R2.15 – Initiation à l'analyse de données

BUT Informatique

UPEC UNIVERSITÉ PARIS-EST CRÉTEIL | iut CRÉTEIL-VITRY

ACTIVITE #1 – LE PROBLEME DES TANKS

Le document présente une activité d'initiation à l'analyse de données dans le cadre du BUT Informatique, centrée sur l'estimation de la taille d'une population à partir d'un échantillon aléatoire. Il s'inspire du "problème des tanks" de la Seconde Guerre mondiale, où les Alliés utilisaient des numéros de série pour estimer la production militaire allemande.

Trois situations sont envisagées pour l'estimation de cette taille. Pour chaque situation, différents estimateurs sont proposés, basés sur la moyenne, la médiane, les valeurs extrêmes, et le nombre d'observations. Le document détaille les formules mathématiques et les étapes de mise en œuvre, incluant la simulation de données et l'application sur des données réelles.

Deux problèmes pratiques sont abordés :

- Estimation du nombre d'étudiants dans une promotion à partir des numéros d'ordinateurs portables.
- Estimation du nombre de Teslas Modèle 3 produites en 2018 à partir des numéros VINs.

Des exemples de code Python sont fournis pour illustrer la lecture des données, le calcul des statistiques, et l'application des estimateurs.

Le problème des tanks

Durant la Seconde Guerre mondiale, les Alliés avaient un besoin criant d'estimer avec précision la quantité de matériel militaire que l'Allemagne nazie produisait. Les estimations provenant des services de renseignements habituels étaient contradictoires et incertaines.

Au début de l'année 1943, la Economic Warfare Division de l'ambassade américaine à Londres commença à analyser divers marquages obtenus à partir d'équipements allemands capturés ou détruits sur le front. Plus particulièrement, les numéros de série pourraient être utilisés pour estimer la force de production de la machine de guerre allemande.

✶ Les statistiques plus fortes que les services de renseignement

Les gouvernements britanniques et américains se tournèrent donc vers des statisticiens pour savoir si leurs estimations pouvaient être améliorées.

Mois	Estimation statistique	Estimation par les services de renseignements	Selon les archives allemandes
Juin 1940	169	1 000	122
Juin 1941	244	1 550	271
Août 1942	327	1 550	342

Comparaison des estimations produites par les estimateurs que nous présenterons, des estimations produites par les services de renseignements ainsi que de la production exacte selon les archives allemandes.

Les formules développées dans cette activité ont aussi été utilisées dans des contextes non militaires. Par exemple, elles ont été utilisées pour estimer le nombre d'ordinateurs Commodore 64 produits. De la même façon, on a estimé à partir des codes IMEI (International Mobile Equipment Identity) de plusieurs utilisateurs avoir vendu 9 190 680 exemplaires du premier modèle d'iPhone commercialisé en 2007 aux États-Unis.

✶ Les trois situations possibles

Dans cette activité, nous nous intéresserons donc à l'estimation de la taille N d'une population à partir d'un échantillon aléatoire prélevé dans celle-ci, dans le cas où les items sont numérotés de façon séquentielle.

Supposons que nous avons une population de N objets numérotés de la façon suivante :

$$s+1,\ \ s+2,\ \ s+3,\ \ \cdots\ ,\ \ s+N$$

Trois situations distinctes peuvent se produire :

1. s est connu et égal à 0 et N est inconnu.
2. s est connu mais différent de 0 et N est inconnu.
3. s est inconnu et N est inconnu.

Dans tous les cas, N est l'inconnue qu'il nous faudra estimer.

Les situations 1 et 2 peuvent être rassemblées en une seule situation pour laquelle : s est connu et égal ou supérieur à 0 et N est inconnu.

✶ Les mathématiques

Pour calculer les diverses mesures statistiques dont nous aurons besoin pour l'estimation de N, nous allons classer les unités statistiques de notre échantillon en ordre croissant. Nous avons :

$$X_{(1)} < X_{(2)} < X_{(3)} < \cdots < X_{(n-1)} < X_{(n)}$$

où les valeurs $X_{(1)}$, $X_{(2)}$,$\cdots$, $X_{(n)}$ sont les valeurs ordonnées des observations de l'échantillon. En particulier, $X_{(1)}$ est la plus petite valeur observée de l'échantillon et $X_{(n)}$ est la plus grande.

