Atelier pédagogique
M21 - Collectionneur de coupons

Combien d'images dois-je acheter pour finir ma collection des joueurs du FC Tours ? Ou encore, combien de temps mettrais-je pour mélanger parfaitement un jeu de cartes ?

COLEAU Victor, COUCHOUD Thomas, GRODECOEUR Clément

Table des matières

1 Introduction

Dans le cadre des ateliers pédagogiques, nous avons été amenés à réfléchir sur le sujet du collectionneur de coupons et autres problèmes similaires.

Ce problème fait partie du domaine des probabilités. Notre objectif est de déterminer le temps (ou nombre de tirages) nécessaire afin d'obtenir tous les éléments d'un ensemble prédéfinit en réalisant des tirages aléatoires d'éléments de cet ensemble. Notre compte-rendu comportera trois parties : la première basée sur des simulations informatiques, la seconde expliquant ces résultats grâce à des modèles mathématiques, et pour finir une comparaison entre les résultats des deux précédentes parties.

2 Simulations informatiques

2.1 Collection d'images

2.1.1 Problème

Nous cherchons à obtenir toutes les images contenues dans un set prédéfinit. Pour cela, on tire ces images une par une aléatoirement. Notre problème est de déterminer combien de ces tirages nous devrons réaliser afin d'acquérir le set complet.

2.1.2 Programme

Variables tailleDeck → Taille du set final, soit le nombre de cartes différentes à obtenir.
nombreTests → Nombre de simulations à réaliser pour obtenir notre moyenne

Fonctionnement On initialise une variable "deck" vide qui représente les cartes dont nous disposons (sans doublons). Tant que notre deck n'est pas complet (toujours sans doublons), nous générons aléatoirement un entier (qui correspond à une carte précise). Si celui-ci n'est pas déjà dans notre deck, nous l'ajoutons, sinon on ne fait rien. A chaque fois que nous générons un entier, nous incrémentons de 1 un compteur qui représente le nombre de tirages réalisés.

Toutes ces étapes sont répétées le nombre de fois comprit dans la variable "nombreTest". Par la suite, nous affichons la moyenne de tous les compteurs ainsi obtenus.

2.1.3 Resultats attendus

On s'attend à ce que lors des premiers tirages, lorsque l'on n'a que peu ou pas de cartes, les images à obtenir soient piochées rapidement. En effet, n'en possédant aucune (ou presque), la probabilité d'en obtenir une nouvelle est très élevée. De cette manière, plus on possède de cartes, moins on a de chances d'en obtenir une nouvelle.

2.1.4 Résultats obtenus

Nombre de répétitions \ Nombre d'images à obtenir	5	10	20	50	100	200	500
10	8,70	21,10	49,90	194,00	394,60	937,60	2889,10
50	8,54	22,94	57,60	177,56	433,46	1036,48	2846,40
100	8,81	22,19	54,42	175,60	419,08	967,56	2842,38
500	8,73	22,16	55,00	177,40	423,98	986,93	2923,09
1000	8,79	22,18	55,91	181,67	421,15	971,40	2907,56
5000	8,66	22,12	55,65	179,60	424,85	979,78	2895,74
10000	8,70	22,40	55,49	180,24	424,52	982,89	2896,80
50000	8,70	22,21	55,56	179,34	423,75	981,07	2902,00
100000	8,69	22,26	55,59	179,39	423,86	981,18	2901,93

TABLE 1 – Résultats des différentes expériences

D'après ce tableau, nous pouvons remarquer deux choses :
- Plus on augmente le nombre de cartes à obtenir, plus le nombre de tirages moyen nécessaire augmente.
- Plus on augmente le nombre de répétitions, plus la valeur moyenne est précise et diffère peu entre les différentes expériences dans les mêmes conditions.

Nous pouvons alors nous demander si, en répétant un nombre de fois infini la simulation, la moyenne tendrait vers une valeur limite. Pourrions-nous trouver une formule mathématique permettant de calculer cette limite ?

2.2 Mélange d'un jeu de cartes

Par analogie de l'expérience précédente, nous pouvons imaginer un problème d'apparence similaire bien que ces conditions initiales soient différentes.

