PROPOSITIONS POUR UN SYSTEME DE MODERATION PLUS EQUITABLE

Résumé: Au delà de l'habituelle technique de redressement systématique des notes sur la base de la moyenne et de l'écart-type, une nouvelle technique de modération est proposée. Cette technique permet une harmonisation variable des notes en fonction des écarts constatés à la moyenne et à l'écart-type généraux ainsi qu'au nombre de notes attribuées. L'ancienne et la nouvelle formule sont appliquées à un cas fictif.

Mots clés : docimologie, redressement, modération, notation, évaluation.

Auteur : Emmanuel Dion, enseignant-chercheur à Audencia, 8 route de la Jonelière, 44300 Nantes, edion@audencia.com

PROPOSITIONS POUR UN SYSTEME DE MODERATION PLUS EQUITABLE

En tout, de la modération.

Inscription dans le pronaos du temple de Delphes

Une égalité d'aigles et de moineaux, de colibris et de chauves-souris, qui consisterait à mettre toutes les envergures dans la même cage, et toutes les prunelles dans le même crépuscule, je n'en veux pas.

Victor Hugo

*
* *

I - EXPOSE THEORIQUE

Quand un même test, une même épreuve d'entretien ou une même dissertation sont corrigés par plusieurs examinateurs indépendants, on se retrouve fréquemment en présence de différences de notation que l'aléa seul ne peut expliquer[1] [2]. Certains correcteurs ayant tendance à sur-noter, d'autres à sous-noter, il a été naturel de mettre en place des systèmes qui redressent les notes pour les rendre compatibles avec le principe d'égalité de tous les candidats devant l'examen. Ces systèmes portent le nom de systèmes de modération.

Le système de modération le plus courant consiste à redresser les notes de chaque correcteur pour les aligner sur la moyenne et l'écart-type généraux [3]. Aucune loi ne permet cependant d'établir les conditions dans lesquelles un tel redressement doit être opéré. Par exemple, on ne sait pas à partir de quelle limite le redressement devient souhaitable.

La recherche de ces conditions permet d'approfondir le problème et de découvrir plusieurs variantes qui offrent chacune un certain nombre de caractéristiques particulières. Des quatre variantes listées ci-dessous par ordre croissant de sophistication, nous n'avons jusqu'à présent trouvé trace que des deux premières dans une littérature savante qui se révèle assez pauvre sur le sujet.

---

[1] Cette situation se retrouve dans tous les cas où il y a un grand nombre d'examinés et qu'il faut partager le travail entre plusieurs examinateurs. L'exemple le plus évident est celui du baccalauréat, mais on peut aussi penser à certains autres examens universitaires, aux entretiens d'entrée dans une grande Ecole, etc.

[2] Les différences les plus sensibles s'observent bien entendu sur les matières non scientifiques: philosophie, histoire, langues étrangères, etc.

[3] Voir par exemple Henri Piéron (Henri Piéron, 1963, p. 20-24), ou Gilbert de Landsheere (Gilbert de Landsheere, 1972, p.145-175).

1 - On ne redresse jamais

Avantage:les distributions atypiques sont convenablement traitées. Un examinateur qui évalue un lot de candidats exceptionnels (en bien ou en mal) conserve l'intégralité du contrôle de ses notes.

Inconvénient:les différences entre les examinateurs ne sont (par définition) pas du tout corrigées. Certains examinés peuvent être injustement avantagés ou désavantagés par le seul hasard qui les a fait corriger par l'un quelconque des examinateurs[4].

2 - On redresse systématiquement

Avantage: pour un candidat donné, la note obtenue ne dépend plus des biais du correcteur.

Inconvénient:En revanche, cette note se met à dépendre de la valeur des autres candidats examinés par ce correcteur. Tout se passe en fait comme si chaque correcteur disposait d'un nombre fini de points à partager entre les candidats qu'il supervise. S'il doit évaluer un bon groupe, chacun est désavantagé. S'il doit évaluer un mauvais groupe, chacun est avantagé. Le risque d'injustice est le plus fort avec des groupes de petite taille.

