DATE :	**……21/12/2020-5/1/2021……**
EPREUVE:	**Statistiques et informatique appliquées aux Sciences sociales**
DIPLOME :	**Licence de Sciences sociales, 3e année** **Étudiant·es de tous parcours, inscrit·es en contrôle continu…**

Sujet obligatoire pour les étudiant·es inscrit·es en contrôle continu : Barême et corrigé

Ce devoir devait être rendu sous forme manuscrite, en main propre ou en version scannée ou photographiée sur le devoir. En cas d'impossibilité, il était indiqué de contacter l'enseignant de CM ainsi que le chargé de TD en expliquant bien la situation.
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'est pas prise en compte dans l'évaluation et vise 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. L'échantillon n'est donc 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. De plus, l'enquête est encore en cours au moment des calculs. Dans un premier temps, nous voulons créer une variable d'âge exploitable pour ensuite regarder comment varient les pratiques en fonction de cette variable. La question posée est « Quelle est votre année de naissance ? » avec la modalité et la répartition ci-dessous parmi les personnes ayant répondu au questionnaire avant le 16/12/2020 :

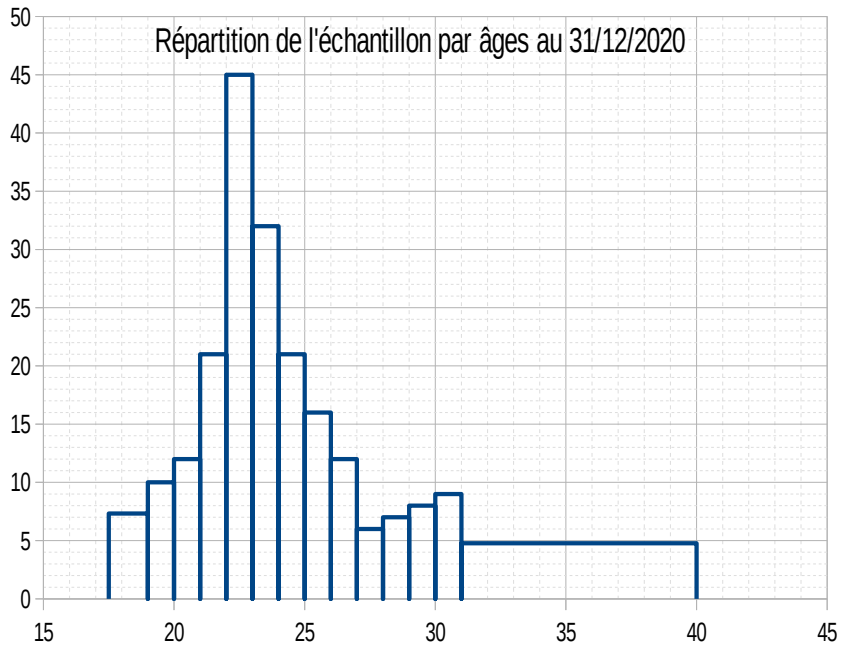
A02An-neenaiss (codes)	Année de Naissance (libellés)	Classe d'âge (au 31/12)	Âge moyen dans chaque classe	Effectifs	Somme des âges dans chaque classe	Somme des arrés des âges dans chaque classe
01	2002 ou après	[17,5;19[	18,25	11	200,75	3663,6875
02	2001	[19;20[	19,5	10	195	3802,5
03	2000	[20;21[	20,5	12	246	5043
04	1999	[21;22[	21,5	21	451,5	9707,25
05	1998	[22;23[	22,5	45	1012,5	22781,25
06	1997	[23;24[	23,5	32	752	17672
07	1996	[24;25[	24,5	21	514,5	12605,25
08	1995	[25;26[	25,5	16	408	10404
09	1994	[26;27[	26,5	12	318	8427
10	1993	[27;28[	27,5	6	165	4537,5
11	1992	[28;29[	28,5	7	199,5	5685,75
12	1991	[29;30[	29,5	8	236	6962
13	1990	[30;31[	30,5	9	274,5	8372,25
15	1989 ou avant	[31;40[	35,5	43	1526,5	54190,75
Total	Ensemble	[17,5;40]	/	**253**	**6499,75**	**173854,1875**
Z	Ne sait pas	Inconnue	/	2	NB : pas de refus	

