UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

DATE :	**……21/12/2020-5/1/2021……**
EPREUVE:	**Statistiques et informatique appliquées aux Sciences sociales**
ENSEIGNANT:	**…Renaud Orain <renaud.orain@univ-paris1.fr> …….**
DIPLOME :	**Licence de Sciences sociales, 3e année** **Étudiant·es de tous parcours, inscrit·es en contrôle continu…**
DUREE conseillée:	**…2. h**
DOCUMENT AUTORISE	**Devoir à la maison (tous documents autorisés)**

Sujet obligatoire pour les étudiant·es inscrit·es en contrôle terminal

Ce devoir devait être rendu sous forme manuscrite, en main propre ou en version scannée ou photographiée sur le devoir.
Il était demandé de traiter les questions qui suivent, en faisant les calculs « à la main » (calculatrices autorisées, y compris sur ordinateur) en indiquant si possible le temps nécessaire pour réaliser le devoir. Cette mention n'était pas prise en compte dans l'évaluation mais visait à mieux calibrer les sujets futurs.

Distribution d'âge et sorties culturelles

Les étudiants en Master Démographie en 2019-2021 ont construit un questionnaire portant sur les sorties et pratiques culturelles des jeunes mais ce questionnaire est accessible en ligne à quiconque.

On se demande d'abord comment varie le nombre de sorties en 2019 en fonction de l'âge en milieu d'année 2019. On se restreint aux personnes ayant répondu sur leur année de naissance et on évalue les corrélations de l'âge avec un décompte de différent type de sortie faites dans l'année.

Coefficients de corrélation de Pearson, N = 253 **Proba > |r| sous H0: Rho=0**							Mo-yennes	**Ec-type**	Coefficients de variation
Âge **Nb sorties**	**Age mi 2019**	**1:cinéma**	**2 :Concert**	**3 :Musee**	**4 :Monuments**	**5 : Cirque, théâtre**			
Age mi 2019	1,00000	-0,03598 0,5689	0,12422 0,0484	-0,06295 0,3186	-0,02928 0,6430	0,15574 0,0131	24,19	5,21	0,22
1:cinéma	-0,03598 0,5689	1,00000	0,21627 0,0005	0,31075 <,0001	0,23937 0,0001	0,29990 <,0001	7,15	6,60	0,92
2 :Concert	0,12422 0,0484	0,21627 0,0005	1,00000	0,15306 0,0148	0,18197 0,0037	0,33887 <,0001	3,40	4,96	1,46
3 :Musee	-0,06295 0,3186	0,31075 <,0001	0,15306 0,0148	1,00000	0,60333 <,0001	0,42451 <,0001	6,80	7,03	1,03
4 :Monuments	-0,02928 0,6430	0,23937 0,0001	0,18197 0,0037	0,60333 <,0001	1,00000	0,29076 <,0001	5,49	5,76	1,05
5 : Cirque, théâtre	0,15574 0,0131	0,29990 <,0001	0,33887 <,0001	0,42451 <,0001	0,29076 <,0001	1,00000	2,33	3,60	1,54

1. À l'aide des sommes et sommes de produits ci-dessous, expliquer les étapes du calcul de la moyenne et de la variance de l'âge moyen en 2019, puis de la corrélation entre l'âge et le nombre de sorties au cirque ou au théâtre dans l'année. **3 pts**

Somme des produits	Âges mi 2019	Cirque, théâtre	Somme	Effectif
Âges mi 2019	154 924,2	15 010,8	6120,25	253
Cirque, théâtre	15 010,8	4 646,0	590	253

La somme des âges moyens étant donnée dans l'énoncé, le plus simple est de l'utiliser :

$$\text{âge moyen} = \frac{\text{sommes des âges}}{\text{effectif}} = \frac{6120{,}25}{253} \approx 24{,}2 \text{ ans}$$ **0,5 pt**

On met ici des points sur la bonne acquisition des réflexes de calcul, en premier celui de tenir compte des effectifs des classes, dont l'impératif est rappelé dans la question.