3 - On ne redresse qu'à partir d'un certain seuil

On peut choisir comme seuil le nombre de candidats. C'est la formule la plus simple. Par exemple on redresse les notes de tous les correcteurs ayant évalué au moins 50 candidats. On peut aussi, plus astucieusement, choisir comme seuil une probabilité donnée que la différence du correcteur à la moyenne et à l'écart-type généraux soit due à autre chose qu'à l'aléa (autrement dit, qu'elle soit "significative", pour employer le terme consacré). Cette mesure doit être faite au moyen des outils statistiques appropriés, le *t* de Student[5] pour la moyenne et la loi du χ^2 pour la variance.

Avantage: Ce système n'est excessif ni dans un sens ni dans l'autre, au contaire de ceux qui précèdent.

Inconvénient: La modération commence brutalement à partir d'un certain stade. Deux correcteurs à la notation presque semblable pourront être traités différemment pour peu qu'ils soient chacun d'un côté différent du seuil.

4 - On redresse de façon variable

---

[4] L'expérience montre qu'il est habituel de constater des différences structurelles de moyenne allant jusqu'à deux points entre le plus sévère et le moins sévère d'un groupe d'une vingtaine de correcteurs. En conséquence, on peut estimer (très approximativement, après un calcul approprié) à plusieurs milliers le nombre d'injustices commises chaque année lors d'un examen comme le baccalauréat (candidats recalés uniquement pour avoir été évalués par un jury anormalement sévère, et candidats admis uniquement pour avoir été évalués par un jury anormalement généreux).

[5] L'utilisation du *t* de Student est préférable à une comparaison directe à la loi normale non seulement parce qu'elle permet de traiter les petits échantillons, mais aussi (et surtout) parce qu'on travaillera en général avec un écart-type estimé.

Le redressement peut enfin être une fonction variable de la probabilité que la différence constatée entre les indicateurs (écart-type et moyenne) du correcteur redressé et les indicateurs généraux soit "significative".

Si un correcteur a vu peu de candidats et que sa notation est voisine de la notation générale (en moyenne et en écart-type), la (petite) différence constatée pourra être mise sur le compte de l'aléa et ses notes seront peu redressées. Dans le cas contraire, l'aléa ne pourra guère être invoqué et les notes seront fortement redressées.

Si on note α la probabilité que la différence constatée avant redressement soit due à l'aléa [6] [7] [8], et a' la même mesure faite après redressement, on peut s'interroger sur la nature de la relation qui doit unir α et α'. En soi, ce problème se révèle d'une grande richesse.

La première qualité qu'il faut demander à α', c'est d'être compris entre α et 1. La condition de supériorité à α signifie que les différences observées après redressement doivent être moins fortes, donc moins significatives, que celles observées avant. La condition d'infériorité à 1 est évidente puisque a' est une probabilité.

On peut d'ailleurs noter que ces limites, α et 1, correspondent à des cas déjà identifiés: l'absence de redressement repose sur la fonction identique où on a toujours $\alpha' = \alpha$. Et le redressement systématique est fondé sur la fonction constante ou quel que soit α, $\alpha' = 1$ (cf. figure 1). Le redressement avec seuil combine ces deux fonctions élémentaires (cf. figure 2).

[6] C'est la notation consacrée dans les tests utilisant *z*, *t* ou le χ^2.
[7] Donc qu'il s'agisse d'une différence significative au seuil de $(100 - \alpha)\%$
[8] Les tests seront toujours bilatéraux (sans que cela soit jamais rappelé dans les calculs ultérieurs), les différences pouvant évidemment surgir en plus ou en moins.