1. Tracez l'histogramme correspondant à ces données d'âge au 31 décembre et commentez le cas échéant le choix des bornes de 17,5 ans et 40 ans. **3 pts**

Rappel : l'histogramme est un diagramme en barres verticales construit en général en positionnant la variable étudiée (ici l'âge) en abscisse et en traçant des rectangles dont la base correspond à chaque intervalle de classe sur cet axe des abscisses et dont la surface est proportionnelle aux effectifs de chaque classe ; pour le tracer, il convient donc de calculer le nombre d'individu moyen à chaque âge c'est à dire l'effectif de la classe divisé par son amplitude. Ce principe est bien connu des étudiant·es qui l'ont notamment vu dans les cours de démographie pour la construction des pyramides des âges et revu en CM et TD cette année.

J'ai fait directement les calculs sous excel mais je reprends ici le petit tableau permettant de faire le graphique. On remarque qu'ici la plupart des classes est d'amplitude 1. Le calcul des densité pour la première et la dernière classe permet d'éviter d'avoir une vision faussée de la distribution, l'oeil étant plus sensible aux surfaces qu'aux hauteurs. ***(1,5 point pour le principe, 1,5 point pour la réalisation)***

Classe d'âge (au 31/12)	Amplitude de chaque classe	Effectif	Densité (nombre moyen d'individu/âge)
[17,5;19[	1,5	11	7,33
[19;20[	1	10	10
[20;21[	1	12	12
[21;22[	1	21	21
[22;23[	1	45	45
[23;24[	1	32	32
[24;25[	1	21	21
[25;26[	1	16	16
[26;27[	1	12	12
[27;28[	1	6	6
[28;29[	1	7	7
[29;30[	1	8	8
[30;31[	1	9	9
[31;40[	9	43	4,78

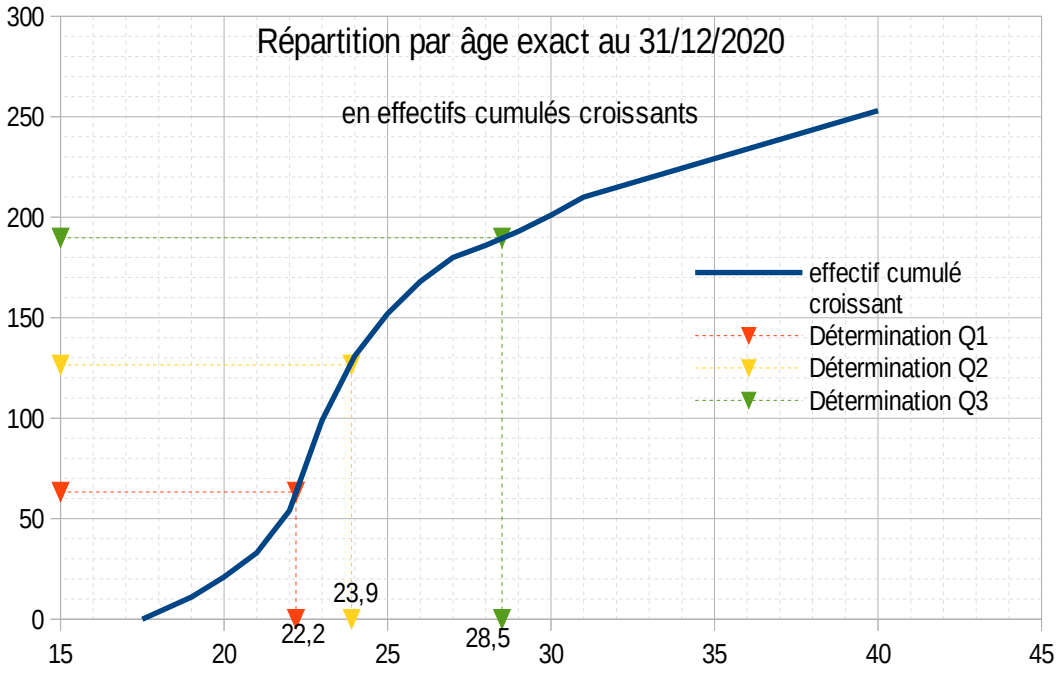


2. Quelle est la médiane, quels sont les premiers et troisième quartiles de cette distribution ? **3 pts**

Remarque : l'énoncé incite à traiter cette question en cohérence avec la question précédente, en travaillant avec l'âge exact au 31/12/2020 (« cette distribution »), c'est à dire par interpolation ; il faut toutefois ne pas trop pénaliser les étudiant·es qui reviendraient à l'âge atteint dans l'année 2020 (en années révolues au 31/12/2020), ce que permettent ici les classes étudiées, du fait que ni la médiane ni les quartiles ne sont dans les classes extrêmes. ***(1,5 pts compréhension notion, 1,5 pts calculs justes)***