2.2.1 Problème

Nous cherchons à mélanger tous les cartes contenues dans un deck prédéfinit. Pour réaliser cela, nous prenons la carte se trouvant au-dessus du tas et l'insérons à une position aléatoire dans le paquet. Notre problème est de déterminer combien de battages nous devons alors effectuer pour avoir un deck parfaitement mélangé. Nous considérerons qu'il sera mélangé lorsque la carte initialement au-dessous (appelée carte témoin) se retrouvera au-dessus du paquet.

2.2.2 Programme

Variables nombreCartes → Taille du paquet de cartes à mélanger
nombreMelanges → Nombre de simulations à réaliser pour obtenir notre moyenne

Fonctionnement Nous initialisons une variable "cartes" vide qui représente le paquet. Puis, nous la remplissons de cartes (autant de cartes que la valeur de la variable "nombreCartes") ordonnées. Cela correspond au deck initial que nous devons par la suite mélanger. Tant que la carte témoin n'est pas arrivée au-dessus du paquet, nous effectuons un battage.

Le battage consiste à générer aléatoirement une position dans le deck. On déplace ensuite la carte du dessus à cette position. A chaque fois que nous générons une position, nous incrémentons de 1 un compteur qui représente le nombre de battages réalisés.

Toutes ces étapes sont répétées le nombre de fois comprit dans la variable nombreMelanges. Par la suite nous calculons et affichons la moyenne de tous les compteurs ainsi obtenus.

2.2.3 Resultats attendus

On s'attend à un effet "inverse" de l'expérience précédente. C'est à dire que les premiers battages n'ont que très peu de chance (faible probabilité) de modifier la position de la carte témoin car la probabilité que la position choisie soit au-dessous de celle-ci est très faible. De cette manière, plus la carte est déjà "haute", plus on a de chances de la faire remonter d'un cran supplémentaire à chaque coup.

2.2.4 Résultats obtenus

Taille du paquet de cartes / Nombre de répétitions	5	10	20	50	100	200	500
10	10,40	27,20	59,30	205,00	489,50	1316,30	3280,70
50	10,38	28,84	74,88	219,86	489,20	1210,62	3381,50
100	10,67	26,74	69,90	227,72	499,59	1157,74	3441,72
500	10,38	28,77	68,86	222,65	518,43	1169,59	3449,83
1000	10,46	28,41	70,87	219,19	519,82	1169,67	3397,03
5000	10,42	28,07	70,98	223,56	518,67	1171,76	3383,32
10000	10,47	28,04	70,84	223,67	518,74	1173,14	3389,48
50000	10,43	28,30	70,98	223,95	516,82	1175,32	3393,94
100000	10,42	28,31	70,92	224,05	517,94	1174,70	3393,99

TABLE 2 – Résultats des différentes expériences

D'après ce tableau, nous pouvons remarquer deux choses :

- Plus on augmente le nombre de cartes à battre, plus le nombre de battages moyen nécessaire augmente.
- Plus on augmente le nombre de répétitions, plus la valeur moyenne est précise et diffère peu entre les différentes expériences dans les mêmes conditions.

Bien que les expériences soient similaires, on peut penser que les modèles mathématiques de ces deux problèmes sont différents.

3 Modèles mathématiques

3.1 Collectionneur de coupons

3.1.1 Modèle mathématique

Commençons par le problème du collectionneur de coupons.

Soit N le nombre d'éléments à obtenir (nombre final de coupons).

On définit T_i comme étant le temps (nombre de tirage(s)) au bout duquel on obtient le i^e nouvel élément depuis le début des tirages.

Dans notre cas, on cherche à calculer T_N. Nous remarquons déjà l'apparition d'une complication avec ce calcul. En effet déterminer T_1 est assez simple puisque la probabilité de tirer la première carte est $T_1 = 1$. Cependant calculer T_2 est déjà beaucoup plus difficile.

Afin d'y remédier, on va s'intéresser au temps séparant deux apparitions de nouveaux éléments. Il est de la forme $X_i = T_{i+1} - T_i$. Par exemple X_0 sera le temps séparant le début des tirages et la première apparition d'une nouvelle carte. On en conclue donc que pour obtenir le N^{eme} élément, soit T_N, le temps nécessaire sera :

$$T_N = \sum_{i=0}^{N-1} X_i$$

La manière dont sont tirées les cartes implique que X_i suit une loi géométrique, à savoir la répétition d'une épreuve de Bernoulli. Le succès est ici représenté par le tirage d'une nouvelle carte et l'échec par le tirage d'une carte que l'on possède déjà. X_i étant le temps séparant deux éléments, on peut calculer sa probabilité qui est de la forme :

$$P(X_i) = \frac{N-i}{N}$$

X_i suit une loi géométrique de paramètre $\frac{N-i}{N}$. Dans notre cas il serai intéressant de calculer la valeur que l'on s'attend à obtenir si l'on répète infiniment, ou du moins un très grand nombre de fois, l'expérience : cette valeur est l'espèrence. Dans le cas d'une loi géométrique celle-ci est l'inverse du paramètre soit $\frac{N}{N-i}$