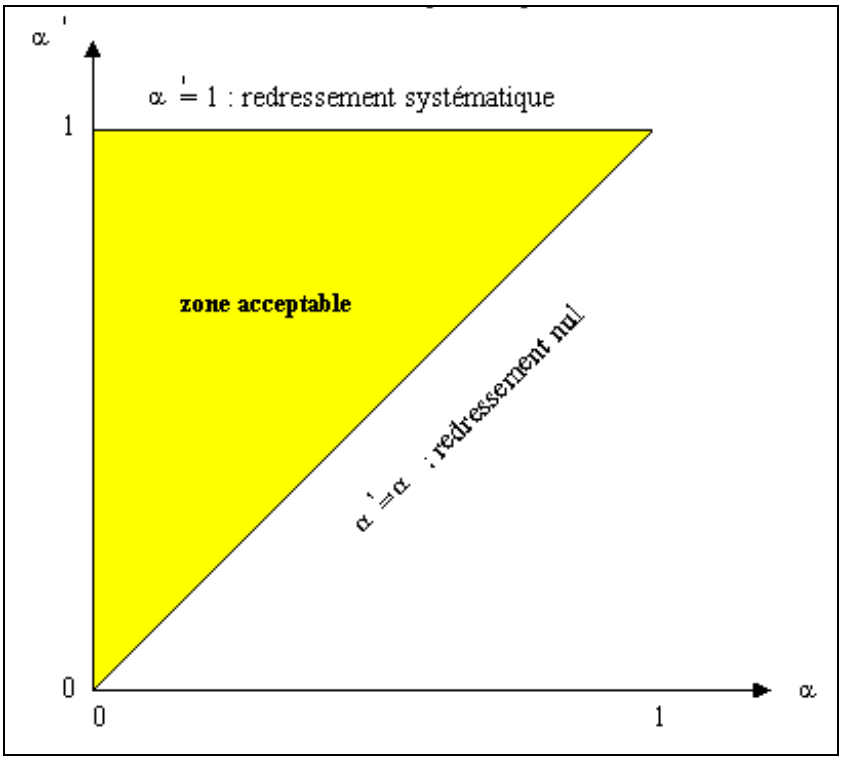

Figure 1 – Le cas des redressements simples

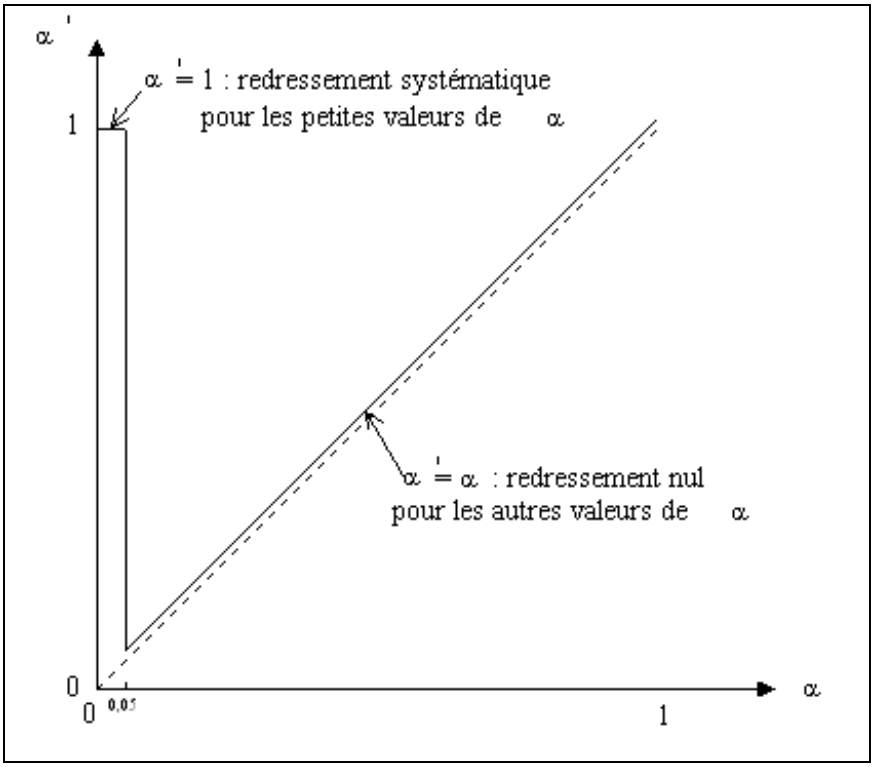

Figure 2 – Le cas des redressements avec seuil

Au delà de ces liaisons simples, voire simplistes, la façon la plus naturelle de voir les choses consiste à penser qu'il faut modifier toutes les probabilités suivant la même règle. On pourra

retenir, par exemple, la relation $\alpha'=2.\alpha$ ou bien $\alpha'=3.\alpha$ selon le degré de redressement désiré (en ayant soin de limiter α' à 1). On rend ainsi deux (ou trois) fois moins improbable chacune des différences constatées entre les correcteurs et les indicateurs généraux. Ce redressement, appelé redressement proportionnel, est illustré par la figure 3.

