A quelques points près : les notes au baccalauréat

Annie Uhry, Claudine Schwartz

On entend régulièrement parler, au vu d'expériences de multi-corrrections de copies d'élèves, du « scandale » suivant : des correcteurs d'une même copie, utilisant le même barème, ne lui attribuent pas la même note, et l'écart est parfois grand.

Si la note mise par différents correcteurs est variable, quel sens attribuer à « *la* » note d'une copie ?

L'existence d'un écart entre notes données à une même copie d'examen est perçue comme une injustice, et on a souvent tendance à oublier que la variabilité peut aussi bien être en faveur d'un candidat qu'en sa défaveur. Quittant le registre de l'émotion, nous analyserons cette variabilité, intrinsèque à toute correction. La question est d'en diminuer l'ampleur et pour cela, il convient de quantifier ce dont il s'agit. Le calcul d'indicateurs de variabilité, à partir de données observées, est ce qui nous occupe ci-dessous.

Après avoir proposé une définition de *la note* d'une copie, nous verrons comment en déduire la précision à accorder à une correction d'une copie de Français au baccalauréat de 1998. Cette précision sera estimée, assez grossièrement, par des méthodes simples et classiques de la statistique, appliquées aux notes données à 30 copies de français de baccalauréat lors de 10 corrections. Les 300 notes recueillies figurent dans le tableau 5 de l'annexe 1.

Nous estimerons de plus la probabilité pour que l'écart entre les notes données par deux correcteurs à une même copie de baccalauréat de français en 1998 soit supérieur ou égal à d, pour d=2,3,4,5 : cet indicateur de la variabilité des notes attribuées à une même copie est particulièrement parlant.

1-Observation des données

Pour chacune des 30 copies, on dispose de la note obtenue au bac et de 9 notes données par 9 correcteurs, tous correcteurs de cette épreuve de français, et qui ont travaillé sur des photocopies de chacune des copies, soit 300 notes en tout. Le tableau 1 donne les statistiques élémentaires pour l'ensemble ces notes.

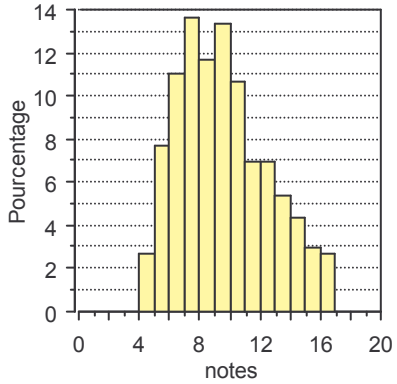
Les données ont été recueillies dans le cadre d'une étude faite par l'école active bilingue Jeanne Manuel (Paris), étude qui n'est pas celle que nous proposons ici. La qualité du travail entrepris au niveau de leur recueil permet de les utiliser ici.

moyenne	écart-type	nombre	minimum	maximum	médiane	étendue
9,1	3,0	300	4	16	9	11

Tableau 1

La figure 1 donne l'histogramme et la distribution de fréquence des 300 notes ; 30% d'entre elles valent 9, 10, ou 11 : un changement d'un point pour ces notes est particulièrement ressenti si 10 est une note couperet ! On constate aussi que 50% des notes sont entre 8 et 12 (8 et 12 inclus).

Le tableau 2-1 donne les statistiques par copies. Les étendues (différences entre la plus grande valeur et la plus petite valeur attribuées à une même copie) varient entre 3 et 11 points. C'est évidemment beaucoup. Notre propos n'est cependant pas ici de nous étonner mais de proposer des éléments au service d'une réflexion commune des enseignants en vue d'établir un jour des procédures visant à diminuer la variabilité des notes.



note	4	5	6	7	8	9	10	11	12	13	14	15	16
nombre	8	23	33	41	35	40	32	21	21	16	13	9	8
fréquence	2,7	7,7	11,0	13,7	11,7	13,3	10,7	7,0	7,0	5,3	4,3	3,0	2,7

Figure 1 : Histogramme des 300 notes et tableau des fréquences.

Résumés Statistiques par copie, avec correcteur 6

Copie	moyenne	Ecart-type	min	max	Etendue
L1	**14,0**	1,5	12	16	4
L2	**8,7**	2,1	5	11	6
L3	**8,9**	1,8	6	11	5
L4	**7,4**	2,0	4	10	6
L5	**8,5**	2,3	5	11	6
L6	**5,8**	1,7	4	9	5
L7	**10,4**	2,7	6	13	7
L8	**10,8**	3,3	5	16	11
L9	**9,6**	2,5	6	13	7
L10	**9,1**	2,3	6	12	6
ES1	**13,5**	2,0	10	16	6
ES2	**8,4**	1,6	5	10	5
ES3	**8,4**	2,1	5	12	7
ES4	**9,3**	2,3	7	15	8
ES5	**5,6**	1,2	4	8	4
ES6	**7,3**	1,8	4	10	6
ES7	**7,8**	1,9	5	11	6
ES8	**6,2**	1,0	5	8	3
ES9	**8,5**	1,4	7	11	4
ES10	**12,2**	2,0	9	15	6
S1	**13,7**	2,5	8	16	8
S2	**9,7**	1,8	7	12	5
S3	**11,5**	2,5	7	14	7
S4	**12,0**	2,6	7	16	9
S5	**6,9**	1,7	5	9	4
S6	**8,3**	2,5	5	14	9
S7	**8,5**	2,0	7	13	6
S8	**7,8**	1,7	5	10	5
S9	**6,5**	1,4	5	9	4
S10	**7,5**	2,0	4	11	7