On en déduit donc que

$$\sum_{i=0}^{N-1} X_i = \sum_{i=0}^{N-1} \frac{N}{N-i} = N \times \sum_{i=0}^{N-1} \frac{1}{N-i} = N \times \sum_{i=1}^{N} \frac{1}{i}$$

D'après les propriétés de la suite Harmonique, on a : $N \times \sum_{i=1}^{N} \frac{1}{i} \approx N \times ln(N)$

3.2 Mélange d'un jeu de cartes

3.2.1 Modèle mathématique

Intéressons-nous maintenant au problème du battage de cartes qui est très similaire.

De la même manière, on pose N comme étant le nombre d'éléments à mélanger (nombre de cartes à mélanger).

On définit T_i comme étant le temps (nombre de battage(s)) au bout duquel la carte témoin est remontée de i étages par rapport à sa position initiale.

Dans notre cas on recherche donc à calculer T_N. Nous remarquons déjà l'apparition d'une complication. En effet déterminer T_1 est très difficile.

Afin d'y remédier, on va s'intéresser au temps séparant deux deux positions(successives) de la cartes témoin. Il est de la forme $X_i = T_{i+1} - T_i$. Par exemple X_0 sera le temps séparant le début du battage et le premier mouvement de la carte témoin. On en conclue donc que pour mélanger totalement le paquet, soit T_N, le temps nécessaire sera :

$$T_N = \sum_{i=0}^{N-1} X_i$$

On est alors dans le cas d'une loi géométrique, à savoir la répétition d'une épreuve de Bernoulli. Le succès est ici représenté par le changement de position de la carte témoin et l'échec par une carte témoin qui n'a pas changé de position. X_i étant le temps séparant deux éléments, on peut calculer sa probabilité qui est de la forme :

$$P(X_i) = \frac{i+1}{N}$$

X_i suit une loi géométrique de paramètre $\frac{i+1}{N}$. Dans notre cas il serai intéressant de calculer la valeur que l'on s'attend à obtenir si l'on répète infiniment, ou du moins un très grand nombre de fois, l'expérience : cette valeur est l'espèrence. Dans le cas d'une loi géométrique celle-ci est l'inverse du paramètre soit $\frac{N}{i+1}$

On en déduit donc que

$$\sum_{i=0}^{N-1} X_i = \sum_{i=0}^{N-1} \frac{N}{i+1} = N \times \sum_{i=0}^{N-1} \frac{1}{i+1} = N \times \sum_{i=1}^{N} \frac{1}{i}$$

D'après les propriétés de la suite Harmonique, on a : $N \times \sum_{i=1}^{N} \frac{1}{i} \approx N \times ln(N)$