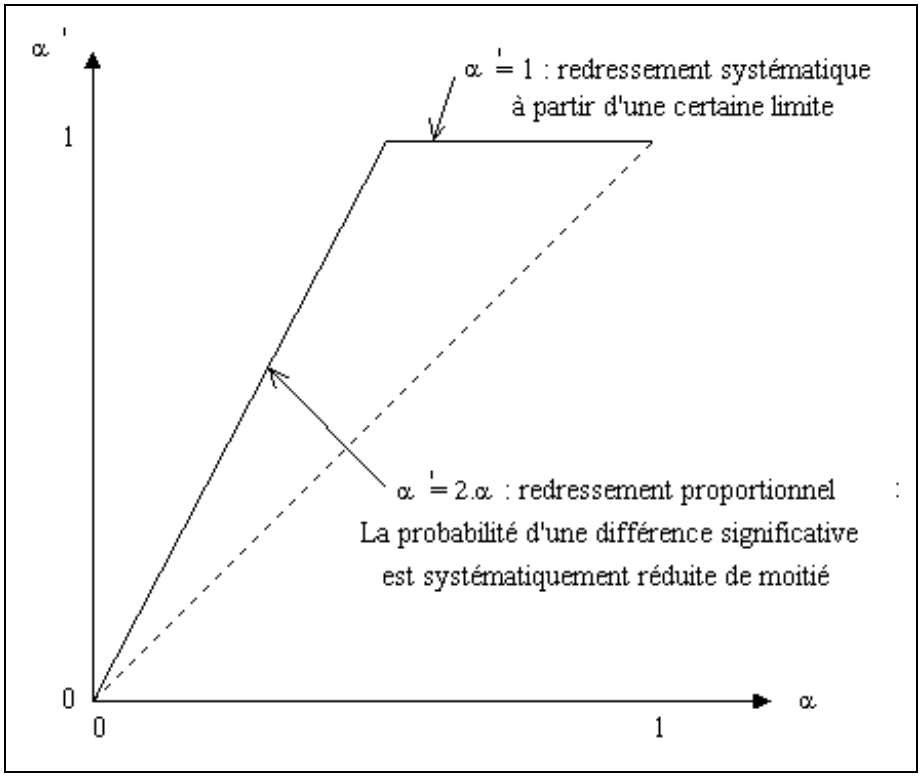

Figure 3 – Le cas des redressements proportionnels

On peut encore considérer une fonction affine du type : $\alpha' = a + (1-a).\alpha$, où a désigne un paramètre donné compris entre 0 et 1 (par exemple, $a = 0,2$). Cette fonction ne redresse fortement que les écarts les plus significatifs en les ramenant à un niveau minimum de vraisemblance (cf. figure 4).

