UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Corrigé et Barême

DATE :	**……5…/…1…/…2022**
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 :	**…3. h**
DOCUMENT AUTORISE	**Calculatrice. Aucun document autorisé (cf. rappels en fin de sujet)**

Total sur 23 points, ramenés à un maximum de 20/20 pour toutes les notes supérieures à 16/20.

Les notes reçues aux épreuves du baccalauréat anticipé se répartissent de la façon suivante parmi les candidat·es à une licence du domaine SHS à l'Université d'Ile de France :

Variable d'analyse : moynbea Moyenne des notes du bac / epreuves anticipees								
Moyenne des notes du bac / epreuves anticipees	**Effectif**	**Moyenne ~~exacte~~** *(arrondies)*	**Somme des notes par classe**	**Amplit class**	***Densité moyenne***	***Valeurs seuils***	***Effectif cumulé***	**Somme des carrés**
]0 ; 5[	**3**	4,69	14,08	*5*	*0,6*	***0***	*0*	327891,97
[5 ; 10[	**292**	8,53	2489,75	*5*	*58,4*	***5***	*3*	
[10 ; 12[(Passable)	**478**	10,91	5213,83	*2*	*239*	***10***	*295*	
[12 ; 14[(AB)	**506**	**12,81785**	*6485,83*	*2*	*253*	***12***	*773*	
[14 ; 16[(B)	**438**	14,79	6478,58	*2*	*219*	***14***	*1279*	
[16 ; 18[(TB)	**213**	16,68	3552,83	*2*	*106,5*	***16***	*1717*	
[18 ; 20[(TB)	**31**	18,54	574,83	*2*	*15,5*	***18***	*1930*	
Total	*1961*	*12,65*	*24809,75*	*20*	*98,05*	***20***	*1961*	

Question 1 : Évaluez la moyenne et l'écart-type de cette distribution, en complétant si besoin le tableau. Tracez un histogramme de cette distribution et indiquez quelle est la classe modale. Évaluez graphiquement au moyen d'un diagramme cumulé croissant ou d'un calcul d'interpolation les quartiles dont la médiane. Enfin, dessinez la boite à moustache de cette distribution. **Total : 10,5 pts**

Pour calculer la moyenne des notes moyennes au baccalauréat anticipé de français, il est nécessaire de déterminer le total de toutes les notes ainsi que le nombre de notes prises en comptes (candidat·es ayant une moyenne renseignées : Effectif = 3+292+…+31 = 1961 ***(1 point)***

*Les totaux des notes sont ici présentés par classe mais il en manque une, à savoir le total des notes entre 12 et 14. Cependant nous disposons d'une moyenne calculée très précisément qui est de 12,8 environ. Ceci permet de retrouver le total des notes entre 12 et 14 soit 12,8178524*506 ≈ 6485,83 et de compléter le tableau : Total de toutes les notes = 14,08+2489,75+…+574,83 ≈ 24809,75* ***(1 point)***

Moyenne ≈ 24809,75 / 1961 ≈ 12,65 sur vingt. ***(1 point)***

Pour le calcul de la variance et de l'écart-type on utilise ici la formule

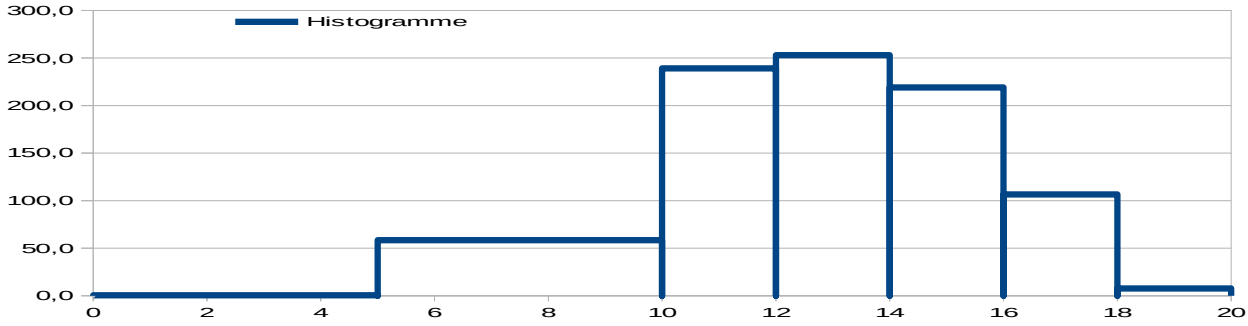
Variance = Moyenne des carrés – Carré de la moyenne = 327891,97 / 1961 – (24809,75/1961)² ≈ 7,144