4 Comparaison des modèles déterminés avec les simulations réalisées

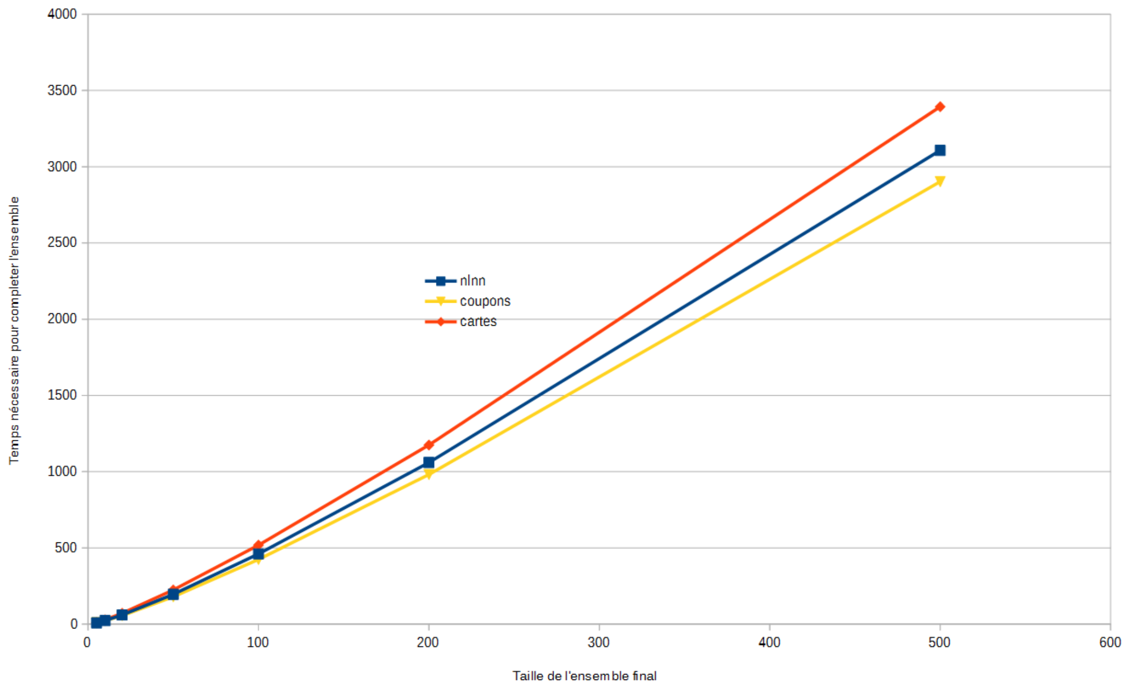

FIGURE 1 – Graphique représentant le temps nécessaire pour compléter un ensemble en fonction de la taille de cet ensemble

Le graphique ci-dessus nous montre les temps nécessaires pour compléter des ensembles de tailles différentes pour le collectionneur de coupons (courbe jaune), pour le battage de cartes (courbe orange) et pour le modèle mathématique (courbe bleue). Les valeurs des simulations proviennent des résultats lors des expériences à 100000 répétitions afin d'avoir la meilleure précision.

Nous pouvons remarquer que dans un cas comme dans l'autre, plus l'ensemble augmente, plus notre modèle mathématique semble s'éloigner de la valeur expérimentale.

Ceci est du à notre approximation avec la suite harmonique. En effet nous avons approché H_n avec $\ln n$. Or ce n'est pas totalement le cas, pour une meilleure approximation nous aurions du poser H_n comme étant $\ln n + c$. A ce moment la nous comprenons bien que lorsque l'on approche $n \times H_n$ on obtient de l'autre coté $n \times c$, ce qui pourrai expliquer cet écart.

Mais ce n'est pas tout, nous savons également que $H_n \underset{n \to +\infty}{\sim} \ln n$ ce qui implique donc que $\lim\limits_{n \to +\infty} H_n - \ln n = 0$ c'est à dire $\lim\limits_{n \to +\infty} \frac{H_n}{\ln n} = 1$. En utilisant cette limite, nous pourrions voir plus clairement que notre modèle mathématique est fiable. En effet la multiplication par n "disparaitrai" puisqu'il sera dans le numérateur et le dénominateur.