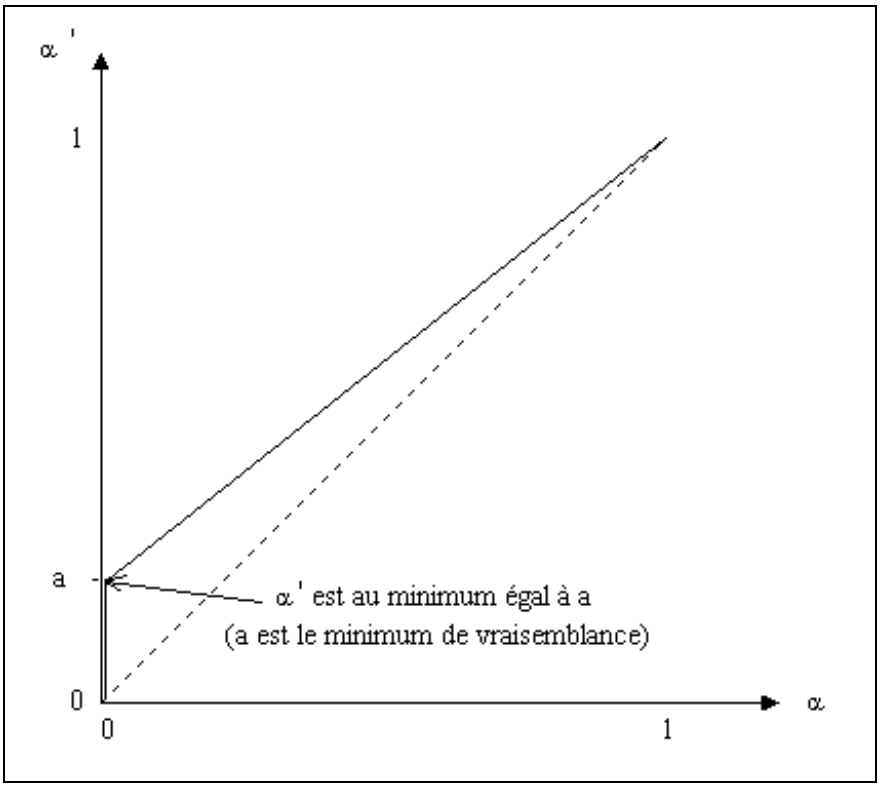

Figure 4 – Le cas des redressements avec minimum de vraisemblance

Il s'agit là d'une caractéristique intéressante. D'une manière générale, nous suggérons en effet qu'une bonne relation devrait surtout s'attacher à transformer les probabilités de différences hautement significatives (une probabilité de 1% étant ramenée à 5% par exemple), de façon à virtuellement interdire les comportements de notation trop déviants. Elle pourrait par contre ne modifier que faiblement les probabilités de différences peu significatives (une probabilité de 20%, par exemple, restant presque inchangée), ce qui, par rapport à un redressement systématique, permet de laisser un certain "jeu" à chacun des correcteurs.

La fonction suivante[9] présente les caractéristiques souhaitées:

$$\alpha' = a + (1-\alpha)a.\ \alpha b$$

où *a* et *b* sont deux paramètres qui commandent la cambrure de la courbe.

La figure 5 montre la forme de la fonction générée par cette relation. Plus *a.b* est petit et plus la courbe s'écarte de la bissectrice (plus la bosse est volumineuse, plus le redressement moyen est fort). Plus *a/b* est grand et plus la bosse se décale vers le bas (plus ce sont les probabilités les plus faibles qui sont redressées en priorité). Pour $a = b$, elle est centrée[10].

[9] Il s'agit de la fonction paramétrable la plus simple réunissant les qualités suivantes :
* $\alpha' = 0$ quand $\alpha = 0$ et $\alpha' = \alpha$ quand $\alpha = 1$
* Quel que soit α compris dans l'intervalle $]\ 0\ ;\ 1\ [$, $\alpha' > \alpha$

[10] Il faut prendre garde à ne pas utitiser un jeu de paramètres rendant par endroits la courbe décroissante. Ce serait évidemment commettre un lourd contresens.