La moyenne, notée $\bar{X}$,des valeurs de l'échantillon est donnée par :

$$\bar{X} = \frac{X_{(1)} + X_{(2)} + X_{(3)} + \cdots + X_{(n-1)} + X_{(n)}}{n}$$

La médiane, notée $\tilde{X}$, des valeurs de l'échantillon est donnée par :

$$\tilde{X} = \begin{cases} \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{si } n \text{ est pair} \\ X_{\left(\frac{n}{2}+\frac{1}{2}\right)} & \text{si } n \text{ est impair} \end{cases}$$

✶ Les estimateurs

Situation 1 : s est connu, avec $s = 0$, et N est inconnu

- **Estimateur #1** – cet estimateur est basé sur la moyenne des valeurs observées :

$$N_1 = 2\bar{X} - 1$$

- **Estimateur #2** – cet estimateur est basé sur la médiane des valeurs observées :

$$N_1 = 2\tilde{X} - 1$$

- **Estimateur #3** – cet estimateur est basé sur les valeurs extrêmes observées :

$$N_3 = X_{(1)} + X_{(n)} - 1$$

- **Estimateur #4** – cet estimateur est basé sur la valeur maximale observée et le nombre d'observations :

$$N_4 = \frac{n+1}{n} X_{(n)} - 1$$

Situation 2 : s est connu, avec $s > 0$, et N est inconnu

Nous pouvons résoudre ce problème en soustrayant la valeur s aux valeurs des observations avant d'utiliser les estimateurs précédents.

Situation 3 : s est inconnu et N est inconnu

- **Estimateur #5** – cet estimateur est basé sur l'étendue des valeurs et le nombre d'observations :

$$N_5 = \frac{n+1}{n-1}\left(X_{(n)} - X_{(1)}\right) - 1$$

Mise en œuvre

Etape #1 – Code source utile

Les fonctions suivantes sont, en partie, fournies. Vous devez les comprendre avant de les utiliser :

- Une fonction `stats(liste)` qui retourne des mesures statistiques calculées sur l'échantillon. Le code source de cette fonction est donné. La variable `liste` contient les valeurs observées non-forcément triées. Les mesures statistiques retournées sont, dans l'ordre, $X_{(1)}$, $X_{(n)}$, $\bar{X}$ et $\tilde{X}$.
- Une fonction par estimateur de N :
  - `estimateur_1(xmoy)` : premier estimateur ;
  - `estimateur_2(xmed)` : deuxième estimateur ;
  - `estimateur_3(xmin, xmax)` : troisième estimateur ;
  - `estimateur_4(xmax, n)` : quatrième estimateur ;
  - `estimateur_5(xmin, xmax, n)` : cinquième estimateur.

  Ces fonctions sont à coder. Les arguments utilisés sont respectivement `xmoy`$= \bar{X}$, `xmed`$= \tilde{X}$, `xmin`$= X_{(1)}$, `xmax`$= X_{(n)}$ et `n` le nombre d'observations dans l'échantillon. Chaque fonction doit renvoyer la valeur estimée de N.

Etape #2 – Vérifications sur des données simulées

Dans cette partie, vous allez tester les différents estimateurs sur des données d'observations simulées par un processus de tirage aléatoire. Nous nous placerons dans chacune des situations décrites précédemment.

- Cas 1 et 2 : s est connu et $s \geq 0$, N est inconnu ;
- Cas 3 : s est inconnu, N est inconnu ;

La procédure à suivre est la suivante :

1. Choisir des valeurs pour s, N et n ;
2. Produire une population de numéro $1+s$, $2+s$, $\cdots$, $N+s$;
3. Extraire un échantillon par un tirage aléatoire de n observations de cette population ;
4. Appliquer sur cet échantillon les 5 estimateurs.

Vous pourrez traiter les cas et situations donnés dans les deux tableaux suivants :

	$N = 500$		$N = 10000$			
	$n = 20$	$n = 100$	$n = 20$	$n = 100$	$n = 500$	$n = 1000$
N_1						
N_2						
N_3						
N_4						
N_5						

Estimations de N en fonction du nombre d'observations avec ici $s = 0$.