$$\text{Variance de l'âge} = \text{Moyenne des Carrés} - \text{Carré de la Moyenne} = \left(\frac{154924{,}2}{253}\right) - \left(\frac{6120{,}25}{253}\right)^2 \approx 27{,}2$$

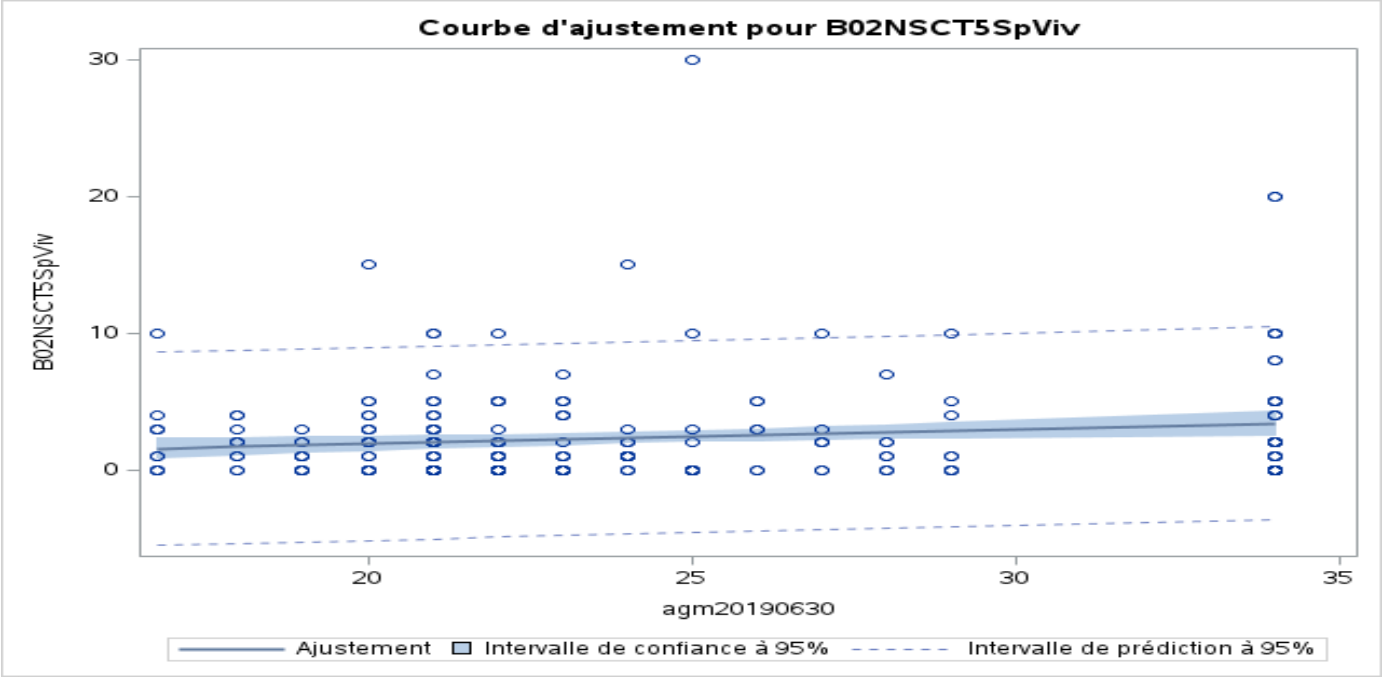
$$\text{Écart-type} = \sqrt{\text{Variance}} = \sqrt{\left(\frac{154924{,}2}{253}\right) - \left(\frac{6120{,}25}{253}\right)^2} \approx 5{,}21$$ *ND*

$$CV = \text{écart} - \frac{\text{type}}{\text{moyenne}} = \frac{\sqrt{\left(\frac{154924{,}2}{253}\right) - \left(\frac{6120{,}25}{253}\right)^2}}{\frac{6120{,}25}{253}} \approx 0{,}215$$ *ND*