Pour l'âge exact, on reconstitue par interpolation linéaire les effectifs cumulés croissants.

âge exact au 31/12/2020	effectif cumulé croissant
17,5	0
19	11
20	21
21	33
22	54
23	99
24	131
25	152
26	168
27	180
28	186
29	193
30	201
31	210
40	253



NB : le graphique n'était pas demandé.

Rappel de la formule : on détermine d'abord la classe où se trouve le quantile recherché Q_i pour i=1,2,3, en calculant l'effectif cumulé croissant correspondant $N\uparrow Q_i$. On note b_{inf} et b_{sup} respectivement les bornes inférieure et supérieure de cette classe et $N\uparrow_{inf}$ et $N\uparrow_{sup}$ les effectifs cumulés correspondants, qui sont tels que $N\uparrow_{inf} \leq N\uparrow Q_i \leq N\uparrow_{sup}$ alors $Q_i = b_{inf} + amplitude \cdot \frac{N\uparrow Q_i - N\uparrow inf}{N\uparrow^{-N}\uparrow inf}$ *et on doit vérifier $b_{inf} \leq Q_i \leq b_{sup}$.*

Pour Q_1, $N\uparrow Q_1 = 253/4 = 63{,}25$; $b_{inf}=22$ et $b_{sup}=23$ ($N\uparrow_{inf} = 54$ et $N\uparrow_{sup} = 99$) donc

$$Q_1 = 22 + 1\cdot\frac{63{,}25-54}{99-54} \approx 22{,}2 \quad ; \quad Q_2 = 23 + 1\cdot\frac{126{,}5-99}{131-99} \approx 23{,}9 \quad ; \quad Q_3 = 28 + 1\cdot\frac{189{,}75-186}{193-186} \approx 28{,}5$$

Pour l'âge en années révolues, on cherche à chaque fois l'âge faisant passer de moins de 25 % à plus de 25 % de l'effectif cumulé pour le 1er quartile (respectivement 50 % pour la médiane, 75 % pour le 3e quartile). Les âges sont transformés et les quartiles déterminés comme suit :

Année de Naissance (libellés)	âge atteint en 2020	effectif cumulé croissant antérieur	effectif cumulé croissant postérieur	Détermination quantiles (effectifs cumulés correspondants)
2002 ou après	18	inconnu	11	
2001	19	11	21	
2000	20	21	33	D1 = 19 ans (25,3)
1999	21	33	54	
1998	22	54	99	Q1 = 22 ans (63,25)
1997	23	99	131	Q2 = 23 ans (126,5)
1996	24	131	152	
1995	25	152	168	
1994	26	168	180	
1993	27	180	186	
1992	28	186	193	Q3 = 28 ans (189,75)
1991	29	193	201	
1990	30	201	210	
1989 ou avant	31	210	inconnu	

Remarque : la détermination des quartiles est plus facile dans ce cas-là ; le résultat inférieur tient au fait qu'au 31/12, l'âge exact est forcément supérieur à l'âge en années révolues. On peut en somme tenir le résultat en années révolues comme moins précis.

3. En faisant comme si chaque individu avait l'âge moyen de sa classe, calculez la moyenne d'âge au 31/12/2020 (de l'ordre de 25-26 ans) , l'écart-type (de l'ordre de 5 ans) et le coefficient de variation (écart-type divisé par la moyenne) de cette distribution. Attention à bien tenir compte de l'effectif de chaque classe. **3 pts**

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

$$\text{âge moyen} = \frac{\text{sommes des âges}}{\text{effectif}} = \frac{6499{,}75}{253} \approx 25{,}7 \text{ ans}$$ ***1 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.

