Document generated on 08/11/2025 1:09 a.m.

McGill Journal of Education
Revue des sciences de l'éducation de McGill

Les dix commandements du nouvel *homo statisticus*
The New *Homo Statisticus*' Ten Commandments

Sébastien Béland, Denis Cousineau and Nathalie Loye

Volume 51, Number 2, Spring 2016

URI: https://id.erudit.org/iderudit/1038612ar
DOI: https://doi.org/10.7202/1038612ar

See table of contents

Publisher(s)

Faculty of Education, McGill University

ISSN

1916-0666 (digital)

Explore this journal

Cite this document

Béland, S., Cousineau, D. & Loye, N. (2016). Les dix commandements du nouvel *homo statisticus*. *McGill Journal of Education / Revue des sciences de l'éducation de McGill*, *51*(2), 947–960. https://doi.org/10.7202/1038612ar

Article abstract

Many discussions, sometimes heated, sometimes critical, were held on the use of statistics since the beginning of the new millennium. The aim of this opinion piece is to tap these criticisms in order to highlight ten ways to analyze a data set. We thus offer our own version of the Ten Commandments. We ultimately hope this will encourage the scientific community to improve its quantitative analysis practices.

All Rights Reserved © Faculty of Education, McGill University, 2016

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

LES DIX COMMANDEMENTS DU NOUVEL *HOMO STATISTICUS*

SÉBASTIEN BÉLAND *Université de Montréal*

DENIS COUSINEAU *Université d'Ottawa*

NATHALIE LOYE *Université de Montréal*

RÉSUMÉ. De nombreuses discussions, parfois vives, parfois critiques, ont eu lieu sur l'utilisation des statistiques depuis le début du nouveau millénaire. Ce court texte d'opinion a comme objectif de puiser à ces critiques pour faire ressortir dix bonnes façons de s'atteler à la tâche d'analyser des données. Nous vous offrons ainsi notre propre version des dix commandements. Nous espérons ultimement qu'ils encourageront la communauté scientifique à réfléchir à ses pratiques d'analyses quantitatives et à les améliorer.

THE NEW *HOMO STATISTICUS*' TEN COMMANDMENTS

ABSTRACT. Many discussions, sometimes heated, sometimes critical, were held on the use of statistics since the beginning of the new millennium. The aim of this opinion piece is to tap these criticisms in order to highlight ten ways to analyze a data set. We thus offer our own version of the Ten Commandments. We ultimately hope this will encourage the scientific community to improve its quantitative analysis practices.

Alors qu'on pourrait penser que bon nombre de pratiques statistiques sont stables depuis leur invention et constituent une valeur sûre, les récents développements dans ce domaine offrent de nombreuses critiques sur des pratiques qui sont les nôtres depuis des décennies. S'en suit une certaine confusion, car certains auteurs prônent l'abandon de ces pratiques, alors que d'autres

voudraient suppléer leurs lacunes par de nouvelles approches plus complexes (voir Trafimow et Marks [2015] pour la première alternative et Wagenmakers [2007] pour la seconde). L'impression qui en ressort peut alors être celle d'une grande désorganisation menaçante (Marmolejo-Ramos et Cousineau, sous presse). Or, voir ces développements comme quelque chose d'excitant est l'avenue que nous avons choisie.

Les statistiques et autres méthodes quantitatives sont, avec les méthodes expérimentales, l'un des deux piliers fondateurs de la méthode scientifique. Ne pas être exposé à ces méthodes peut avoir de graves conséquences. En effet, une étude troublante (Bédard, 2013) montre que les décideurs publics les moins exposés à ces méthodes lors de leurs études sont aussi les moins susceptibles de rechercher des données probantes avant d'élaborer de nouvelles politiques. Ces personnes sont aussi les plus à même de considérer des données probantes comme une opinion parmi d'autres. Cette étude touchait plusieurs domaines, incluant celui de l'éducation, nous confortant si besoin est dans notre idée qu'il est indispensable d'avoir la capacité d'être critique relativement aux pratiques statistiques quand on est un chercheur ou un décideur !