$$\text{Covariance}(\text{âge}, \text{sorties}) = \underset{\text{des Produits}}{\text{Moyenne}} - \underset{\text{des Moyennes}}{\text{Produit}} = \frac{15010{,}8}{253} - \left(\frac{6120{,}25}{253}\right) \cdot \left(\frac{590}{253}\right) \approx 2{,}92$$ *1 pts*

$$\text{Corr}(\text{âge}, \text{sorties}) = \frac{\text{Covariance}(\text{âge}, \text{sorties})}{\text{Produit des écarts-types}} = \frac{\frac{15010{,}8}{253} - \left(\frac{6120{,}25}{253}\right) \cdot \left(\frac{590}{253}\right)}{\sqrt{\frac{154924{,}2}{253} - \left(\frac{6120{,}25}{253}\right)^2} \cdot \sqrt{\frac{4646}{253} - \left(\frac{590}{253}\right)^2}} \approx 0{,}156$$ *1 pts*

2. Déterminer les coefficients et le coefficient de détermination dans la régression linéaire du nombre de sorties au cirque et au théâtre sur l'âge. **3 pts**



*Équation de la régression linéaire : sorties = A * âge + B + Résidu*

avec $$A = \frac{cov(\text{sorties}, \text{âge})}{var(\text{âge})} = \frac{\frac{15010{,}8}{253} - \left(\frac{6120{,}25}{253}\right) \cdot \left(\frac{590}{253}\right)}{\frac{154924{,}2}{253} - \left(\frac{6120{,}25}{253}\right)^2} \approx 0{,}107$$ *1 pt*

et $B = \left(\begin{matrix}Moyenne\\Sorties\end{matrix}\right) - A \cdot \left(\begin{matrix}Moyenne\\âges\end{matrix}\right) \approx -0{,}267$ *0,5 pt*

On rappelle que le coefficient de détermination mesure la part de la variance de la variable expliquée par le modèle, la variance expliquée étant égale à la variance totale diminuée de celle des résidus suivant l'équation de décomposition de la variance :

$$Variance\ des\ sorties = A^2 \cdot (variance\ de\ l'\ âge) + Variance\ des\ résidus$$

d'où il vient $R^2 = \dfrac{A^2 \cdot (Variance\ de\ l'\ âge)}{Variance\ des\ sorties} \approx 0{,}0243$ *1,5 pts*

Suite aux résultats qui précèdent, on décide de construire le tableau suivant pour mieux appréhender le lien entre l'âge des enquêté·es et la fréquentation des cirques et des théâtres.

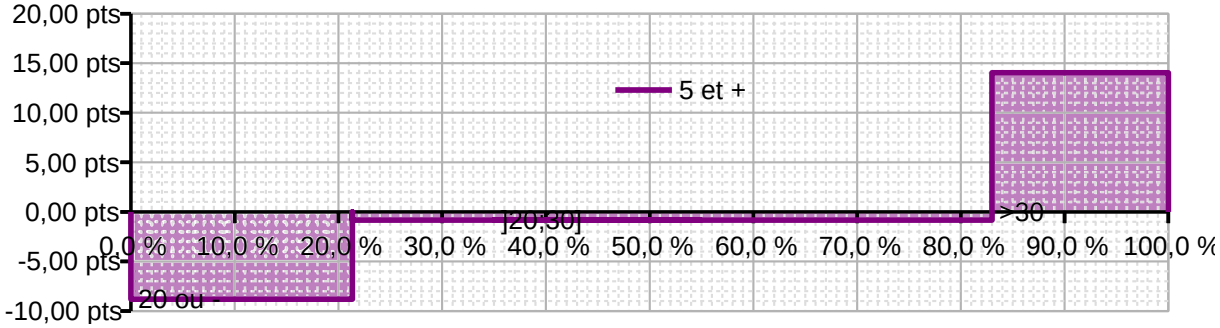
Effectifs	**Nombre de séances théâtre ou cirque**			
Âge 2019	**aucune**	**1-4 fois**	**5 et +**	**Total**
20 ou -	17	33	4	54
]20;30]	63	69	24	156
>30	14	16	13	43
Total	94	118	41	253
Fréquence manquante = 2				

