L2 MASS Université Lyon 1, Probabilités-statistique 2
Année Universitaire 2020 – 2021

TD 1 : Introduction à la statistique

Ce TD a pour objectif :

1. de faire une introduction à la statistique descriptive
2. d'introduire par l'exemple les définitions des quantités empiriques (moyenne empririque, variance empirique, quantile empirique...)
3. de vous aider à différencier quantité empirique associée à un échantillon, et quantité théorique associée à une variable aléatoire.

On rappelle aussi quelques définitions élémentaires :

- Population : Ensemble étudié (ex : La population de la région Rhône-Alpes)
- Individu : Un élément de la population (ex : Vous)
- Échantillon : Un sous-ensemble de la population (ex : 500 individus intérrogés)
- Caractère, variable : Une caractéristique mesurable chez chaque individu (ex : Niveau de revenu caractère quantitatif, pour ou contre... caractère qualitatif)

Exercice 1 ***Calcul de quantité empiriques***
Voici les températures moyennes observées, pour certains pays d'Europe (dans les capitales), en 2018[1].

Pays	*Temperature*
Albanie	*15.2*
Belgique	*10.5*
Croatie	*10.7*
Danemark	*9.1*
Finlande	*5.9*
France	*12.3*
Allemagne	*10.3*
Islande	*4.3*
Italie	*15.2*
Norvege	*5.7*
Portugal	*17.5*
Espagne	*15.0*
Suede	*6.6*
Royaume-Uni	*9.3*

1. *Quelle est la population ? Quels sont les individus ? Quel est la variable étudiée ?*
2. *Calculer la moyenne empirique* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *des températures.*
3. *Calculer la variance empirique* $\bar{s}_b^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$*, et l'écart type empirique* s_b *des températures.*
4. *On définit ici un quantile empirique d'ordre* p *comme le nombre* $\bar{q}_p$ *donné par* $\bar{q}_p = x_{j+1} + r(x_{j+2} - x_{j+1})$*, où* j *est la partie entière de* $(n-1)p$ *et* r *est la partie fractionnelle, et* x_j *désigne la* j*-ieme plus petite donnée. Cette définition est en partie une convention, et des définitions alternatives sont possibles, du moment que le quantile d'ordre* p *correspond à un nombre tel qu'une proportion* p *de l'échantillon se trouve au dessous de ce nombre. La médiane correspond au quantile d'ordre* $1/2$*, et les premiers et troisièmes quartiles aux quantiles d'ordres respectifs* $1/4, 3/4$ *(idem pour les déciles, centiles...). Donner la médiane, le premier et troisième quartile des températures.*
5. *Pour dessiner une "boîte à moustaches" (boxplot), une convention usuelle consiste à représenter la boîte entre les* 1*-er et* 3*-eme quartiles, ajouter une barre pour la médiane, illustrer les bornes correspondant à* $x_{min} = \max(\min_i\{x_i\}, \bar{q}_{1/4} - 1.5e)$ *et* $x_{max} = \min(\max_i\{x_i\}, \bar{q}_{3/4} + 1.5e)$*, où* $e = \bar{q}_{3/4} - \bar{q}_{1/4}$ *désigne l'écart interquartiles. On peut aussi ajouter les données plus "extrèmes" avec seulement un point. Représenter la boîte à moustache des températures.*

[1]Source : `https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature#Europe`

6. *On définit la fonction de répartition empirique de l'échantillon x au point t par*

$$\bar{F}_x(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \leq t}, \text{ où } \mathbb{1}_{bool} \text{ vaut 1 si bool est vrai et 0 sinon.}$$

Cela correspond à la proportion des valeurs de l'échantillon plus petites où égales à t. Calculer la fonction de répartition empirique au point $13°C$, $\bar{F}(13)$.