Un seul remède : au lieu d'être inquiet, le chercheur doit se muer en un *homo statisticus* moderne et profiter pleinement du potentiel accru des ordinateurs, des rapprochements entre les différentes communautés de recherche et des nouvelles techniques d'analyse offertes qui lui ouvrent tout un monde de nouvelles perspectives prometteuses et permettent de garder une attitude d'ouverture envers les données quantitatives.

Les récentes critiques ne vont pas éliminer les statistiques ou les remplacer par de nouvelles écoles de pensées. Elles ont plutôt pour but d'utiliser les outils existants avec discernement, de mettre à mal les automatismes non fondés et d'éviter d'abuser de certaines statistiques si abstraites que leur sens échappe à l'utilisateur. Pour tenter de convaincre notre lecteur, nous vous offrons, ici, différentes recommandations qui ressortent de la récente polémique. Pour les présenter, nous avons adopté une posture ludique en proposant dix commandements que le nouvel analyste avisé pourra faire siens, et que l'analyste d'expérience pourra utiliser pour dépoussiérer ses pratiques.

Nous espérons, ultimement, encourager nos collègues à améliorer leurs pratiques entourant l'examen de leurs données quantitatives et même, qui sait, à avoir autant de plaisir que nous.

1. EN STATISTIQUES, TU TE FORMERAS

Comprendre les statistiques demande un effort que certains peuvent juger considérable. On sait que nous ne naissons pas *homo statisticus*, tout comme nous ne sommes pas particulièrement enclins à prendre les décisions qui maximisent notre satisfaction personnelle (Tversky et Kahneman, 1974, 1979). Simon (1956) résume ceci en disant que la plupart des gens préfèrent être des

« satisficers » plutôt que des « optimizers ». Nous sommes des « satisficers » en procédant par essais et erreurs et en utilisant des règles approximatives très générales, ce qui permet de réagir rapidement sans investir beaucoup d'effort. De même, lorsqu'on est anxieux, fatigué ou pressurisé, on utilise plus spontanément des « règles euristiques », plutôt que de prendre le temps et l'énergie nécessaires pour résoudre un problème complexe (Chanquoy, Tricot et Sweller, 2007). Ainsi, la facilité et les habitudes gouvernent nos prises de décision et une large part de notre compréhension du monde, bien plus que la rationalité et la réflexion. Dès lors, il n'est pas étonnant que plusieurs individus éprouvent de la résistance face aux statistiques et, s'ils ont réussi à apprendre les méthodes de base, aux changements et aux innovations dans cette discipline (Sharpe, 2013).

La clé d'une meilleure compréhension des statistiques repose sur une formation de qualité (Giguère, Hélie, Cousineau, 2004 ; Sijtsma, 2015). Malheureusement, la place réservée à cette discipline est généralement fort limitée dans la formation en recherche, notablement en éducation, mais aussi dans plusieurs disciplines des sciences sociales. À l'Université de Montréal en sciences de l'éducation, par exemple, un seul cours de statistiques est offert aux étudiants inscrits aux cycles supérieurs. Hunt (2013) rapporte un problème similaire dans les programmes en psychologie.

Toutefois, il serait trop facile de se retrancher derrière le manque de cours de statistiques dans les programmes universitaires pour justifier un manque de formation. Nous vivons une époque qui donne la part belle aux apprentissages alternatifs. Ainsi il existe de nombreux réseaux parallèles permettant de bénéficier d'une formation portant sur les plus récents développements en statistiques. Le premier est sans doute la rencontre de laboratoire où l'apprenant a la chance de côtoyer des doctorants qui exposent leurs démarches quantitatives. De nombreux groupes de recherche et associations sont actifs et offrent rencontres, conférences et causeries. Les opportunités d'autoformation sont aussi devenues plus accessibles ces dernières années. Il existe, sur internet, un grand nombre d'articles ou de vidéos en accès libre qui expliquent comment réaliser une certaine analyse à l'aide d'un logiciel précis. Finalement, les réseaux sociaux offrent de nombreuses plateformes pour échanger des idées sur les statistiques. Ainsi, il n'existe plus de bonne raison pour ne pas se former dans cette discipline.