$$\text{Variance de l'âge} = \text{Moyenne des Carrés} - \text{Carré de la Moyenne} = \left(\frac{173854{,}1875}{253}\right) - \left(\frac{6499{,}75}{253}\right)^2 \approx 27{,}2$$ *1 pt*

Compter 1,5 pour l'écart-type si le calcul est fait directement sinon 0,5 pour la racine carrée.

$$\text{Écart-type} = \sqrt{\text{Variance}} = \sqrt{\left(\frac{173854{,}1875}{253}\right) - \left(\frac{6499{,}75}{253}\right)^2} \approx 5{,}21$$ *0,5 pt*

Coefficient de variation

$$CV = \text{écart-type} \div \text{moyenne} = \frac{\sqrt{\left(\frac{173854{,}1875}{253}\right) - \left(\frac{6499{,}75}{253}\right)^2}}{\frac{6499{,}75}{253}} \approx 0{,}203$$ *0,5 pt*

On se demande comment varie le nombre de sorties en 2019 en fonction de l'âge en milieu d'année. Ceci revient à décaler les âges de 1,5 ans dans le passé. 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							Moyennes	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 :Monu-ments	-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

4. À l'aide des sommes de produits ci-dessous, expliquer les étapes du calcul 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

$$Covariance(âge, sorties) = \underset{des\ Produits}{Moyenne} - \underset{des\ Moyennes}{Produit} = \frac{15010,8}{253} - \left(\frac{6120,25}{253}\right)\cdot\left(\frac{590}{253}\right) \approx 2,92$$ *1,5 pts*

$$Corr(âge, sorties) = \frac{Covariance(âge, sorties)}{Produit\ des\ \text{écarts-types}} = \frac{\frac{15010,8}{253} - \left(\frac{6120,25}{253}\right)\cdot\left(\frac{590}{253}\right)}{\sqrt{\frac{154924,2}{253} - \left(\frac{6120,25}{253}\right)^2}\cdot\sqrt{\frac{4646}{253} - \left(\frac{590}{253}\right)^2}} \approx 0,156$$ *1,5 pts*

5. Déterminer les coefficients et le coefficient de détermination dans la régression linéaire du nombre de sorties au cirque et au théâtre sur l'âge. **4 pts**

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

avec $A = \frac{cov(sorties, âge)}{var(âge)} = \frac{\frac{15010,8}{253} - \left(\frac{6120,25}{253}\right)\cdot\left(\frac{590}{253}\right)}{\frac{154924,2}{253} - \left(\frac{6120,25}{253}\right)^2} \approx 0,107$ *1,5 pts*

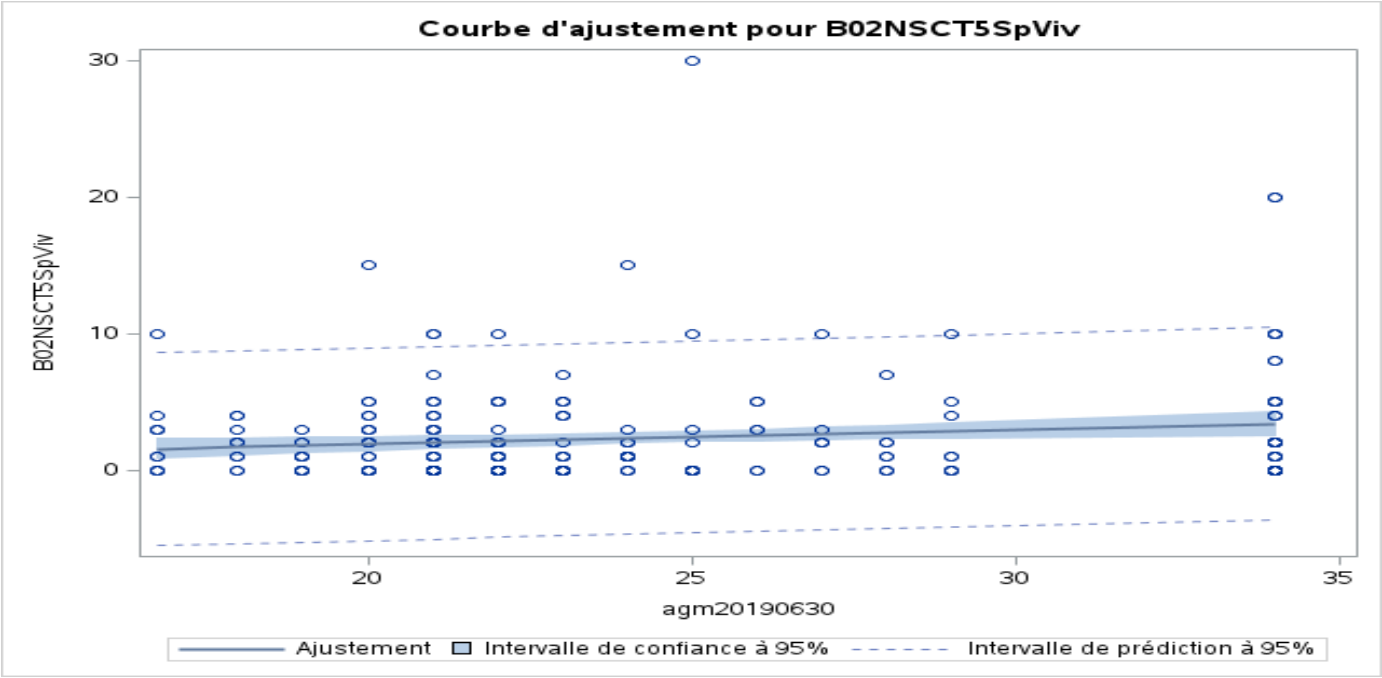
et $B = \left(\underset{Sorties}{Moyenne}\right) - A\cdot\left(\underset{âges}{Moyenne}\right) \approx -0,267$ *1 pt*