$$\text{écart type} = \sqrt{\text{variance}} \approx \sqrt{7{,}144} \approx 2{,}67$$ ***(2 pts)***

La présentation des données dans ce sujet ayant dérouté un certain nombre d'étudiant·es, les étapes du calcul de la moyenne ont été valorisées (effectif, total des notes, application numérique finale). Il restait néanmoins essentiel de bien avoir compris la définition et le calcul de cet indicateur de base, en ayant

correctement défini la population statistique, constituée ici de 1961 candidat·es sur Parcoursup ayant une moyenne renseignée aux épreuves du baccalauréat anticipé. Ainsi un calcul reposant sur la somme des moyennes de classe divisé par le nombre de classe n'était pas acceptable puisqu'il revenait à supposer que chaque classe avait sensiblement le même effectif, ce qui était évidemment faux pour résumer la distribution des notes des candidat·es. Toutefois, les personnes ayant fait cette grossière erreur ont néanmoins pu recevoir les points correspondant à l'une ou l'autre des étapes du calcul correct si elles l'avaient réalisée.

*Pour tracer un histogramme, dont on rappelle qu'il représente chaque classe par un rectangle dont la **surface** doit être proportionnelle à **l'effectif**, il est nécessaire de calculer la hauteur des rectangles qu'on obtient en divisant l'effectif de classe par son amplitude qui donne elle la largeur du rectangle. Cet indicateur indique combien on a en moyenne de personnes par point gagné dans la classe considérée. Par exemple dans la classe des personnes ayant eu entre 5 et 10 on compte en tout 253 individus, ce qui fait qu'on comptera en moyenne 253/5=50,6 personnes par point gagné c'est à dire entre 5 et 6, 6 et 7, etc. On obtient la figure suivante donnant la répartition des candidat·es selon leur moyennes aux épreuves anticipées du baccalauréat, ce qui permet de visualiser immédiatement cette distribution :*



(2 pts)

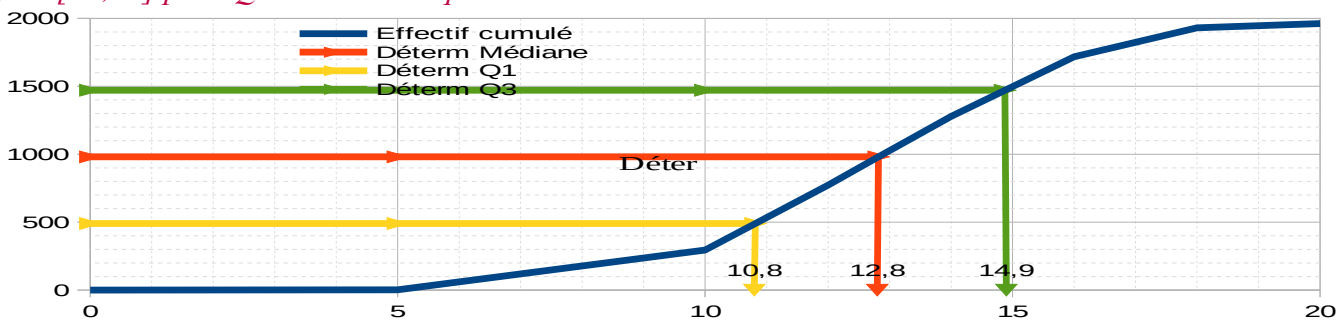
On constate que la classe modale (celle où l'on rencontre la plus grande concentration de personnes) est celle des candidat·es ayant reçu entre 12 et 14 sur 20. En l'occurrence, cette classe se trouvait également être la plus nombreuse dans l'absolu. ***(0,5 point)***