Pour accroitre la maitrise des statistiques, il faut réduire l'anxiété engendrée par ces techniques, il faut accroitre sa présence dans des réseaux parallèles pour multiplier les occasions de découvrir ces méthodes et de s'y frotter (voir Commandements 7 et 10), il faut rendre les analyses plus concrètes et plus visuelles (voir Commandements 2, 4 et 5) ; finalement, il faut cesser de se commettre par procuration comme lorsque l'on se fie aveuglément à une seule valeur *p* (voir Commandements 3, 6 et 9). Finalement, il faut éradiquer l'idée que les logiciels sont LA solution aux statistiques (voir Commandement 8).

2. L'UTILISATION DE GRAPHIQUES SIMPLES ET EFFICACES, TU AIMERAS

Des données, il peut y en avoir beaucoup, et même, trop ! Comment peut-on « faire parler ces données » ? Elles cachent des tendances, des effets, des revirements, voire des déceptions... En paraphrasant Rodin, pour dégager l'œuvre d'art qui se cache dans un bloc de marbre, il faut l'exposer de la bonne façon, avec le bon angle et le bon éclairage. Si le bloc de marbre devient les données, les graphiques permettent de trouver les bonnes expositions.

L'exploration de données à l'aide de graphiques peut aussi se comparer au travail d'un enquêteur. L'idée est de découvrir le bon point de vue afin de mettre en évidence les ressemblances et les dissemblances (Tukey, 1977). Un champ de recherche s'est d'ailleurs développé pour optimiser l'utilisation de graphiques. Wickham (2010) parle même d'une « grammaire » pour décrire les composantes d'un graphique (voir aussi Wilkinson, 2005).

Un bon graphique doit contenir juste assez d'information pour être lisible sans être rébarbatif. Il faut éviter de surcharger les graphiques, qui sont alors appelés graphiques poubelles (*chartjunk*). La Figure 1, panneau de gauche, montre un exemple qu'il faudrait éviter ; le panneau de droite conviendra mieux.

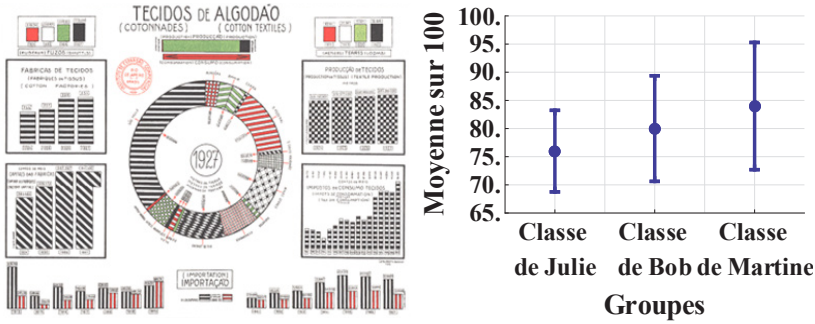

FIGURE 1 : *Deux exemples de graphiques, un exemple de graphique poubelle à gauche, et un exemple de graphique convenable à droite*[1] *(Source : davidgiard.com/2011/05/13/DataVisualizationPart6ChartJunk.aspx)*

Tufte a publié plusieurs livres sur la question des représentations graphiques (par exemple, 1983) ; il formule aussi de nombreuses recommandations : la représentation des nombres doit être proportionnelle aux quantités rapportées ; il faut des étiquettes claires et détaillées ; il est important de montrer la variation dans les données, etc. Il faut, dans tous les cas, privilégier les graphiques simples et efficaces, ce qui évitera de mal interpréter ses résultats.