3. Faire le tableau des % en ligne. Ajouter une colonne à droite des totaux avec la répartition par âge. **2 pts**

Nombre de sorties au théâtre ou au cirque en 2019 en fonction de l'âge

Âge \ Sorties	aucune	1-4 fois	5 et +	Total	Profil colonnes
20 ou -	31,5 %	61,1 %	7,4 %	100,0 %	21,3 %
]20;30]	40,4 %	44,2 %	15,4 %	100,0 %	61,7 %
>30	32,6 %	37,2 %	30,2 %	100,0 %	17,0 %
Total	37,2 %	46,6 %	16,2 %	100,0 %	100,0 %

4. Tracer le diagramme des écarts à l'indépendance pour les personnes qui se sont rendues 5 fois ou plus au théâtre ou au cirque. Les autres diagrammes des écarts à l'indépendance sont reproduits en annexe à la fin du sujet. **2 pts**



5. Calculer les effectifs à l'indépendance (théoriques). **1 pt**

Âge \ Sorties	aucune	1-4 fois	5 et +	Total
20 ou -	20,1	25,2	8,8	54
]20;30]	58,0	72,8	25,3	156
>30	16,0	20,1	7,0	43
Total	94	118	41	253

6. Calculer directement les écarts à l'indépendance représentés graphiquement à la question 4. **1 pt**

écarts	aucune	1-4 fois	5 et +	Total
20 ou -	-3,1	7,8	-4,8	0
]20;30]	5,0	-3,8	-1,3	0
>30	-2,0	-4,1	6,0	0
Total	0	0	0	0

7. Construire le tableau des effectifs représentant la situation où l'écart à l'indépendance est maximal en respectant le plus possible les écarts apparus dans le tableau des effectifs observés puis calculer le pourcentage de l'écart maximal (PEM) de ce tableau. **2 pts**

Eff EM	aucune	1-4 fois	5 et +	Total
20 ou -	0	54	0	54
]20;30]	92	64	0	156
>30	2	0	41	43
Total	94	118	41	253

Écarts EM	aucune	1-4 fois	5 et +	Total
20 ou -	-20,1	28,8	-8,8	0,0
]20;30]	34,0	-8,8	-25,3	0,0
>30	-14,0	-20,1	34,0	0,0
Total	0	0	0	0

Écarts + *18,8853754940711*
écart EM *96,8853754940712*
PEM *19,49 %*

8. Calculer les écarts pondérés à l'indépendance (contributions au Khi²) puis calculer le Khi² total du tableau et le degré de liberté. Si l'on admettait avoir ici un échantillon représentatif, pourrait-on en déduire que le lien entre les deux variables est significatif ? Quelles sont les cases du tableau qui contribuent le plus au Khi² ? **2 pts**

écarts pondérés	aucune	1-4 fois	5 et +	Total
20 ou -	0,47	2,42	2,58	5,47
]20;30]	0,44	0,19	0,06	0,70
>30	0,24	0,82	5,22	6,29
Total	1,15	3,44	7,87	12,45

DL	*4*
Seuil 5 %	*9,48772903678116*
Prob > khi²	*0,014275772614858*
Khi²	*12,4540357776998*
Phi²	*0,049225437856521*
Phi	*0,221868064075299*
V de Cramer	*0,110934032037649*

9. En utilisant le V de Cramer et/ou le PEM, le lien entre les deux variables apparaît-il fort ? **1 pt**

Lien relativement fort mais ce n'est pas massif

10. Commenter l'ensemble des résultats. Ceux-ci étant issus d'une enquête encore inachevée dont l'échantillon n'est pas tiré aléatoirement à partir d'une base de sondage dans la population cible (jeunes résidant en Ile de France, mais aussi dans d'autres régions), mais d'un appel à participation à partir de plusieurs réseaux sociaux au sens large, vous pourrez vous demander ce que ces résultats vous apprennent non seulement sur la population touchée par cet appel mais aussi sur le profil des personnes les plus susceptibles de répondre. **3 pts**

Ce point n'a pas fait l'objet d'un corrigé rédigé.

Annexe : diagrammes des écarts à l'indépendance non demandés dans le sujet