Pour la détermination des quantiles un graphique n'était pas exigé mais il restait nécessaire de calculer les effectifs cumulés correspondant aux valeurs seuils. Si beaucoup de personnes ont fait ce calcul, très peu ont correctement indiqué la valeur seuil correspondant à chaque effectif cumulé, ce qui les a alors souvent conduites à une mauvaise utilisation de la formule d'interpolation donnée en annexe. Beaucoup notamment ont calculé un centre de classe et fait comme si c'était à ce centre de classe que correspondait l'effectif cumulé calculé, ce qui conduisait à décaler la distribution de manière fautive.

Il suffisait pourtant d'interpréter littéralement les données en ayant bien compris la définition d'un effectif cumulé croissant (i.e. le nombre d'individus ayant moins que la note considérée) et de raisonner ensuite de proche en proche : personne n'avait moins de 0 donc l'effectif cumulé pour 0 était de 0, 3 personnes avaient de 0 à 5 ce qui faisait 3 personnes ayant moins de 5, 292 avaient entre 5 et 10 ce qui faisait en tout 295 personnes ayant moins de 10 et ainsi de suite jusqu'à la valeur maximale de 20 pour laquelle on retrouvait comme effectif cumulé le total de la population, personne n'ayant plus de 20.

*Cette étape une fois réalisée, il fallait rechercher l'effectif cumulé correspondant à chaque quantile : 1961/2=980,5 pour la médiane, 1961/4=490,25 pour Q1 et 3*1961/4=1470,75 pour Q3.*

Enfin on allait rechercher l'abscisse correspondant à chacun des effectifs cumulés ainsi calculés sur le graphique cumulé croissant ou l'on pouvait effectuer directement le calcul, ceci en prenant soin de déterminer dans quelle classe se trouvait chacun des quartiles : [12;14] pour la médiane, [10;12] pour Q1 et [14;16] pour Q3 le troisième quartile.



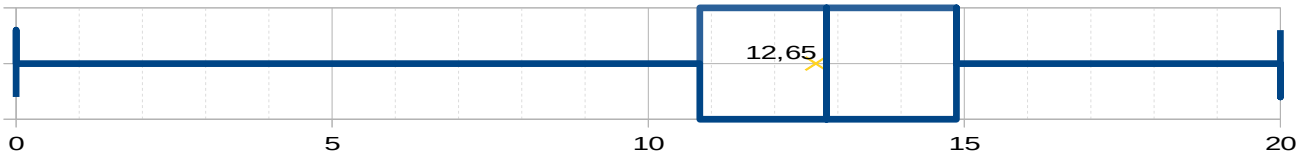
Interpolation des quartiles par le calcul : $Qtl = \text{b.inf} + \frac{n\uparrow Qtl - n\uparrow \text{inf}}{n\uparrow \text{sup} - n\uparrow \text{inf}} \cdot (\text{b.sup} - \text{b.inf})$ ***(2 pts)***

Médiane : $Q_2 \in [12;14]$ car $773<980,5<1279$, $Q_2 = 12 + \frac{980,5-773}{506} \cdot 2 \approx 12,82$

Premier quartile : $Q_1 \in [10;12]$ car $295<490,25<773$, $Q_1 = 10 + \frac{490,25-295}{478} \cdot 2 \approx 10,82$

Troisième quartile : $Q_3 \in [14;16]$ car $1279<1470,75<1717$, $Q_3 = 14 + \frac{1470,75-1279}{438} \cdot 2 \approx 14,88$