(1)

Résumés Statistiques par copie, sans correcteur 6

Copie	moyenne	Ecart-type	min	max	Etendue
L1	**14,2**	1,4	12	16	4
L2	**9,1**	1,7	6	11	5
L3	**9,2**	1,6	6	11	5
L4	**7,6**	2,1	4	10	6
L5	**8,9**	2,1	5	11	6
L6	**6,0**	1,7	4	9	5
L7	**10,9**	2,3	6	13	7
L8	**11,4**	2,7	7	16	9
L9	**9,7**	2,6	6	13	7
L10	**9,4**	2,1	6	12	6
ES1	**13,8**	1,9	10	16	6
ES2	**8,8**	1,1	7	10	3
ES3	**8,8**	1,8	6	12	6
ES4	**9,4**	2,4	7	15	8
ES5	**5,8**	1,1	4	8	4
ES6	**7,3**	1,9	4	10	6
ES7	**8,1**	1,8	6	11	5
ES8	**6,3**	1,0	5	8	3
ES9	**8,7**	1,4	7	11	4
ES10	**12,6**	1,7	10	15	5
S1	**14,3**	1,7	12	16	4
S2	**10,0**	1,6	7	12	5
S3	**12,0**	2,1	8	14	6
S4	**12,3**	2,5	7	16	9
S5	**7,1**	1,7	5	9	4
S6	**8,4**	2,6	5	14	9
S7	**8,7**	2,1	7	13	6
S8	**8,1**	1,5	6	10	4
S9	**6,7**	1,4	5	9	4
S10	**7,9**	1,7	6	11	5

(2)

Tableau 2 : Résumés statistiques par copie, avec 10 corrections (tableau 2-1) ou en éliminant le correcteur 6 (tableau 2-2). Les écart-types varient de 1 à 2,7 : ça peut sembler beaucoup, mais cela reste cependant dans l'ordre de grandeur de la fluctuation d'échantillonnage, vu de petit nombre de données sur lesquelles ces écarts type sont calculés (voir le dernier paragraphe de ce chapitre et la note de bas de page n°8).

Regardons les statistiques par correcteur (figure 2 et tableau 3).

Face au problème de la variabilité -incontournable- des notes, on croit souvent qu'harmoniser moyenne et écart-type entre les jurys garantit une faible variabilité de la note. Ce n'est pas le cas. Ainsi, si on considère les colonnes 1 et 2 (resp.1 et 4) du tableau 5 des notes (voir annexe 1), la moyenne et l'écart type des 30 notes sont proches (voir tableau 3) et cependant, l'écart va jusqu'à 7 points pour la copie L7 (resp. 4 points pour la copie L9).

Résumé Statistique par correcteur

	moyenne	Ecart-type	min	max	Etendue
Bac	**9,3**	3,1	4	16	12
C2	**8,8**	3,1	4	16	12
C3	**10,0**	2,7	5	14	9
C4	**9,3**	2,7	5	16	11
C5	**10,3**	2,4	6	15	9
C6	**6,5**	2,0	4	12	8
C7	**9,3**	3,4	4	16	12
C8	**8,9**	2,5	5	15	10
C9	**9,9**	3,3	6	16	10
C10	**8,8**	3,0	4	15	11

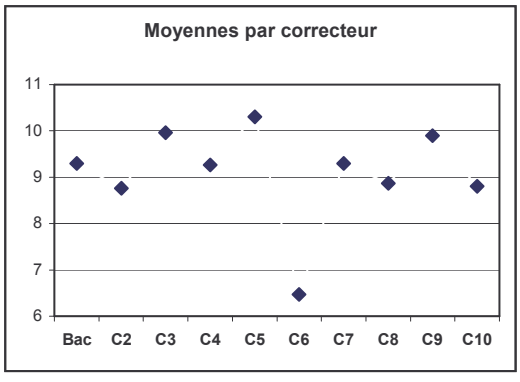
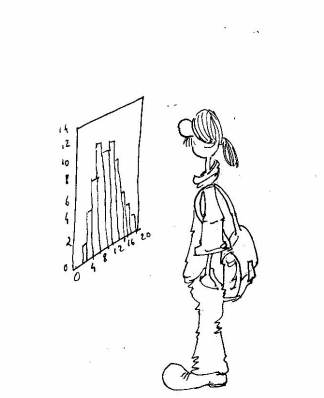


Tableau 3 Figure 2

Le sixième correcteur a une moyenne de notes particulièrement basse (il « note sec »). On peut aussi constater qu'il donne la note minimale pour 24 copies sur les 30. Dans la suite de ce texte, tous les calculs seront faits sans tenir compte de ce correcteur.