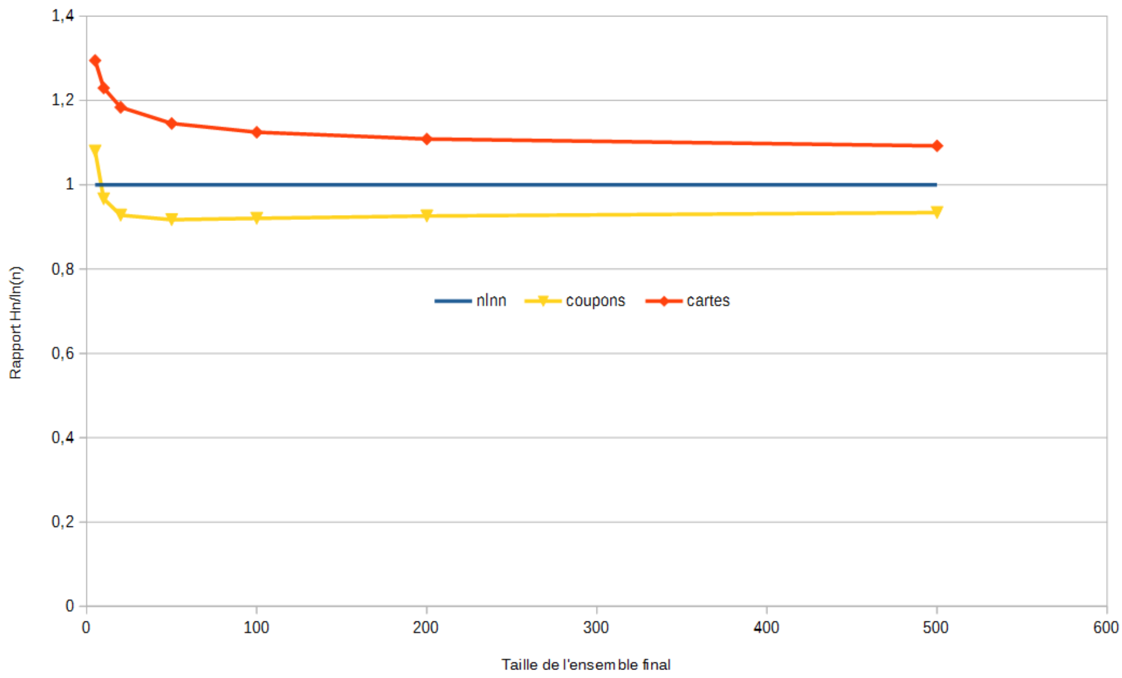

FIGURE 2 – Graphique représentant le rapport $\frac{T_N}{\ln n}$ en fonction de la taille de l'ensemble

En traçant ce graphique, nous remarquons bien que les valeurs semblent tendre vers 1 et donc que l'approximation de H_n par $\ln n$ est de plus en plus "correcte" et que l'écart entre les valeurs du modèle et les expériences se réduisent plus la taille de l'échantillon augmente.

En reprenant une simulation informatique, nous avons obtenu ce graphique suivant :

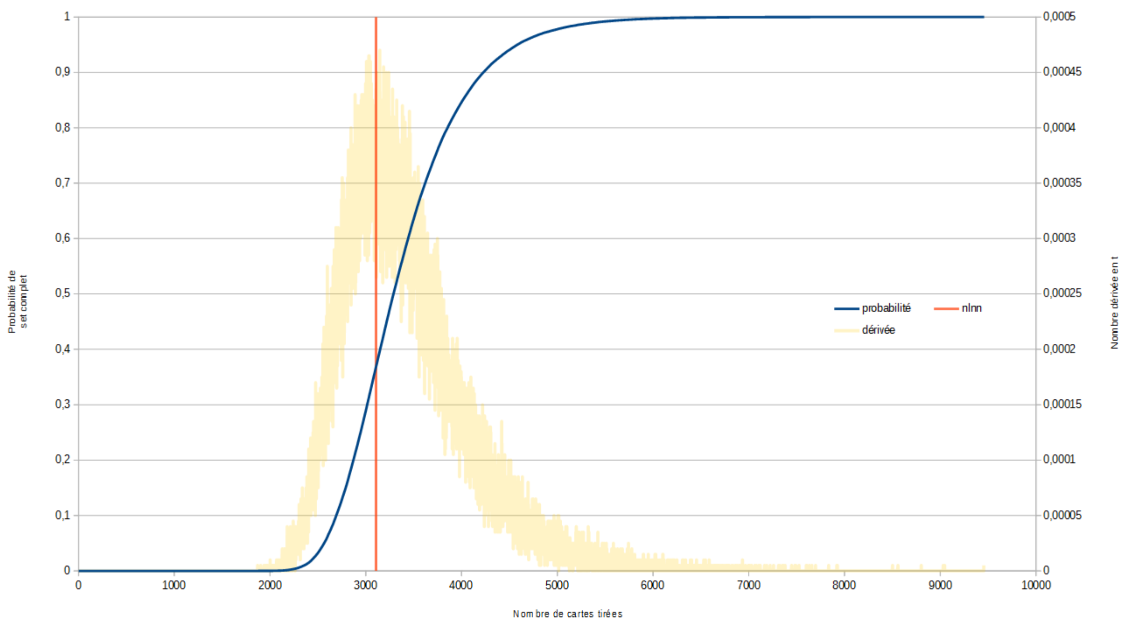

FIGURE 3 – Graphique représentant la probabilité d'obtenir un ensemble complet en fonction du nombre de tirages, réalisés sur une simulation avec un ensemble de taille 500

Pour obtenir ce graphique, nous avons calculé le pourcentage des expériences qui se sont terminées en t tirages ou moins par rapport au nombre total d'expériences avec t étant seuil maximal de tirages (axe des abscisses sur le graphique).

