cela, on va utiliser un test-t de Welch [WEL47].
Après vérification de la normalité des échantillons (Figure 4), on va tester l'hypothèse nulle suivante

$$H_0 : m_L = m_W \text{ vs } H_1 : m_L \neq m_W$$

avec m_L et m_W les moyennes empiriques des échantillons de lentilles et de champs.
Dans notre cas, il s'avère qu'on ne peut pas rejeter l'hypothèse nulle. On conclura alors que les 2 échantillons suivent la même loi (Listing 1, Listing 2 Figure 5). Ainsi, il ne semble pas y avoir de biais d'observation selon le seeing.

1.2 Étude du biais selon le temps d'exposition

Cette fois, on observe une légère sur-densité pour les expositions entre 80 et 90.

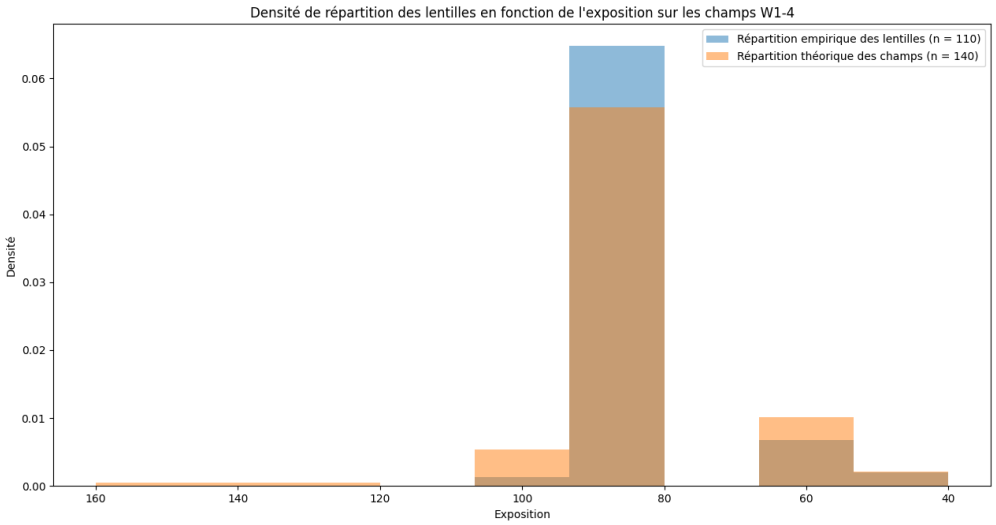

Fig. 3 – Caption

Source: Données CFHTLS [Hud+12]

Ici, les analyses de nos échantillons (Figure 6) montrent que ces derniers ne suivent pas une loi normale. Pour comparer les lois de répartitions, nous ne pouvons donc pas utiliser les tests paramétriques de Student ou de Welch qui présupposent de la répartition gausienne des échantillon. Nous allons donc utiliser des tests non paramétriques, qui ne font pas d'hypothèse sur la forme des distributions.
En particulier, nous allons utiliser le test des rangs signés de Wilcoxon [Wil45] qui permet la comparaison de deux distributions de variables continues.
Plus précisément, pour 2 échantillons $X = (X_1, \cdots, X_n)$ et $Y = (Y_1, \cdots, Y_n)$, on calcule la différence $D = (X_1 - Y_1, \cdots, X_n - Y_n)$. On note θ la médiane de l'échantillon D et on effectue le test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

On ordonne alors les $|D_i|$ et on note R_i le rang de chaque D_i dans l'échantillon ordonné. On défini l'indicatrice de signe ψ_i qui vaut $+1$ si $D_i > 0$ et -1 si $D_i < 0$. Si $D_i = 0$, on l'exclu de l'échantillon. Après avoir effectué toutes les exclusions, notre échantillon est de taille $m \leq n$.
On a alors la statistique de test $T = \sum_i^n R_i \psi_i$. Sous l'hypothèse nulle, T ne suit pas une loi usuelle mais une distribution spécifique d'espérance nulle et de variance $\frac{m \times (m+1) \times (2m+1)}{6}$.
Le test consiste alors en le rejet de H_0 si $T > T_{\text{critique}}$ avec T_{critique} disponible dans des tables de références [Low99].
Dans notre cas, on ne peut pas rejeter l'hypothèse nulle. On conclura alors que les 2 échantillons suivent la même loi (Listing 3, Listing 4). Ainsi, il ne semble pas y avoir de biais d'observation selon le seeing.