Le tableau 2-2 donne les statistiques par copie pour les 9 corrections restantes : toutes les moyennes obtenues sont légèrement plus élevées que les précédentes et les écarts-type diminuent pour presque toutes les copies. Entre le tableau 2-1 et le tableau 2-2, l'étendue maximale des 30 notes passe de 11 à 9, le maximum est encore atteint pour la copie L8, mais il est aussi atteint pour la copie S4. Notre objectif est ici de faire un étude optimiste, c'est-à-dire qui risque plutôt de sous-estimer la variabilité des notes ; il faudrait bien sur disposer de plus de données pour voir si la présence de correcteurs tels que celui que nous avons exclu est fréquente ou non.

Nous verrons en annexe 3 une procédure d'harmonisation des notes qui permettrait d'intégrer ce correcteur.



2- La note d'un copie

On pourrait ici décider que la note d'une copie est la moyenne des notes attribuées. Si on faisait corriger les copies par d'autres groupes de correcteurs, les notes moyennes seraient encore variables mais l'amplitude de leur variation serait plus faible que celle des notes individuelles. Il est cependant inenvisageable, au niveau du baccalauréat, de faire corriger les copies du bac par tant de correcteurs !

On aimerait se laisser aller à dire qu'un barème étant fixé, *la note* d'une copie est un nombre « objectivement » défini : ce serait la note d'un *correcteur idéal*. Cela ne définit pas pour autant *la* note puisqu'on ne sait pas non plus définir un *correcteur idéal* ! Cette vision incite à penser qu'il existe une note *juste*, celle du *correcteur idéal*, toute autre note étant *injuste* : ce point de vue nous parait…injuste.

Cette vision contient cependant l'idée intéressante que *la* note d'une copie est une « idéalité », une *note théorique* qu'on ne peut pas directement observer expérimentalement. Nous allons partir de là, en imposant toutefois que la définition choisie de *la note théorique* d'une copie permette de faire le lien entre ce nombre abstrait et la réalité observable. On doit

en particulier pouvoir énoncer des propriétés sur l'écart entre la note d'un correcteur quelconque et *la note théorique*.

3- Un modèle pour réfléchir

La seule réalité observable pour une copie donnée est l'ensemble des notes de tous les correcteurs possibles de cette copie. Nous allons modéliser cette situation en associant à la correction d'une copie C une loi de probabilité P_C, une note x donnée par un correcteur étant alors la réalisation d'une variable aléatoire de loi P_C .

Un modèle classique (qu'il conviendrait de valider sur les données) pour une telle situation consiste prendre pour loi P_C une loi de Gauss[1] . Soit μ_C la moyenne théorique, ou espérance, de cette loi.

Par convention, nous dirons que μ_C est *la note théorique* de la copie *C*.

On fait ici l'hypothèse que la *variabilité* ne dépend pas de la copie si celle-ci est suffisamment loin des notes extrêmes (0 ou 20). Autrement dit, on suppose que l'écart type σ de cette loi de Gauss est le même pour toutes les copies dont la note théorique est disons entre 5 et 15 : notre étude se limitera à l'ensemble E de telles copies.

Dans le cadre ainsi choisi, on peut calculer un intervalle de confiance associé à une note (cette note n'étant pas nécessairement arrondie à l'entier le plus proche). Si on se place au niveau de confiance 0,95, et si x est la note attribuée par un correcteur quelconque à la copie C, l'intervalle de confiance de μ_c est $[x-2\sigma, x+2\sigma]$, ce que nous écrirons :

Si x est la note d'un correcteur, la note théorique de la copie corrigée est égale à x avec une précision de 2σ, au niveau de confiance 0,95.

Comme on a supposé que l'écart-type σ est le même pour chaque copie d'un sous-ensemble E des copies du bac (celles dont la note théorique est entre 5 et 15), on en déduit :

- pour environ 95% des copies de E, leur note théorique μ sera dans l'intervalle $[x-2\sigma, x+2\sigma]$,
- pour environ 2,5% des copies de E , on aura $x+2\sigma < \mu$,
- pour environ 2,5% des copies de E , on aura $x-2\sigma > \mu$.

Si on a 300 000 copies (ordre de grandeur raisonnable pour des copies de bac entre 5 et 15), sous réserve que les corrections puissent être considérées comme indépendantes :

- environ 285 000 corrections s'écarteront de la note théorique d'au plus 2σ points,
- environ 15 000 corrections s'écarteront de la note théorique de plus de 2σ points[2] ; dans environ la moitié des cas cette « injustice du hasard » sera ressentie durement tandis que les autres candidats n'auront sans doute pas conscience que *le sort* leur a été favorable.

[1] Les notes étant ici données en nombre entier, les calculs relatifs à ce modèle devront être ajustés : on parle de corrections de continuité.

[2] Si on abaisse le niveau de confiance, la précision est meilleure. La précision est égale à σ au niveau de confiance 0,66.

Il convient maintenant d'attribuer une valeur à σ, à partir des 300 notes observées.

Comme nous l'avons dit, les données observées seront utilisées pour réfléchir, avoir des ordres de grandeur, élaborer des points de vue et définir d'autres études.