	$N = 500$		$N = 10000$			
	$n = 20$	$n = 100$	$n = 20$	$n = 100$	$n = 500$	$n = 1000$
N_1						
N_2						
N_3						
N_4						
N_5						

Estimations de N en fonction du nombre d'observations avec ici $s = 2000$.

Que pouvez-vous conclure d'après les résultats obtenus en fonction des différentes situations rencontrées ?

Etape #3 – Application sur des données réelles

Pour chacun des 2 problèmes posés ci-dessous, identifiez la situation dans laquelle l'estimation doit s'appliquer. Que pouvez-vous conclure d'après les résultats obtenus ? En particulier, que représente réellement l'estimée de N pour chaque problème ?

- Problème #1 : Estimation du nombre d'étudiants dans une promotion, ou un groupe, à partir d'un échantillon de cette population.

  > Ici, l'observation porte sur les numéros attribués aux ordinateurs portables mis à disposition de chaque étudiant. Ces ordinateurs sont numérotés de manière incrémentale, à partir de 1, et distribués dans les différents groupes de TD (environ 30 étudiants par groupe).
  >
  > On suppose que, pour chaque groupe de TD, on dispose d'une fraction (échantillon) des numéros des ordinateurs (observations). A partir de ces observations, nous allons produire les différentes estimations du nombre total N d'étudiants.

Mise en œuvre – Afin d'estimer la taille du groupe, la mise en œuvre est la suivante :

1. Votre groupe de TD est partagé en 2 sous-groupes de TP (A1 et A2, ou B1 et B2, ou C1 et C2) pour obtenir 2 échantillons que l'on nommera SG1 et SG2.
2. Dans chaque sous-groupe, relever les numéros des ordinateurs.
3. Utiliser cette liste de numéros afin d'estimer la valeur de N à l'aide des différents estimateurs.
4. Comparer les estimations obtenues de N d'un sous-groupe à l'autre.
5. Recommencer l'opération à partir de l'étape 1 en subdivisant, cette fois, chaque sous-groupe de TP en 2 sous-sous-groupes (1/2 TP) pour obtenir 4 échantillons que l'on nommera SSG11, SSG12, SSG21 et SSG22.

Vous pourrez reporter les différentes estimations dans le tableau donné ci-dessous :

	SG1	SG2	SSG11	SSG12	SSG21	SSG22
N_1						
N_2						
N_3						
N_4						
N_5						

Estimations de la population d'étudiants dans un groupe.

Conclure…

- Problème #2 : Estimation du nombre de Tesla produites

> La Tesla Modèle 3 est une berline familiale haut de gamme et 100 % électrique, construite par la société Tesla. Présentée au public le 31 mars 2016, les 30 premières livraisons ont eu lieu le 28 juillet 2017 aux États-Unis. Il s'agit du quatrième modèle de voiture commercialisé par Tesla, après la Tesla Modèle X.
>
> En raison d'un rythme de production déficient, plusieurs acheteurs mécontents attendent leur modèle 3 avec impatience. Pour estimer le nombre de voitures qui sortent des usines de production, certains acheteurs ont décidé de partager leurs « Vehicle Identification Numbers » (VINs), des codes numériques uniques qui sont attribués à tous les nouveaux véhicules vendus aux États-Unis. La société ne communique pas sur le nombre de véhicules produits.
>
> Ces VINs sont une séquence d'entiers débutant à 1 et augmentant à chaque nouvelle voiture produite. Nous allons exploiter ces VINs afin d'estimer la production de Tesla Modèle 3.

Une petite recherche des chiffres de production en 2018 de la Tesla 3 aboutie à des chiffres quelque peu différents si l'on compare les différentes sources :

Sources	2017	2018
Tesla Wikipedia (DE)	-	145 864
Tesla Wikipedia (FR)	1 764	146 055
Tesla Wikipedia (EN)	1 764	145 846

Mise en œuvre : Les données réelles utiles sont disponibles sur Eprel (fichier `tesla2018.csv`). Le fichier contenant les numéros VINs est un fichier au format CSV (consultable à l'aide de l'application Excel). La structure de données est composée de 3 colonnes : ID = identifiant du propriétaire de la Tesla, DATE = date d'assignation du VINs au format `jj/mm/aaaa`, VIN = numéro Vins.