A Annexe - biais d'observation

A.1 Seeing

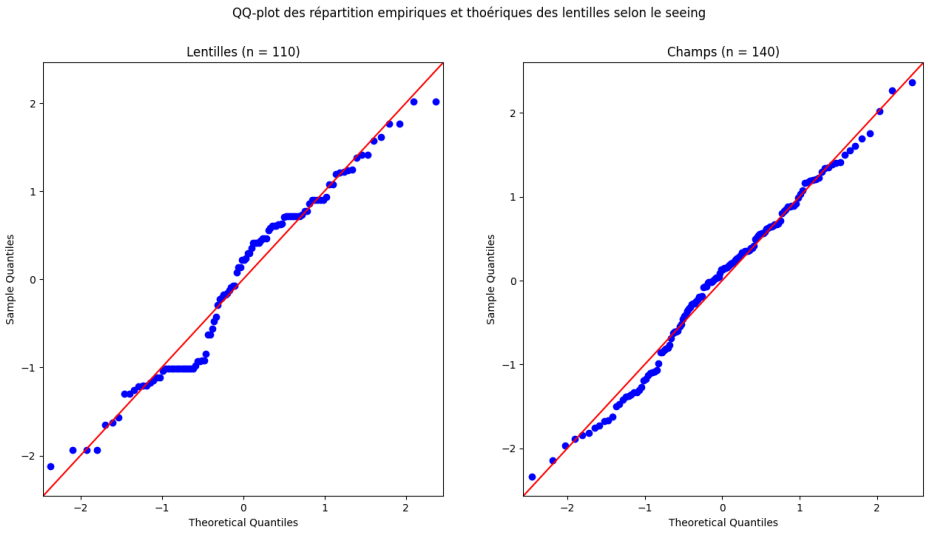

Fig. 4 – Normalité des échantillons en fonction du seeing

Source: Données CFHTLS [Hud+12]

```python
# T test de Welch
_, p_value = scipy.stats.ttest_ind(seeing_list_lentille_01,
                                   seeing_list_obj_01, equal_var=False)
p_value
```

Listing 1 – t-test de welch entrée python

```
>>> _, p_value = scs.ttest_ind(seeing_list_lentille_01,
...                            seeing_list_obj_01, equal_var=False)
>>>
>>> p_value
0.9999999999999961
>>> # on ne rejette pas H_0
```

Listing 2 – t-test de welch sortie terminal

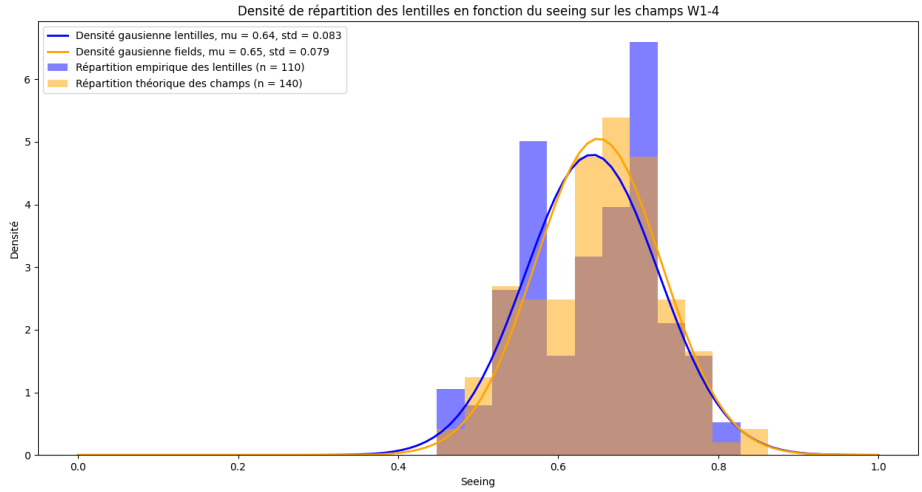

Fig. 5 – Densités Gausiennes des échantillons de lentille et de fields en fonction du seeing