Il y a ici trop peu de copies pour vérifier avec une bonne marge de sécurité certaines hypothèses, comme celle d'une variance commune aux copies dont les notes ne sont pas trop basses ou trop élevées.

Nous donnons dans ce paragraphe les résultats en ne tenant pas compte des écarts entre correcteurs. On verra dans l'annexe 3 qu'ils sont faibles et qu'en tenir compte modifie très peu les résultats. La note du bac, qui n'est pas le fait d'un seul correcteur est traitée ici au même titre que les autres.

On calcule les moyennes m_i des notes de chaque copie ; m_i est une estimation de la note théorique μ_i de la copie i. On note m la moyenne des 270 notes :

$$m_i = \frac{1}{9}\sum_k x_{ik} \qquad m = \frac{1}{270}\sum_{i,k} x_{ij} = \frac{1}{30}\sum_i m_i$$

On trouve ici : m=9,4.

Notons $\hat{s}_i^2 = \frac{1}{8}\sum_k (x_{ik} - m_i)^2$; les valeur de $\hat{s}_i$ sont données dans le tableau 2-2, sous la colonne écart-type[3].

Un estimateur classique $\hat{s}^2$ de σ^2 consiste à prendre la moyenne arithmétique des estimations des variances avec chacune des copies, ce qui donne

$$\hat{s} = 1{,}89.$$

Pour les copies dont la note théorique est entre 5 et 15, la précision de la note donnée au baccalauréat est d'environ $\pm 2\sigma$, soit $\pm 3{,}8$ points, au niveau de confiance 0,95.

On a vu précédemment que 80% des notes observées sont entre 8 et 13 ; pour toutes ces notes x, la fourchette [x-3,8, x+3,8] de la note contient la « note barrière » 10 : voilà des chiffres qui plaident en faveur d'une réflexion sur les possibilités d'améliorer la correction, c'est-à-dire de diminuer σ !

Le cas de la double correction

---

[3] Il y a en statistique deux versions de l'écart-type d'une série de données. L'une est $\hat{s}_i$ (formule avec n-1au dénominateur, où n est le nombre des données) et l'autre serait ici $s_i = \sqrt{\sum_k (x_{ik} - m_i)^2 / 9}$ (formule avec n au dénominateur). Les deux formules correspondent à des usages différents. En l'absence de modélisation, c'est-à-dire tant qu'on reste dans la description des données, on peut utiliser s_i qui permet des décompositions de variance telles la formule (2) de l'annexe 3.Pour certains calculs théoriques et pour estimer un paramètre d'un modèle, c'est $\hat{s}_i$ qu'on doit utiliser. Heureusement, en pratique, n est *grand* et les valeurs numériques obtenues pour $\hat{s}_i$ et s_i sont proches.

Imaginons une double correction de chaque copie à l'issue de laquelle la note retenue est la moyenne y des deux notes. Dans le modèle choisi, la loi de la note y sera une loi normale de moyenne μ et d'écart type $\sigma/\sqrt{2}$. La précision sur la moyenne des deux notes, au niveau de confiance 0,95 vaut $2\,\sigma/\sqrt{2}=\sqrt{2}\,\sigma$; le rapport des précisions est donc $1/\sqrt{2}\approx 0{,}7$:

La double correction, suivie d'une moyenne des deux notes, améliore la précision de 30%

Dans notre exemple, σ est voisin de 2. Au niveau de confiance 0,95, la double correction fait passer d'une précision d'environ ± 4 points à une précision d'environ ± 3 points. Il serait ici intéressant de trouver une méthode de correction du baccalauréat de français moins coûteuse et plus efficace !

5- Un autre indicateur de variabilité des notes

Prendre σ comme indicateur de variabilité est classique et très parlant….pour ceux qui connaissent un peu de statistique. Nous allons maintenant nous intéresser à une autre manière d'appréhender la variabilité des notes, en considérant, pour d=2,3,4,5 les probabilités $\pi(d)$ pour que deux correcteurs, choisis au hasard dans la population des correcteurs potentiels, donnent pour une même copie des notes dont l'écart $|x'-x|$ est supérieur ou égal à d.

Nous allons estimer ces probabilités à partir des données, de deux façons différentes. Un premier estimateur, dit non paramétrique, ne fait intervenir aucun modèle des données. Le second, paramétrique est calculé en estimant les paramètres du modèle choisi puis en faisant les calculs dans ce modèle :

(1) Une première estimation, notée $\tilde{\pi}(d)$, consiste à dénombrer, parmi les N=30n(n-1)/2 doubles corrections des trente copies faisant intervenir deux correcteurs parmi n=9 correcteurs, le pourcentage de celles dont l'écart absolu est supérieur à d. On a ici N=1080.

On notera que comme les N doubles corrections ne sont pas indépendantes (chaque note intervient 8 fois dans le calcul de $\tilde{\pi}(d)$), la précision de cette estimation n'est pas en $1/\sqrt{N}$.

(2) Une deuxième estimation de $\pi(d)$, que nous noterons $\hat{\pi}(d)$, consiste, dans le modèle considéré, à estimer σ puis à faire ensuite le calcul de $\pi(d)$ en remplaçant σ par son estimation.