7. *Représenter la fonction de répartition empirique.*

8. *Un histogramme est une représentation des données, où, après les avoir regroupées par classes de valeurs (par exemple par tranche de $1°C$ ou $2°C$ ici), on représente pour chaque classe une barre dont la hauteur correspond au nombre de données dans cette classe. Représenter un histogramme des données (largeur des barres : $2°C$).*

Exercice 2 ***Variable aléatoire vs réalisations***
On considère un dé équilibré à 6 faces, et des résultats de 12 lancers de ce dé donnés par le tableau suivant :

Valeur	*1*	*2*	*3*	*4*	*5*	*6*
Nombre	*0*	*3*	*1*	*4*	*3*	*1*

1. *Quelle est la loi de la variable aléatoire X correspondant au lancer d'un dé équilibré ? Proposer une construction explicite d'un espace de probabilité et de X sur cet espace.*

2. *Calculer $\mathbb{E}[X]$, $Var(X)$, F_X, et la quantile théorique d'ordre $\frac{1}{4}$ (on rappelle que le quantile théorique d'ordre p est défini par $q_p = \inf\{x, p \leq F(x)\}$.)*

3. *On note $x_1, \cdots x_{12}$ le résultat des 12 lancers donnés par le tableau précédent. Donner la fréquence de chaque résultat, puis $\bar{x}_{12}$, $\bar{s}_b^2$, $\bar{F}$, et le quantile empirique d'ordre $\frac{1}{4}$.*

4. *On considère maintenant $\bar{X}_{12}$ comme une variable aléatoire. Expliquer ce qu'elle représente, proposer un espace de probabilité sur lequel la construire, donner l'ensemble des valeurs qu'elle peut prendre, et calculer $\mathbb{P}(\bar{X}_{12} = 1)$, et $\mathbb{P}(\bar{X}_{12} = \frac{13}{12})$, et $\mathbb{P}(\bar{X}_{12} = 1.25)$ pour les plus courageu-x-ses.*

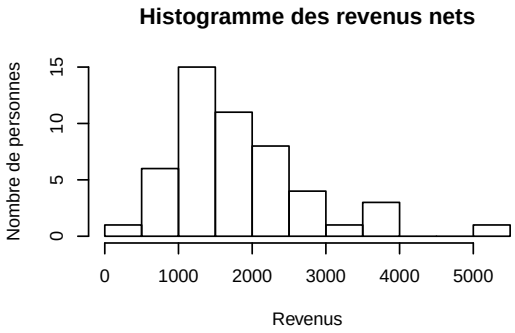
5. *Que vaut $\mathbb{E}[\bar{X}_{12}]$?*

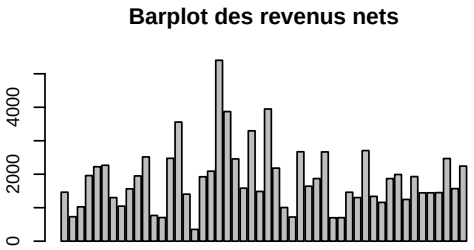
Exercice 3 ***Qualitatif ou quantitatif***
Pour chacune des variables suivante, dire si elles sont qualitative (catégorielle) ou quantitative et discrètes ou continues

- *La race d'un chien*
- *Le salaire d'un individu*
- *La marque d'un téléphone portable*
- *L'âge d'un individu*
- *Le nombre de députés d'un pays*
- *La température d'un réacteur nucléaire*
- *Le nombre de frères et soeurs d'un individu*

Exercice 4 ***Comment représenter et interpréter des données ?***
On a sondé aléatoirement 50 foyers pour lesquels on a estimé le revenu moyen du foyer par adulte actif, et le nombre moyen d'années d'études (après la scolarité obligatoire) des enfants. Ces résultats ont été résumés dans les graphes suivants.

- *Dire quelles représentations il est pertinent ou non de fournir.*
- *Expliquer ce que représentent les différents graphiques.*
- *Lister les propriétés qui semblent apparaître sur les graphiques. Peut-on tirer des conclusions sur les les causes des différences entre le nombre d'années d'études des enfants de l'étude ?*


Histogramme des revenus nets
Nombre de personnes
0
5
10
15
0
1000
2000
3000
4000
5000
Revenus