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

6. Commentez ces résultats. Vous pourrez utiliser le graphique ci-dessous pour visualiser l'ajustement linéaire présenté à la question précédente. **4 pts**



Les âges des personnes enquêtées jusqu'ici sont concentrés en dessous de 23 ans (révolus) pour plus de la moitié de l'effectif interrogé. Au delà, on observe toutefois un éventail d'âges plus ouvert, plus du quart des enquêté·es ayant 28 ans ou plus. Cette distribution asymétrique décale l'âge moyen au delà de 25 ans (25,7 ans exactement), la moyenne étant sensible aux valeurs extrêmes, alors que la médiane (de 23,9 ans exactement) leur est insensible. Dans tous les cas, cette distribution d'âge apparaît relativement dispersée pour une enquête initialement centrée sur les jeunes avec un écart-type de plus de 5 ans et un coefficient de variation de l'ordre de 20 %.

On s'intéresse ici à la variation du nombre de sorties culturelles en fonction de l'âge : les différents types de sorties sont corrélés entre eux ; en revanche, la plupart ne semble pas faire apparaître de corrélation linéaire significative avec l'âge, à l'exception des concerts et des sorties au cirque et au théâtre, qui semblent augmenter avec l'âge. Les sorties au cirque ou au théâtre étant les mieux corrélées (positivement) à l'âge, on a cherché ici à mesurer l'impact de celui-ci : il en ressort que le nombre moyen de sorties au cirque ou au théâtre augmente d'environ 1 sortie supplémentaire en 2019 pour une différence d'âge de dix ans de plus dans cet échantillon. Un tel effet (linéaire) apparaît très limité et c'est bien ce que montre le calcul du coefficient de détermination, établissant que l'âge n'explique qu'environ un quarantième (2,5%) de la variance du nombre de sorties au cirque ou au théâtre, la variance résiduelle étant donc, et de loin, beaucoup plus importante. On le visualise bien sur le graphique du nuage de points : on trouve à tout âge des personnes sortant peu ou beaucoup, mais le calcul montre que cela va quand même dans le sens de cette augmentation graduelle.

Il serait donc ici utile de prendre en compte d'autres facteurs, comme le genre, le niveau d'étude, la profession des parents, l'activité (étudiant, salarié ou autre) et d'autres éléments sur le mode de vie, pour voir si le lien demeure. On pourrait aussi souhaiter une approche par tranches d'âge en comparant celles qui sortent plus souvent au moyen d'intervalles de fréquences. Cette approche est proposée dans le sujet pour le contrôle terminal avec des tableaux croisés et des résultats plus nets. On pourrait également restreindre l'étude aux personnes sortant au moins une fois, l'absence de sortie entraînant peut-être une non-linéarité de la distribution.