Pour une même copie, les notes x et x' des deux correcteurs choisis au hasard sont, des réalisations de deux variables X et X' indépendantes et de loi N(μ_c , σ). En fait les notes dont nous disposons sont arrondies à l'entier le plus proche et si nous voulons comparer cet estimateur avec le précédent, il convient de faire une correction de continuité. Une manière simple de faire la correction de continuité consiste à remplacer $\pi(d)$ par :

$$\pi(d) = \text{Prob}(|X'-X| > d\text{-}0{,}5) = 1\text{-Prob}(-d\text{-}0{,}5 \le X'\text{-}X \le d+0{,}5)\ ,$$

qui se calcule aisément puisque la variable aléatoire X-X' suit la loi N(0, $\sqrt{2}\sigma$).

L'intervalle de confiance de $\pi(d)$ est relié à la précision de l'estimation de σ, et est délicat à déterminer. Nous ne le calculons pas ici.

Les résultats de ces deux estimations sont donnés dans le tableau 4. Ils donnent des résultats sensiblement égaux et on voit ainsi qu'il y a environ une chance sur trois pour que l'écart des notes entre deux copies soit supérieur ou égal à 3, presque une chance sur 5 qu'il soit supérieur ou égal à 4.

d	2	3	4	5	6
$100\,\tilde{\pi}(d)$	**56,8**	**33,4**	**18,7**	**9,3**	**4,1**
$100\,\hat{\pi}(d)$	**57,5**	**35,0**	**19,0**	**9,2**	**4,0**

Tableau 4 : Deux estimations des probabilités que deux correcteurs, choisis au hasard parmi les correcteurs du bac, mettent à une même copie des notes s'écartant de d points ou plus. Les calculs sont faits en tenant compte des 8 correcteurs et de la note du bac.

On pourra consulter l'annexe 2 pour les calculs de ces estimateurs lorsqu'on ne tient pas compte de la note obtenue au baccalauréat : les résultats obtenus sont très voisins.

On pourra aussi consulter l'annexe 3 qui donne quelques résultats relatifs à une étude avec effet *correcteur*.

Conclusion

Les données dont nous disposons nous ont conduit à estimer la précision des notes à environ 4 points, au niveau de confiance 0,95, et à estimer à environ une chance sur 3 la probabilité que deux correcteurs donnent à une même copie des notes dont l'écart est au moins 3 points. Si ces résultats donnent un ordre de grandeur il ne faut pas les prendre pour argent comptant. Il s'agissait ici plus d'avancer dans la réflexion que de prouver.

Nous n'avons parlé que de copies du bac de français de 1998, et les résultats numériques obtenus sont à valider, section par section, avec une étude plus importante. Ils ne sont pas extrapolables à d'autres matières, ni aux copies de la même discipline à d'autres examens ou concours, et ils peuvent évoluer selon les années.

Nous avons considéré des copies et non des élèves : ceux-ci peuvent plus ou moins réussir leur examen, ce qui introduit à leur niveau une autre source de variabilité de leur note au baccalauréat.

Notre propos était ici de montrer le type de concept et de méthodes qui permettent de définir et quantifier ce dont on parle, à savoir *la note* et sa *précision*. Par ailleurs, cette note vise à « mesurer » des qualités du candidat : définir ces qualités et savoir si la note en est une mesure pertinente relève d'autres études. De nombreuses études sont d'ailleurs régulièrement faites sur les multi-corrections dans le cadre de la docimologie.

Il est évidemment délicat, mais pas impossible, de trouver un consensus pour définir un niveau de variabilité à ne pas dépasser, autrement dit un maximum pour σ, ou pour $\pi(2)$ par exemple.

Faut-il avoir des barèmes très détaillés, quel type de sujet concevoir pour être en deçà d'un tel maximum?

Et avant de chercher de tels seuils, il faut aussi regarder ce qui se passe dans chaque matière et aussi estimer l'écart type de la note finale du baccalauréat. [4]

Enfin, on peut penser que des méthodes visant à diminuer l'écart type (ou $\pi(2)$) conduiraient de plus à améliorer la corrélation entre la note du bac et la moyenne pendant l'année : c'est une hypothèse à vérifier soigneusement.