Barplot des revenus nets
0
2000
4000

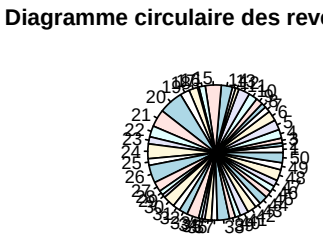
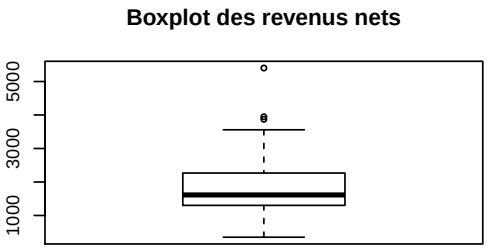
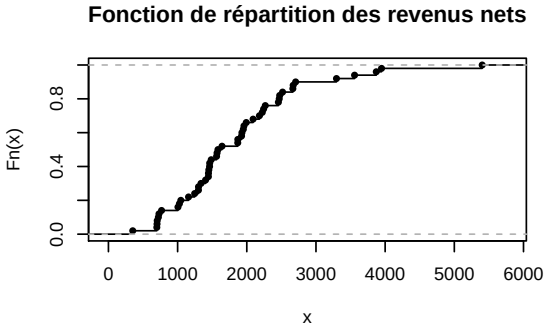
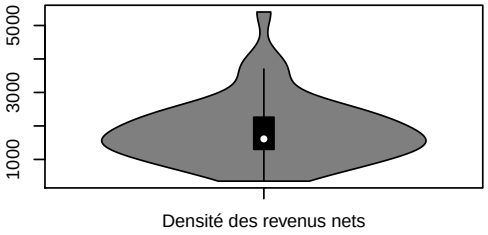
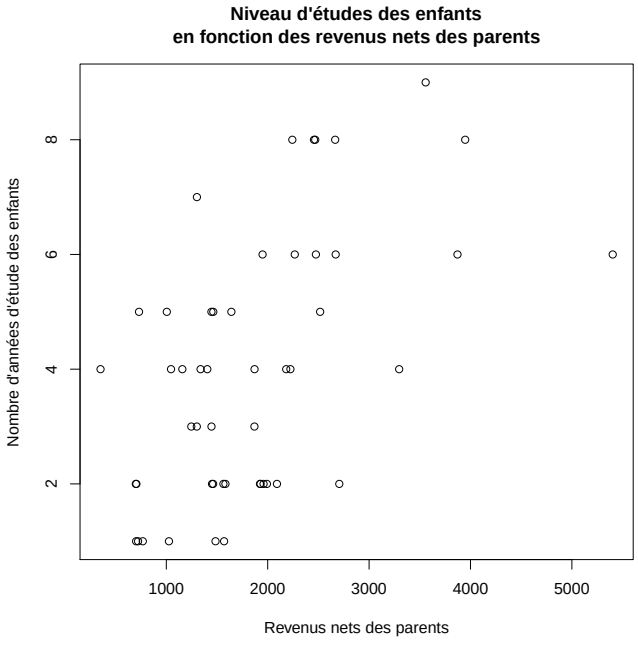


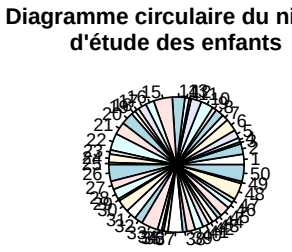
Diagramme circulaire des revenus nets

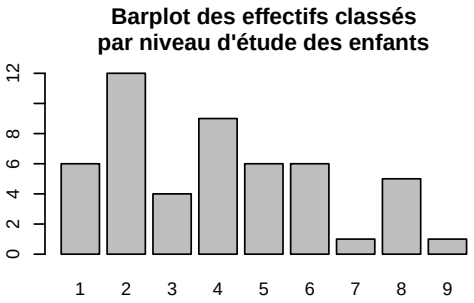

Boxplot des revenus nets
1000
3000
5000

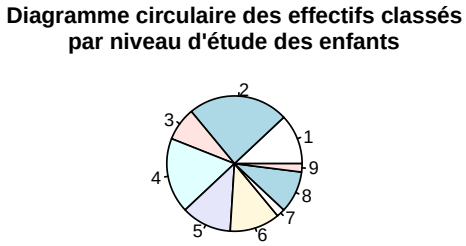

Fonction de répartition des revenus nets
Fn(x)
0.0
0.4
0.8
0
1000
2000
3000
4000
5000
6000
x