Nous allons utiliser les VINs afin d'obtenir une estimation du nombre N de Teslas produites au cours de l'année 2018. La procédure à suivre est la suivante :

1. Lire le fichier CSV en Python à l'aide de Pandas ;
2. Pour chaque mois de production (de janvier à décembre) faire :
   a. Sélectionner la liste des VINs du mois courant ;
   b. Pour cette liste de VINs, appliquer les 5 estimateurs afin de mesurer le nombre de voitures produites au cours du mois ;
3. Estimer la production totale pour l'année 2018.
4. Supplément : Estimer la production par trimestre de l'année.

Un extrait du code Python permettant la lecture et la sélection des VINs d'un mois de l'année 2018 depuis le fichier CSV est donné.

Le code illustre également l'affichage graphique, sous forme d'un nuage de points, des VINs en fonction de la date d'assignation de ces VINs.

Vous pourrez compléter le tableau suivant :

	N_1	N_2	N_3	N_4	N_5
Janvier					
Février					
Mars					
Trimestre 1					
Avril					
Mai					
Juin					
Trimestre 2					
Juillet					
Août					
Septembre					
Trimestre 3					
Octobre					
Novembre					
Décembre					
Trimestre 4					
Total 2018					

Estimations de la production de Teslas Modèle 3 au cours de l'année 2018.

Conclure...

Suppléments – Vous pourrez également estimer la production pour l'année 2019 (fichier `tesla2019.csv`).

Code source pour l'étape #1 : fonction utile

```python
def stats(liste):
    # Tri de la liste pour le calcul de la médiane
    liste_triee = sorted(liste)

    # Taille de la liste (nombre d'observations)
    n = len(liste_triee)

    # Calcul des valeurs demandées
    minimum = liste_triee[0]
    maximum = liste_triee[n-1]
    moyenne = int(sum(liste_triee) / n)

    # Calcul de la médiane (attention l'indice démarre à 0 et non pas à 1)
    if n % 2 == 1:
        # Si la taille est impaire, la médiane est l'élément du milieu
        mediane = liste_triee[n // 2]
    else:
        # Si la taille est paire, la médiane est la moyenne des deux éléments du milieu
        mediane = int((liste_triee[n // 2 - 1] + liste_triee[n // 2]) / 2)

    # Retour des 4 mesures sous forme de valeurs séparées
    return minimum, maximum, moyenne, mediane


# On teste la fonction sur un jeux de données quelconques
valeurs = [5, 3, 100, 8, 1, 7, 20, 33, 4]
minimum, maximum, moyenne, mediane = stats(valeurs)

# Affichage des résultats
print(f"Minimum : {minimum}")
print(f"Maximum : {maximum}")
print(f"Moyenne : {moyenne}")
print(f"Médiane : {mediane}")
```

Code source pour l'étape #2 : vérification sur des données simulées

```
import random

# Population totale (inconnue)
N = 100000

# Décalage 1+s,2+s,...,N+s
s = 10

# Nombre d'observations (échantillon)
n = 200

######################################################################
# Génération d'une population et tirage aléatoire d'un échantillon
######################################################################

# Population complète
population = list(range(s+1,s+N+1))

# Echantillon tirer aléatoirement, n valeurs sans remise
echantillon = random.sample(population, n)

print(echantillon)
```

Code source pour l'étape #3 : application sur des données réelles

```
import pandas as pd
import matplotlib.pyplot as plt

# Charger les données au format CSV dans un DataFrame Pandas
df = pd.read_csv('tesla2018.csv', sep=';')

# Convertir la colonne DATE en format datetime (aaaa-mm-jj)
df['DATE'] = pd.to_datetime(df['DATE'], dayfirst=True)

# Créer le nuage de points avec la date en abscisse et la valeur du VIN en ordonnée
plt.figure(figsize=(7,6))
plt.scatter(df['DATE'], df['VIN'], alpha=0.25)

# Ajouter des labels et un titre
plt.title('VINs pour l\'année 2018')
plt.xlabel('Date')
plt.ylabel('VIN')

# Afficher le graphique
plt.show()

# Extraire le mois de la date
df['MOIS'] = df['DATE'].dt.month

# Filtrer les VINS en spécifiant le mois retenu, par exemple 6 (juin)
mois_cible = 6
vins = df[df['MOIS'] == mois_cible]['VIN'].to_list()
print(vins)
```