Annexe 1

Tableau des notes et graphiques associés

	BAC	Cor2	Cor3	Cor4	Cor5	Co6	Cor7	Cor8	Cor9	Cor10
L1	15	12	13	16	14	12	16	13	15	14
L2	6	9	10	10	9	5	11	9	11	7
L3	6	8	9	10	11	6	9	11	9	10
L4	6	4	10	8	9	6	8	9	9	5
L5	5	11	11	11	9	5	10	7	8	8
L6	4	5	9	5	7	4	4	7	7	6
L7	13	6	13	13	12	6	9	11	10	11
L8	10	9	14	10	12	5	13	7	16	12
L9	12	6	13	8	11	9	12	11	7	7
L10	12	8	12	10	9	6	11	6	7	10
ES1	16	12	13	15	14	11	14	10	15	15
ES2	10	9	8	9	8	5	8	7	10	10
ES3	9	6	7	10	8	5	9	8	12	10
ES4	10	9	7	7	10	8	15	9	10	8
ES5	6	4	6	6	6	4	5	8	6	5
ES6	7	8	9	6	9	7	6	10	7	4
ES7	9	9	6	7	11	5	10	7	6	8
ES8	7	5	7	6	8	5	7	5	6	6
ES9	9	7	10	7	9	7	9	9	11	7
ES10	13	14	14	10	12	9	10	15	12	13
S1	13	16	12	12	15	8	16	14	16	15
S2	8	12	11	7	10	7	11	10	10	11
S3	13	14	12	13	12	7	13	8	14	9
S4	12	12	14	11	14	9	7	13	16	12
S5	9	6	5	9	9	5	5	7	8	6
S6	6	9	8	10	14	7	5	8	8	8
S7	9	8	7	11	13	7	7	8	8	7
S8	10	6	10	7	9	5	8	7	9	7
S9	6	9	8	8	7	5	5	5	6	6
S10	8	10	11	6	8	4	6	7	8	7

Tableau 5 : Les copies L1...L10 (resp. ES1...ES10 et S1…S10) sont des copies de bac de français d'élèves de la section L (resp. ES et S) venant d'un même lycée. Les 9 correcteurs ont corrigé des copies du bac de français 1998. Ils ont corrigé les 30 copies anonymées (photocopies de celles du baccalauréat) dans les mêmes conditions.
La colonne *Bac* donne les notes effectivement attribuées au bac pour ces copies. Elles ne sont pas le fait d'un unique correcteur, mais dans une première étude, nous mettrons sur le même niveau les 10 notes de chacune des copies. Nous excluons cependant cette note dans les calculs de l'annexe 3.

[4] Pour se faire une idée, si on prend la moyenne arithmétique de 9 matières ayant toutes un écart-type de 2, alors au niveau de confiance 0,95, la note finale sera à environ 1,3 point près, ce qui semble acceptable.

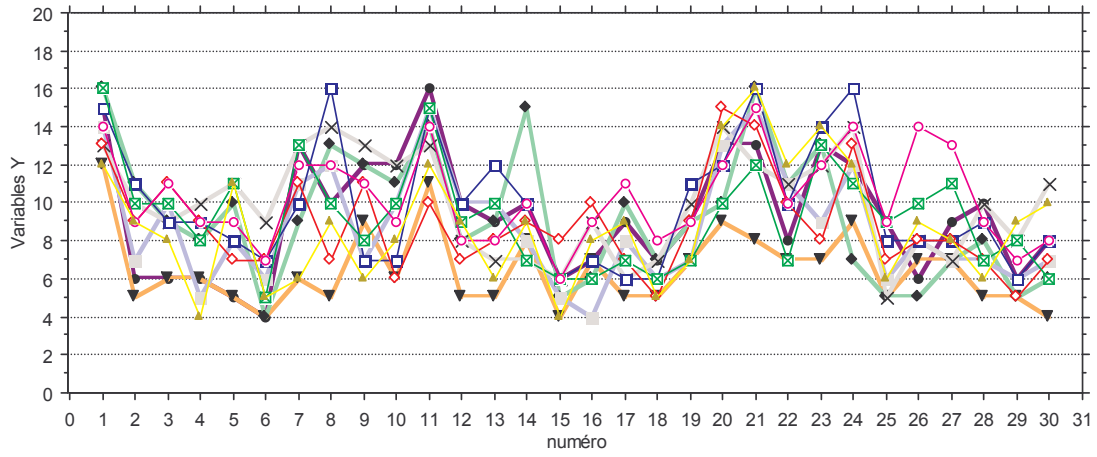


Figure 3 : représentation graphique associée au tableau 1.
Les 10 courbes correspondent aux 10 corrections. La courbe codée avec ▼ est associée au correcteur 6, qui a été ensuite exclu de l'étude. La courbe associée à la note du bac se repère par la note maximale, 16 donnée à la copie 11 (copie ES1) cette courbe est « dans le tas » formé par les autres courbes.

Annexe 2

Estimations des probabilités que deux correcteurs, choisis au hasard parmi les correcteurs du bac, mettent à une même copie des notes s'écartant de d points ou plus. Les calculs sont faits en tenant compte des 8 correcteurs (le correcteur 6 et la note obtenue au baccalauréat ont été exclues pour ces calculs).

d	2	3	4	5	6
$100\,\tilde{\pi}(d)$	**56,1**	**32,9**	**19.3**	**9,4**	**4,2**
$100\,\hat{\pi}(d)$	**57,6**	**35,1**	**19,2**	**9,4**	**4,0**

Tableau 7 : calculs en excluant les notes du baccalauréat.

Annexe 3

Dans cette annexe, notre objectif est de regarder l'importance de la variabilité due aux écarts moyens entre les différents correcteurs. Les calculs sont faits à partir des résultats de 8 correcteurs (numérotés ici de 1 à 8) : les notes du bac ne sont pas prises ne compte car elles ne sont pas le fait d'un unique correcteur.