Source: Données CFHTLS [Hud+12]

A.2 Temps d'exposition

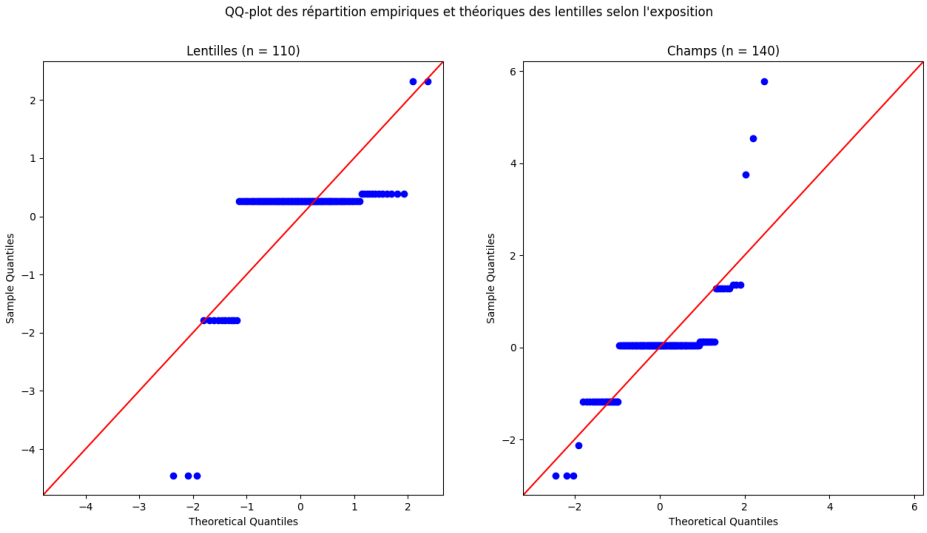

Fig. 6 – Caption

Source: Données CFHTLS [Hud+12]

```python
N = 10_000
p_values = np.zeros(N)
for i in range(N):
    np.random.shuffle(expo_list_obj_noNA)
    expo_list_obj_noNA_len = expo_list_obj_noNA[0:len(expo_list_lentille_noNA)]

    _, p_values[i] = scs.wilcoxon(x=expo_list_lentille_noNA, y=expo_list_obj_noNA_len)
np.mean(p_values)
```

Listing 3 – test non paramétrique des rangs signés de Wilcoxon **entrée python**

```
>>> N = 10_000
>>> p_values = np.zeros(N)
>>> for i in range(N):
...     np.random.shuffle(expo_list_obj_noNA)
...     expo_list_obj_noNA_len = expo_list_obj_noNA[0:len(expo_list_lentille_noNA)]
...     _, p_values[i] = scs.wilcoxon(x=expo_list_lentille_noNA, y=
    expo_list_obj_noNA_len)
...
>>> np.mean(p_values)
0.5645733799735497
>>> # on ne rejette pas H_0
```

Listing 4 – test non paramétrique des rangs signés de Wilcoxon **sortie terminal**

Bibliographie

[Hud+12] Patrick Hudelot et al. *T0007 : The Final CFHTLS Release, Executive Summary*. Report. TERAPIX-CFHTLS, 2012. URL : https://cfhtls.calet.org/T07/doc/T0007-doc.pdf.

[Low99] Richard Lowry. *Concepts & Applications of Inferential Statistics - The Wilcoxon Signed-Rank Test*. 1999. URL : http://vassarstats.net/textbook/ch12a.html.

[WEL47] B. L. WELCH. "The Generalization of "Student's problem when several different population variances are involved". In : *Biometrika* 34.1-2 (jan. 1947), p. 28-35. ISSN : 0006-3444. DOI : 10.1093/biomet/34.1-2.28. eprint : https://academic.oup.com/biomet/article-pdf/34/1-2/28/553093/34-1-2-28.pdf. URL : https://doi.org/10.1093/biomet/34.1-2.28.

[Wil45] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In : *Biometrics Bulletin* 1.6 (1945), p. 80-83. ISSN : 00994987. URL : http://www.jstor.org/stable/3001968.