Nous remarquons trois parties majeurs sur le graphique :

— Le début de la courbe où la probabilité est proche de 0. Cela s'explique car nous ne pouvons pas compléter un ensemble en 5 coups si la taille de l'ensemble est 50. De plus il y a une certaine durée (environ quand $0 < t < 2000$ sur le graphique) où il est "impossible" de finir le deck. Finir un ensemble de taille 500 en 500 coups reviendrai à obtenir à chaque fois une nouvelle carte. Autant dire que nous pouvons essayer un bon nombre de fois avant que cela n'arrive.

— La partie "du milieu" est la plus intéressante. En effet, nous voyons que la portion de courbe sur cet intervalle évolue fortement et rapidement. C'est dans cet interval que la majorité des résultats des simulations se trouverons. En effet c'est ici que la probabilité commence à augmenter. Ainsi on a environ 50% de chances de compléter notre deck en environ 3300 tirages, 25% de chances à environ 3000 tirage et 75% de chances à environ 4000 tirages. On voit bien que les valeurs sont concentrées : passer de 25% à 75% de chances de finir le deck s'étale sur une différence de 1000 tirages alors que le graphique complet s'étend sur plus de 9000 tirages ! C'est d'ailleurs dans cette zone que nous retrouvons notre approximation par le modèle mathématique. Proche des 40% et au maximum de la dérivée, on peut estimer que ce modèle se comporte un peu comme la valeur moyenne nécessaire pour compléter notre ensemble.

— La fin de la courbe qui tend vers 1. Effet inverse de la première partie, cette fois-ci, au bout d'un certain temps, on est quasiment certain d'avoir obtenu l'ensemble. Si l'ensemble est de taille 50, il est clair qu'au bout de 500000 tirages il y a très peu de chances que nous n'ayons pas compléter notre deck.

D'après les différentes observations et comparaisons effectuées, nous pouvons supposer que le modèle mathématique est assez fiable par rapport aux différentes expériences menées. Ainsi nous pouvons répondre à la question posée lors des simulations (à savoir s'il est possible de déterminer une valeur dans le cas où l'expérience est répétée un nombre infini de fois) par le fait que si l'on réalise l'expérience infiniment, nous devrions nous rapprocher de plus en plus de la valeur trouvée avec le modèle mathématique. C'est à dire que la limite à l'infini de l'écart entre le modèle mathématique et le résultat de la simulation est 0.

5 Conclusion

Suite aux différents résultats obtenus, aussi bien par le biais de simulations informatiques que par les calculs algébriques, nous avons pu déterminer que les problèmes du type "Collectionneur de coupons", qui paraissent extrêmement aléatoires, suivent en réalité un modèle mathématique qui permet d'avoir une estimation assez précise du temps pour résoudre le problème.

Le modèle mathématique permettant d'obtenir l'estimation en question serait alors $N \ln(N)$ où N représente le nombre d'éléments du set final à obtenir.

Prenons un exemple. Si l'on veut collectionner toutes les cartes des joueurs du FC de Tours, sachant que chaque paquet contient une carte et qu'il y a au total 23 cartes (saison 2011-2012), on peut effectuer le calcul suivant : $23 \ln(23) \approx 72,11$. Si l'on achète une carte par jour, en moyenne il nous faudrait deux mois et onze jours afin de compléter totalement notre collection.

Il pourrait être intéressant de poursuivre les recherches en s'intéressant aux cas où l'on achète un paquet contenant plusieurs cartes et ainsi essayer de déterminer le nombre moyen de paquets qu'il nous faudrait acheter pour obtenir toute la collection.

A Codes source

Simulation du collectionneur de coupons (Python)

Simulation du battage de cartes (Python)

Simulation du collectionneur de coupons avec différentes tailles de paquets (Python, programme non utilisé)

Simulation du collectionneur de coupons (Java, exécution plus rapide, utilisé quand le nombre d'expériences est grand)

Simulation pour les probabilités du collectionneur de coupons (Java)

B Autres annexes

Approximation de la suite harmonique

Fichier Libre Office Calc des différents graphiques

Fichier Libre Office Calc des résultats des probabilités pour la figure 3, page 8