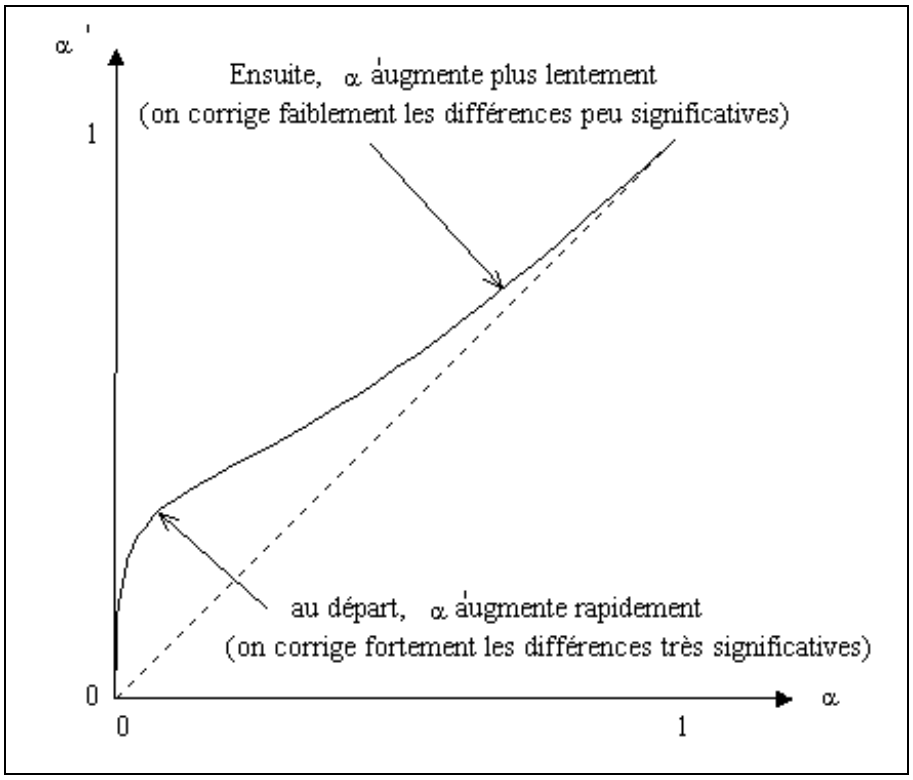

Figure 5 – Le cas des redressements variables

Un jeu de paramètres intéressant est: $a = 5$ $b = 0{,}5$. Ce jeu, défini arbitrairement, permet un redressement modéré axé surtout sur la correction des différences importantes. Il est adapté a cas d'un examen comportant quelques centaines ou quelques milliers d'examinés, et quelques dizaines de correcteurs. C'est lui que nous utiliserons dans l'exemple chiffré qui suit.

Cette fonction n'a aucune base théorique. C'est simplement parce qu'elle permet un redressement à puissance variable suivant les critères requis qu'elle a été choisie. Il faut donc souligner qu'elle ne constitue pas une solution exacte au problème posé. Si elle est supérieure aux autres systèmes de redressement, c'est surtout parce que ces derniers sont plus schématiques encore.

Malgré son caractère approximatif, elle conserve cependant un certain intérêt car on ne peut définir exactement la fonction devant lier α et α' de la façon la plus juste en l'absence d'une loi décrivant la distribution théorique de α [11], c'est-à-dire la nature de la tendance des correcteurs à s'écarter des indicateurs généraux. Cette tendance, dont l'analyse relève de la psychologie expérimentale, suit peut-être une loi normale, mais cette condition seule ne suffit pas à résoudre le problème.

[11] Cette distribution permettrait, par comparaison avec les lois susnommées, de savoir quels sont les aberrations qu'il faut redresser le plus. On pourrait d'ailleurs toujours classer les correcteurs par ordre d'éloignement aux indicateurs généraux, et redresser leurs notes de manière à ce que leur distribution après redressement corresponde à ces lois. Cette disposition ne pourrait malheureusement être retenue que dans les cas où le nombre de correcteurs est suffisamment grand pour que la reconstitution d'une distribution artificielle ne soit pas trop aliénante. La formule que nous suggérons plus haut est plus approximative, mais a le mérite de fournir une solution d'ordre général.

En fait, il nous semble qu'il faut réunir au moins quatre hypothèses pour qu'une solution exacte puisse être donnéeau problème:
- Les correcteurs doivent être nombreux (au moins 30)
- Ils doivent avoir tendance à s'écarter structurellement de la moyenne générale suivant une distribution normale
- Ils doivent cependant avoir tendance à utiliser structurellement le même écart-type
- Ils doivent avoir effectué chacun le même nombre de corrections

Si ces quatre hypothèses sont réunies, alors la variance constatée entre les moyennes des différents correcteurs (variance inter-groupe, notée V_b) peut être considérée comme la somme d'une variance d'échantillon légitime (de valeur V_w/n où V_w désigne la variance intra-groupe moyenne et n le nombre de corrections effectuées par chacun) et d'une variance parasite qu'il faut imputer à la différence structurelle de notation entre correcteurs.

Pour que cette variance parasite soit nulle, il faut transformer les moyennes de chaque correcteur (notées Mi) en moyennes redressées (notées M'_i) de façon à modifier la variance inter-groupe V_b en V'_b telle que $V'_b=V_w/n$ [12] [13].

Si on désigne par M la moyenne générale, le redressement à opérer s'obtient dès lors par la formule :

$$M'_i = M + (M_i - M) \cdot (V_w/(n.V_b))0{,}5$$

Ce cas est cependant fort rare, les quatre hypothèses évoquées étant difficiles à réunir. Aussi, la formule suggérée plus haut lui sera en général préférée, malgré son imperfection sur le plan mathématiquement, à cause de la facilité de sa mise en place.

Avantage du redressement progressif:Toutes les variables intervenant dans la probabilité que les différences observées soient dues au correcteur plutôt qu'à l'aléa sont par définition prises en compte dans la formule. On s'attaque en effet directement au coeur du problème. On dispose de plus d'un outil de redressement paramétrable qui permet de doser le redressement à volonté. On peut par exemple rendre ce dosage fonction du résultat d'une analyse de la variance (ANOVA) préalable [14]. Si cette ANOVA laisse deviner des différences importantes entre les correcteurs, on réglera les paramètres de façon à opérer un redressement important. Dans le cas contraire, on se contentera d'un redressement mineur.