Le modèle choisi, celui de l'analyse de la variance à 2 facteurs, implique d'en estimer les paramètres par des formules utilisant les degrés de liberté des variables aléatoires utilisées. La notion de degré de liberté n'est pas simple à appréhender, c'est pourquoi nous

commençons par aspect purement descriptif de l'échantillon des 240 données et une décomposition simple de leur variance ; ceci pour donner à la fois une première intuition de ce qu'est une analyse de la variance à deux facteurs et un ordre de grandeur de l'effet correcteur observé sur les données.

1-Quelques calculs sur l'échantillon de 240 notes (30 copies, 8 correcteurs)

1-1 Moyennes empiriques par copies ou par correcteur

On calcule les moyennes m_i des notes de chaque copie ; m_i est une estimation de la note théorique μ_i de la copie i. On note m la moyenne des 240 notes :

$$m_i = \frac{1}{8}\sum_k x_{ik} \qquad m = \frac{1}{240}\sum_{i,k} x_{ij} = \frac{1}{30}\sum_i m_i$$

On trouve ici : m=9,4 (à 0,05 près la moyenne générale est égale à celle où la note du bac est inclue).

Les moyennes par correcteurs sont :

$c_k = \frac{1}{30}\sum_{i=1}^{30} x_{ik}$ (voir ligne 2 du tableau 8).

1-2 Résidus empiriques

On définit les *résidus* e_{ik} par la formule :

$$x_{ik} = m_i + (c_k - m) + e_{ik}$$

On remarque que pour tout k : $\sum_i e_{ik} = 0$, et pour tout i : $\sum_k e_{ik} = 0$.

L'histogramme des résidus est donné figure 4.

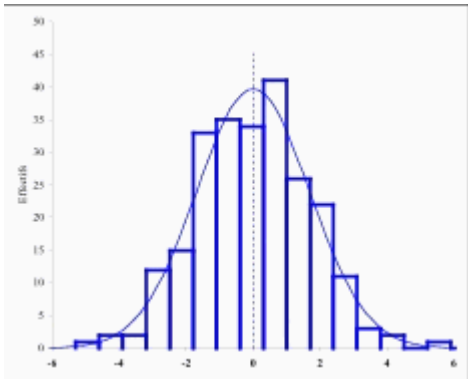


Figure 4 : histogramme des 240 résidus et densité de la loi de Gauss ayant même moyenne et même écart-type.

1-3 Décomposition de la variance entre les notes

On démontre aisément que la somme totale des écarts des observations autour de leur moyenne se décompose ainsi :

$$\sum_{i,k}(x_{ik} - m)^2 = 8\sum_i (m_i - m)^2 + 30\sum_k (c_k - m)^2 + \sum_{i,k} e_{ik}^2 \qquad (1)$$

Nous allons passer aux variances empiriques, prises ici égales à la somme des carrés des écarts à la moyenne divisée par le nombre d'observations.

D'où :

$$\frac{\sum_{i,k}(x_{ik}-m)^2}{240}=\frac{\sum_{i}(m_i-m)^2}{30}+\frac{\sum_{k}(c_k-m)^2}{8}+\frac{\sum_{i,k}e_{ik}^2}{240} \qquad (2)$$

Le premier membre de l'équation représente la *variance empirique* [5] entre toutes les notes. Le deuxième membre donne une décomposition de cette variance en somme de plusieurs termes : analyser une variance, c'est la décomposer en éléments explicables par des facteurs de variabilité (ici facteur copie et facteur correcteur) et en un terme résiduel. Ici la variance totale se décompose donc en trois termes :

- le premier est la *variance empirique* entre les moyennes des copies,
- le deuxième la *variance empirique* entre les moyennes des correcteurs,
- le troisième la *variance empirique* « résiduelle »[6], variabilité qui n'est expliquée ni par les écarts entre copies, ni par les écarts moyens entre correcteurs.

Les valeurs numériques sont ici :

$$\sum_{i,k}(x_{ik}-m)^2 \approx 2033 \qquad 8\sum_{i}(m_i-m)^2 \approx 1277$$

$$30\sum_{k}a_k^2 \approx 73 \qquad \sum_{i,k}e_{ik}^2 \approx 682$$

Et, pour les variances empiriques :

Variance totale	= 8,47
Variance entre les moyennes des copies	= 5,32
Variance entre les moyennes des correcteurs	= 0,31
Variance résiduelle	= 2,84

On constate que la variabilité entre les moyennes de ces correcteurs est faible par rapport à la variabilité résiduelle « inexpliquée ».
Pour pouvoir généraliser les calculs faits sur cet échantillon, il convient de définir un modèle et d'en estimer les paramètres, ce que nous allons faire maintenant.

2- Modèle de l'analyse de variance à deux facteurs

La note x_{ik} donnée par le correcteur k à la copie i, i = 1… 30 , k = 1.. 8 est ici considérée comme étant une réalisation d'une variable aléatoire X_{ik} , avec :

$$X_{ik} = \mu_i + \alpha_k + Z_{ik} \quad , \quad \text{où} \ \sum_k \alpha_k = 0.$$

[5] Les variances empiriques calculées ici sont égales à la somme des carrés des écarts à la moyenne divisée par le nombre d'éléments N de cette somme. On démontre que cette estimation des variances est « biaisée » (ici trop faible en moyenne), et on doit corriger en divisant par un nombre inférieur à N, appelé « degré de liberté », qui dépend du nombre d'observations et du nombre de paramètres estimés dans le modèle.