On utilise ces résultats pour la boîte à moustache ***(1 point)***



Question 2 : on constate que 129 candidat·es n'ont pas de notes renseignées aux épreuves anticipées du baccalauréat et que cela est plus fréquent parmi les personnes déjà bachelières (en réorientation ou reprise d'études) que parmi les personnes en train de passer le baccalauréat. Cependant, 36 personnes sans notes du baccalauréat anticipé ont une note moyenne du baccalauréat et 67 ont une mention renseignée. Ceci peut être lié une série professionnelle ou technologique où ces épreuves n'existaient pas. **Total : 7,5 pts**

Étudiez le tableau suivant puis évaluez la covariance et la corrélation, pour les personnes dont à la fois les notes du baccalauréat anticipé et la moyenne au baccalauréat sont renseignées. NB : du fait de cette restriction, les moyennes et variances calculées pour cette question peuvent donc différer de celles calculées à la question 1.

Déduisez-en les coefficients de la régression linéaire des notes au baccalauréat anticipé en fonction de la moyenne de l'ensemble du baccalauréat, ainsi que le coefficient de détermination de la régression.

Variable : notes	Somme des produits		Statistiques simples					
	Bac anticipé	Moy. Bac.	N	Moyenne	Ecart-type	Somme	Min.	Max.
Bac anticipé	97412,9412	100075,7310	628	12,21012	2,45738	7668	6,000	18,66667
Moy. Bac.	100075,7310	104874,4635	628	12,73627	2,18922	7998	8,570	19,58000

On peut utiliser la formule Covariance = Moyenne des produits – Produit des moyennes
Moyenne des produits = 100075,7310/628 ≈ 159,356
*Produit des moyennes = 12,21012*12,73627 ≈ 155,511*
Covariance ≈ 159,356-155,511 ≈ 3,845 ***(2 pts)***

$Corrélation = \frac{covariance}{Produit\ des\ écarts\ types} \approx \frac{3,845}{2,457 \cdot 2,189} \approx 0,715$ ***(1,5 point)***

Pour la régression linéaire, on utilise $Y = a \cdot X + b + U ; a = \frac{cov(X,Y)}{var(X)} ; b = \bar{Y} - a \cdot \bar{X}$ *avec X moyenne du baccalauréat et Y moyenne des notes du baccalauréat anticipées,*

$Var(X) \approx \frac{104874,46}{628} - 12,73627^2 \approx 2,18922^2 \approx 4,785$ *donc* $a \approx \frac{3,845}{4,785} \approx 0,80$ ***(2 pts)***

et $b \approx 12,21 - 0,80 * 12,74 \approx 2,0$ ***(1 point)***

Coefficient de détermination $R^2 = \frac{Var\ expl}{Var\ tot} = \frac{a^2 \cdot Var(X)}{Var(Y)} = correlat^2$ *donc* $R^2 \approx 0,715^2 \approx 51\ \%$ ***(1 point)***

Le choix des variables expliquée et explicative de cette régression est expliqué dans le commentaire ci-dessous : il était peu intuitif et beaucoup ont fait leurs calculs pour un choix inverse. Cette petite erreur d'inattention par rapport à l'énoncé n'a pas été pénalisée dans la notation du moment que les calculs menés étaient cohérents.

Question 3 : Commentez l'ensemble de vos résultats. Pour la corrélation, vous pourrez utiliser le résultat du test de Student qui indique que la probabilité d'obtenir la corrélation observée si le lien était dû au seul hasard de l'échantillonnage statistique est inférieure à 0,0001. **Total : 5 pts**

On étudie ici en premier lieu les notes obtenues par les personnes candidatant sur Parcoursup à une formation du domaine SHS (Sciences humaines et sociales) aux épreuves anticipées du baccalauréat. Il nous est indiqué que la plupart (94%) de ces candidat·es ont une note moyenne renseignée soit un effectif de 1961 personnes dont la plupart avait au moins 10/20, près des trois-quarts dépassant la note de 11/20 tandis que moins d'un quart de ces personnes avaient dépassé la note de 15/20, avec une moyenne de 12,6, un écart type de 2,7 points et une médiane proche de 13/20.