Inconvénient: Cette technique de modération redresse de la même manière les notes d'un correcteur "normal" qui est tombé sur un groupe atypique et les notes d'un correcteur atypique qui est tombé sur un groupe "normal". Cependant, ce problème nous paraît impossible à résoudre par la statistique pure[15]. Seule la formation des correcteursle pourrait. Un système de modération, fût-il le meilleur, n'est pas un devin!

---

[12] Notons que la variance intra-groupe, qu'on ne redresse pas, reste inchangée par cette opération.
[13] Cette condition pourrait également s'énoncer au travers d'un F de Fisher-Snedecor qu'on cherche à égaler à 1.
[14] Mais la condition d'homoscédasticité (identité de la variance), rarement respectée, rend en général une telle ANOVA problématique.
[15] Les tests ne peuvent que constater des différences plus ou moins improbables, ils ne peuvent pas les attribuer plutôt à la qualité de groupe évalué ou plutôt au correcteur.

II - EXEMPLE CHIFFRE

Supposons que six correcteurs s'attachent à corriger une même épreuve, les copies étant partagées aléatoirement et en nombre inégal entre eux. Le tableau 1 présente les données recueillies.

i : numéro du correcteur	**1,000**	**2,000**	**3,000**	**4,000**	**5,000**	**6,000**		**Total**
ni : nombre de copies corrigées	16,000	16,000	41,000	41,000	101,000	101,000	**nt** :	**316,000**
Mi : moyenne (/10) du correcteur i	5,400	6,400	5,400	6,000	6,100	5,600	**Mt** :	**5,820**
σi : écart-type du correcteur i	1,100	1,700	1,200	1,800	1,250	1,600	**σt** :	**1,460**
N1i : note de la première copie	5,000	6,000	3,000	3,000	6,000	4,000		
N2i : note de la deuxième copie	7,000	7,000	8,000	4,000	8,000	7,000		
N3i : note de la troisième copie	4,000	4,000	5,000	7,000	5,000	6,000		
etc.								
1 Les ni ont, dans cet exemple, des valeurs artificielles qui facilitent la lecture des tables statistiques								
Attention : il s'agit de l'écart-type moyen intra-groupe								

Tableau 1 - Données brutes

Le tableau 2 présente le résultat d'un redressement systématique (les notes ainsi redressées sont notées N'_i) [16].

i : numéro du correcteur	**1,000**	**2,000**	**3,000**	**4,000**	**5,000**	**6,000**
N'1i = Mt +(σt/σi).(N1i - Mi)	5,290	5,470	2,900	3,380	5,700	4,360
N'2i = Mt +(σt/σi).(N2i - Mi)	7,940	6,330	8,980	4,190	8,040	7,090
N'3i = Mt +(σt/σi).(N3i - Mi)	3,960	3,750	5,330	6,630	4,530	6,180

Tableau 2 - Redressement systématique

Le tableau 3 présente le résultat d'un redressement avec seuil. Le seuil retenu est celui où il devient probable à plus de 95% que la différence observée entre la moyenne (resp. l'écart-type) du correcteur concerné et la moyenne (resp. l'écart-type) générale n'est pas due à l'aléa (α=0,05). Les notes ainsi redressées sont notées N''_i.

[16] Il faut toujours faire attention d'éviter qu'après le redressement, on se retrouve avec des notes en dehors de la plage de notation, par exemple inférieures à 0 ou supérieures à 10 (l'application univoque de la formule risque parfois d'y conduire).