[6] Un *bon estimateur* de cette variance, en tenant compte des degrés de libertés est la somme des carrés des résidus divisé par (r-1)(n-1), où r est le nombre de correcteurs (r=8) et n le nombre de copies (n=30).

• μ_i représente la note théorique de la copie *i*.
• α_k représente l'effet de sévérité (ou d'indulgence) moyen du correcteur *k* par rapport à l'ensemble des 8 correcteurs.
• les Z_{ik} sont des variables aléatoires indépendantes, de même distribution, gaussienne centrée, d'écart-type[7] σ.

2-1 Estimations des moyennes

Les moyennes m_i par copies sont des estimateurs des notes μ_i des copies. Elles sont très voisines des moyennes obtenus avec la note du bac, nous ne les redonnons pas ici.

L'effet moyen α_k du correcteur *k* est alors estimé par $a_k=c_k-m$, k=1…8 (ligne 3 du tableau 8). (on a aussi $\sum_k a_k = 0$).

	1	2	3	4	5	6	7	8
c_k	8,77	9,97	9,27	10,30	9,30	8,87	9,90	8,80
a_k	-0,63	0,57	-0,13	0,90	-0,10	-0,53	0,50	-0,60

Tableau 8 : moyenne et écart moyen des 8 correcteurs

2-2 Harmonisation des notes

Au niveau de jurys de bac, une procédure d'harmonisation assez complexe permet de réduire en moyenne la sévérité excessive ou insuffisante des correcteurs. Pour les notes données, on essaye en pratique (au niveau des jurys de bac) de s'affranchir, de la *sévérité moyenne* plus ou moins grande des correcteurs.

Nous allons utiliser ici une procédure simple qui ramène la moyenne de chaque correcteur à la même valeur *m*. Les nouvelles notes y_{ik} harmonisées sont définies par :

$$y_{ik} = x_{ik} - a_k = m_i + e_{ik}$$

Pour ces notes harmonisées y_{ik}, , la note moyenne de la copie *i*, i=1…30, vaut m_i (puisque $\sum_k a_k = 0$) , et est donc égale à la moyenne des notes non harmonisées x_{ik} .

Avec ces nouvelles notes, la somme des carrés des écarts à la moyenne se décompose ainsi :

$$\sum_{i,k}(y_{ik}-m)^2 = 8\sum_i (m_i - m)^2 + \sum_{i,k} e_{ik}^2$$

Le premier terme du second membre de l'égalité est la partie de la somme des carrés des écarts qui est expliquée par les variations entre copies. Le deuxième terme est le « terme résiduel ».

2-3 Estimation de σ^2

Pour chaque copie $\hat{s}_i^2$ est une estimation de σ^2, avec :

$$\hat{s}_i^2 = \frac{\sum_k (y_{ik} - m_i)^2}{7}$$

[7] Cet écart-type σ n'est pas le même que celui du modèle initial (sans effet correcteur) ; il est a priori plus petit.

Les 30 valeurs de $\hat{s}_i^2$ trouvées varient de $0{,}7^2$ à $2{,}6^2$, ce qui pourrait faire douter de l'hypothèse d'un même σ pour toutes les copies. Mais chaque écart-type est calculé avec 8 données, ce qui conduit à une estimation peu précise[8]. En fait, on montre que cette disparité n'est pas significative au risque 0,05 et on admet donc l'hypothèse d'un même σ pour toutes les copies.

L'idée *naturelle* serait alors d'estimer la variance en prenant le moyenne arithmétique des 30 variances estimées, ce qui donnerait 3,24 et un écart-type de 1,80. Mais on a « perdu des degrés de liberté » dans la somme des carrés des résidus en imposant que les moyennes empiriques des correcteurs soient identiques ; les calculs théoriques montrent qu'il convient d'estimer l'écart-type σ par $\hat{s}$, avec[9] :

$$\hat{s} = \sqrt{\sum_i \hat{s}_i^2 / 29} \approx 1{,}83.$$

Enfin, un test classique de l'analyse de la variance montre que l'effet correcteur est non nul, au risque 0,05, mais il cependant de faible ampleur.

Conclusion

L'estimation de l'écart type σ, en excluant la note du bac mais sans négliger l'effet correcteur donne un écart-type du même ordre de grandeur que celui qui est calculé dans le paragraphe 4 de ce chapitre.

On peut envisager d'autres calculs, par exemple inclure le correcteur éliminé dès le départ puisqu'ici, on harmonise les notes : cela change peu les résultats numériques.

[8] A titre d'illustration, l'intervalle de confiance à 95 % d'un écart-type dont l'estimation avec 8 données vaut 1 est [0,66 ; 2,04]))

[9] $\hat{s}^2$ s'écrit aussi $\sum_{i,k} e_{ik}^2 \Big/ 203$ et est l'estimateur sans biais pour la variance résiduelle, mentionné dans la note 6.