De manière assez prévisible, ces résultats anticipés, qui interviennent dans le calcul de la note globale au baccalauréat, apparaissent fortement et significativement corrélés positivement avec ce résultat définitif pour les 628 personnes dont on connaît les deux à la fois (avec une corrélation positive de 0,715) : on peut noter au passage que cette information n'est connue que pour environ un tiers des personnes candidates, essentiellement suite à une réorientation. À l'inverse, on a sans doute une proportion de néo-bacheliers de l'ordre des deux tiers dans ce vivier, dont les résultats définitifs ne sont donc pas encore connus.

Ceci étant précisé, on entreprend ici d'effectuer une imputation d'une note au baccalauréat anticipé pour ces personnes déjà bachelières pour lesquelles cette note est inconnue ou n'a pas de sens (pour certains diplômes étrangers ou pour les baccalauréats professionnels par exemple), sachant que la moyenne générale au baccalauréat permet de prédire 51 % des variations des notes au baccalauréat anticipé lorsque les deux sont connues et qu'un gain d'un point sur la note globale se traduit en moyenne par un gain de 0,8 points prévisible sur la moyenne des épreuves anticipées. Ceci permet donc de construire ce qu'on appelle une proxy de la note du baccalauréat anticipé pour les personnes qui n'ont pas une telle note, qui a l'avantage d'être déjà connue avec précision pour l'essentiel de l'échantillon pour des cas résiduels où l'on a d'autres informations, plus complètes, à savoir le résultat global à l'examen.

Il apparaît assez probable que la commission examinant les vœux sur cette formation ait pris en compte ces notes ou une proxy à défaut puisque l'issue de la procédure apparaît significativement et fortement liée à ce résultat. On constate en effet que les personnes reçues le sont plus fréquemment que la moyenne avec une moyenne au baccalauréat anticipé supérieure à 14/20 et elles ont également plus souvent les moyens de refuser la formation considérée avec ce type de notes, tandis que les personnes ayant des résultats inférieurs à ce seuil ont plus tendance à renoncer à ce vœu au profit d'une formation plus accessible ou à attendre en vain la proposition. Chez les personnes ayant eu moins de 12 voire de 10, le phénomène constaté est davantage un abandon de la plate-forme elle-même sans doute faute de réponses satisfaisantes. Une partie pourrait correspondre à un échec à l'examen lui-même.

Répartition des candidat·es selon l'issue de la procédure, détail des % en ligne selon la note réelle ou imputée au baccalauréat anticipé

%ligne	Renonce	Refuse	Reçue	Non reçue	Sorties de route	Ensemble
<10	58,8 %	0,0 %	0,0 %	19,3 %	21,9 %	100,0 %
[10 ; 12[(Passable)	59,0 %	3,6 %	4,2 %	18,8 %	14,4 %	100,0 %
[12 ; 14[(AB)	60,2 %	8,1 %	4,8 %	15,8 %	11,1 %	100,0 %
[14 ; 16[(B)	53,1 %	25,3 %	7,3 %	5,2 %	9,1 %	100,0 %
16 et + (TB)	20,7 %	62,2 %	6,9 %	0,4 %	9,8 %	100,0 %
Total	53,5 %	16,0 %	4,8 %	12,9 %	12,9 %	100,0 %

Issue de la procédure Parcoursup (modalités détaillées)
Renonce au vœu, avant d'avoir reçu la proposition
Refuse après réception de la proposition
Reçue : Personne acceptée et venue s'inscrire
Non-reçue : maintien du vœu mais sans proposition
Sorties de route : diverses démissions de la plate-forme à différents stades dont personnes acceptées non inscrites

Khi²	662,33	
DL	16	
Prob>Khi²	1,33E-130	
Significativité	5,00 %	1,00 %
Seuil	26,296	32,000
V de Cramer	0,28147	