i : numéro du correcteur	1,000	2,000	3,000	4,000	5,000	6,000
$\chi^2 i = ni \,.\, \sigma i2/\sigma t2$ ou $ni \,.\, \sigma t2/\sigma i2$	9,080	21,680	27,690	62,290	74,010	121,250
$\alpha\,(\chi^2 i)$ (se lit avec la table du $\chi 2$ à (ni-1) degrés de liberté)	0,270	0,260	0,140	0,040	0,050	0,150
Décision de redresser σi	non	non	non	oui	oui	non
σ'i (σ'i=σt ou σ'i=σi selon la décision qui précède)	1,100	1,700	1,200	1,460	1,460	1,600
ti = |Mi -Mt| . ni 0,5/ σi	1,510	1,370	2,220	0,650	2,280	1,360
α (ti) (se lit avec la table du t de Student à (ni-1) degrés de liberté)	0,160	0,200	0,030	0,520	0,020	0,180
Décision de redresser Mi	non	non	oui	non	oui	non
M'i (M'i=Mt ou M'i=Mi selon la décision qui précède)	5,400	6,400	5,820	6,000	5,820	5,600
N''1i = M'i +(σ'i/σi).(N1i - Mi)	5,000	6,000	3,420	3,570	5,700	4,000
N''2i = M'i +(σ'i/σi).(N2i - Mi)	7,000	7,000	8,420	4,380	8,040	7,000
N''3i = M'i +(σ'i/σi).(N3i - Mi)	4,000	4,000	5,420	6,810	4,530	6,000

Tableau 3 - Redressement avec seuil

Le tableau 4 présente le résultat d'un redressement variable. La progressivité retenue est celle obtenue en transformant la probabilité a en a' par l'application de la formule citée plus haut: $\alpha' = \alpha + (1-\alpha)a.\ \alpha b$. Les notes ainsi redressées sont notées N'''_i.

i : numéro du correcteur	1,000	2,000	3,000	4,000	5,000	6,000
$\alpha(\chi^2 'i) = \alpha(\chi^2 i) + \alpha(\chi^2 i)5.(1-\alpha(\chi^2 i))0,5$	0,380	0,370	0,320	0,200	0,220	0,320
$\chi^2 'i$ (se lit avec la table du $\chi 2$ à (ni-1) degrés de liberté, en correspondance de $\alpha(\chi^2 'i)$)	10,000	18,800	31,000	52,000	83,000	115,000
$\sigma''i = \sigma t \,.\, (\chi^2 'i / ni)0,5$ ou $\sigma t \,.\, (ni/\chi^2 'i)0,5$	**1,150**	**1,580**	**1,270**	**1,640**	**1,320**	**1,560**
$\alpha(t'i) = \alpha(ti) + \alpha(ti)5.(1 - \alpha(ti))0,5$	0,330	0,350	0,190	0,540	0,160	0,340
t'i (se lit avec la table de Student à (ni-1) degrés de liberté en correspondance de α(t'i))	1,020	0,960	1,350	0,620	1,460	0,980
M''i = Mt +/- t'i.σ''i / ni 0,5	**5,540**	**6,220**	**5,560**	**5,990**	**6,000**	**5,660**
N'''1i = M''i +(σ''i/σi).(N1i - Mi)	5,120	5,850	3,020	3,250	5,890	4,100
N'''2i = M''i +(σ''i/σi).(N2i - Mi)	7,210	6,780	8,310	4,160	8,010	7,020
N'''3i = M''i +(σ''i/σi).(N3i - Mi)	4,070	3,990	5,140	6,900	4,830	6,050

Tableau 3 - Redressement variable

*
* *

Nous pensons que le redressement variable permet une amélioration notable de l'équité d'une épreuve à plusieurs correcteurs. Il n'est pas très difficile à mettre en place, les correspondances entre les ti et les t'i ainsi qu'entre les $c^2 i$ et les $c^2 'i$ pouvant facilement être placées en mémoire résidente dans un logiciel adapté.

Son défaut principal réside dans son caractère encore partiellement inexact et arbitraire. S'il représente un progrès par rapport aux systèmes les plus simples, il n'est pas une panacée. Sa

seconde faiblesse tient à sa sophistication apparente. Les formules utilisées risquent de paraître proprement incompréhensibles à la plupart des examinés qui, si ils ont connaissance du redressement sans en saisir le principe, pourront avoir l'impression que leurs notes sont manipulées à tort et à travers.

Ces deux défauts sont bien réels. Mais justifient-t-ils à eux seuls que de simples différences de notation entre correcteurs, échappant à tout contrôle, créent chaque année des milliers d'erreurs d'évaluation qui sont parfois autant d'années scolaires perdues ?

Références bibliographiques

DE LANDSHEERE, Gilbert. *Evaluation continue et examens - précis de docimologie*. Paris : Nathan, 1972.

PIERON Henri. *Examens et docimologie*. Paris : PUF, 1963