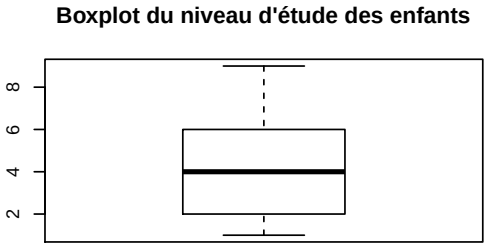

1000
3000
5000
Densité des revenus nets

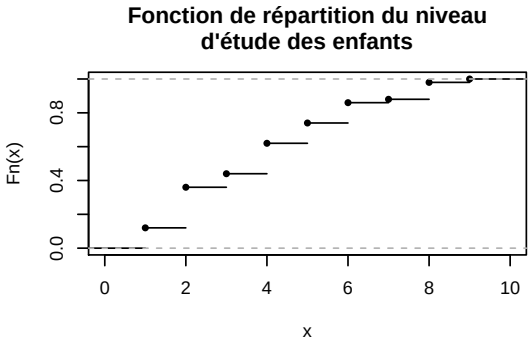

Niveau d'études des enfants
en fonction des revenus nets des parents
Nombre d'années d'étude des enfants
2
4
6
8
1000
2000
3000
4000
5000
Revenus nets des parents

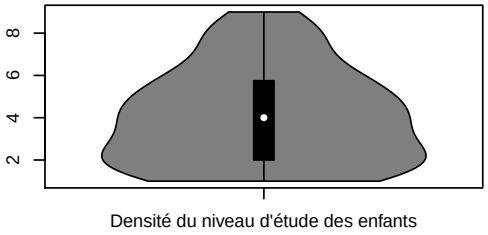












Exercice 5 ***Paradoxe de Simpson***

On s'interesse à l'efficacité de 2 *traitements pour les calculs rénaux. Le nombre de patients guéris et traités pour chacun des traitements A et B est indiqué par le tableau suivant, détaillé en fonction de la taille des calculs.*

Taille des calculs \ *Traitement*	*Traitement A*	*Traitement B*
Petits calculs	93%(81/87)	87%(234/270)
Grands calculs	73%(192/263)	69%(55/80)
Total	78%(273/350)	83%(289/350)

1. *Quel traitement est-le meilleur à votre avis. Pourquoi parle-t-on d'un paradoxe ?*
2. *Proposer une explication à ce phénomène.*
3. *(Question de culture scientifique) Comment peut-on éviter cette situation avec une méthode usuellement utilisée pour les statistiques ?*

Exercice 6 ***Problème de protocole***
Lister les problèmes de protocole que vous voyez dans l'expérience suivante :
On veut montrer que une personne sur deux qui occupe Lyon n'y habite pas. Pour cela, on va interoger des personnes devant la gare de la PartDieu. La première expérience n'a pas été concluante (sur 80 *personnes intérogées, seulement* 30 *n'habitaient pas à Lyon). On réitère donc l'expérience une nouvelle fois, en choisissant les personnes intérogées "au feeling", et on obtient dans l'ordre (O : habite Lyon, N : n'habite pas Lyon) :* $O/O/N/O/N/O/N/N/N$*. On s'arête après ce* 5 *non Lyonnais sur* 9 *personnes intérogées : c'est prouvé, plus de la moitié les Lyonnais n'habitent pas à Lyon !*

Exercice 7 ***Dessiner une boîte à moustaches***
On donne le tableau suivant [2]
Distribution des salaires mensuels nets en EQTP dans le secteur privé selon le sexe et la catégorie socioprofessionnelle en 2016***.***

Catégorie	*Ensemble*	*Cadres*	*Employés*
1*eme décile*	1190	1970	1130
1*er quartile*	1410	2520	1270
Médiane	1790	3270	1470
3*eme quartile*	2460	4470	1780
9*eme décile*	3580	6430	2220

Dessiner la boite à moustache correspondant à chaque catégorie (on prendra arbitrairement $x_{\min}$ *et* $x_{\max}$ *égal au* 1*er et* 9*eme décile).*

Exercice 8 ***Vers un projet***

1. *Trouver une ou plusieurs questions qui vous intéresseraient pour lesquelles une démarche statistique pourrait contribuer à trouver une réponse.*
2. *Restreindre les questions à des exemples pour lesquelles vous pourriez trouver des données disponibles et/ou organiser un "sondage".*
3. *Expliciter les éventuels biais méthodologiques.*

Exercice 9 ***À quoi sert la statistique***
Auxquelles de ces questions pensez-vous qu'un protocole statistique adapté pourrait apporter une réponse (avec une certitude de 99%*) ? Dans le cas négatif, préciser la raison.*

- *Le vaccin A a un taux de protection plus haut que le vaccin B ?*
- *Y a-t-il exactement* 18054 *étudiant-e-s de Lyon* 1 *qui travaillent en parallèle des études ?*
- *Le risque d'accident nucléaire majeur en France chaque année est-il inférieur à* 1 *chance sur* 10000 *?*
- *La télékinésie existe-t-elle ?*
- *Le café est-il meilleur que le thé ?*
- *Le salaire moyen chez les femmes est-il inférieur au salaire moyen chez les hommes ?*

Exercice 10 ***Les données à problèmes***
À partir des données présentes dans le tableau traitant de vaccination contre la rage (cf poly de cours), lister les problèmes rencontrés sur les données, et les questions que vous poseriez à la personne ayant produit ces données, et vous demandant une analyse.

Exercice 11 ***Du travail de groupe***

1. *À partir des données présentes dans le tableau traitant de vaccination contre la rage (cf poly de cours), en vous répartissant les rôles par groupes, dessiner (vous choisirez : vaccin -* 1*/*2*/ensemble, date -* 2 *semaines/*1 *mois/ moyenne)*
    - *Un histogramme des données*

[2] source: Insee, 2016 `https://www.insee.fr/fr/statistiques/2407703#tableau-figure2`

- *Un boxplot des données*
- *La fonction de répartition des données*

2. *Si vous disposez d'un ordinateur, essayer d'utiliser le logiciel de votre choix pour faire le même type de représentations.*

Exercice 12 ***Un problème ouvert*** *En tant que statisticien, votre rôle pourra souvent dépasser le seul cadre du traitement mathématique des données, en particulier, il se peut qu'on vous demande de participer*

- *À la collecte de données*
- *Au "nettoyage des données"*
- *À l'analyse descriptive des données (représentation graphique, résumé...)*
- *...*

L'objectif de cet exercice est de vous confronter à ces différents problèmes. De plus en plus de données sont accessibles au public (en particulier, certaines issues d'organismes publics).

1. *En visitant le site* `https://www.ecad.eu/`*, trouver les données quotidiennes de température sur Lyon (ou la ville de votre choix).*
2. *Charger ces données, avec le logiciel de votre choix*
3. *Observer les données, et fournir* 3 *représentations graphiques de votre choix que vous trouverez pertinentes.*
4. *Proposer des façons de résumer les données (indices caractéristiques que vous trouvez pertinents).*

Exercice 13 *Une étude médicale a récolté les données de plusieurs patients. Proposer une loi de probabilité pouvant modéliser ces variables :*

(a) Genre (femme, homme) *(b) Age (années)* *(c) Fumeur (oui ou non)*

(d) Pression (millimétre de mercure) *(e) Niveau de calcium dans le sang (microgr. par ml)*

Exercice 14 *Nous présentons les pourcentages de femmes parmi les docteurs de 1997-1998 aux Etats-Unis : psychologie, 67.5 %; pédagogie, 63.2 %; sciences de la vie, 42.5 %; affaire, 31.4 %; sciences physiques, 25.2 %; ingénierie, 12.2 %.*

1. *Expliquer pourquoi nous ne pouvons pas utiliser un camembert pour présenter ces données.*

2. *Faire un graphique en barres des données.*