3. MAL INTERPRÉTER LES RÉSULTATS, TU ÉVITERAS

Les statistiques se composent de nombreuses quantités plus abstraites les unes des autres. Même la moyenne, qui semble la plus simple des statistiques descriptives, peut se révéler en réalité fort complexe (Watier, Lamontagne et

Chartier, 2011). Et que dire de l'écart-type ? Plusieurs étudiants croient à tort qu'avec un plus gros échantillon, l'écart-type sera plus petit. Or la taille de l'écart-type dépend bien plus de la manière de contrôler les impondérables qui entourent les données ou de la nature de l'information qu'on cherche à obtenir.

Le paradoxe de Simpson

Le *paradoxe de Simpson* (1951) est intéressant pour montrer que même la moyenne peut être un casse-tête. Examinons un exemple célèbre provenant du monde du baseball : la moyenne au bâton de Derek Jeter et de David Justice, au milieu des années 1990.

La *moyenne au bâton* est calculée en divisant le nombre de coups surs produit par un joueur par le nombre de présences au bâton de ce même joueur. En 1995, David Justice, des Braves d'Atlanta, présentait une moyenne au bâton de ,253, alors que Derek Jeter, des Yankee de New York, présentait une moyenne légèrement inférieure de ,250. Comme on le voit au Tableau 1, Justice présentait encore une meilleure moyenne que Jeter en 1996.

TABLEAU 1: *Moyennes au bâton*

Joueur	1995		1996		Total	
David Justice	104/411	,253	45/140	,321	149/551	,270
Derek Jeter	12/48	,250	183/582	,314	195/630	,310

La réflexion devient intéressante lorsque nous calculons le total des deux années (voir la dernière colonne du tableau 1) : Jeter présente maintenant la meilleure moyenne au bâton. Paradoxal, non ? Le paradoxe vient du fait que les joueurs n'ont pas le même nombre de présences au bâton (les *n* sont inégaux) et, dans ce cas, il ne convient pas de moyenner les moyennes (ou alors il faut les pondérer par la taille des échantillons respectifs, 48 et 582 dans le cas de Jeter). C'est un exemple simple qui permet de voir qu'un calcul de moyenne peut ne pas faire état de la réalité et qu'il faut rester vigilant !

La valeur p

Un autre exemple de statistique très mal interprétée est la valeur *p* (Cohen, 1994). La croyance populaire veut que le *p* indique la probabilité de l'hypothèse nulle. Ainsi, obtenir un *p*, disons, de ,01 voudrait dire que l'hypothèse nulle a 1 % de chance d'être vraie, et ce serait la raison pour laquelle on la rejette. Or, cette interprétation est incorrecte. La valeur *p* donne la probabilité d'obtenir des données semblables à celles observées si l'hypothèse nulle était vraie. Une faible valeur *p* suggère donc que les données ont quelque chose d'inusité, *si on croit l'hypothèse nulle*.

Ces deux exemples montrent que l'*homo statisticus* doit se méfier des raisonnements hâtifs et des mécanismes. Il doit au contraire prendre le temps de regarder en détail ses données et ne jamais se contenter d'une statistique globale.

4. LORS DE COMPARAISONS, LES TAILLES D'EFFET TU RAPPORTERAS

Le chercheur en éducation procède régulièrement à des comparaisons : entre élèves, entre enseignants, entre écoles, etc. Plusieurs chercheurs semblent se satisfaire de rapporter uniquement la valeur *p*, qui ne fait qu'attester que la différence entre deux moyennes observées peut être ou ne pas être due au hasard. Une information plus importante et qui devrait être considérée bien avant le *p* concerne la taille de l'effet (*effect size*). Quelle est l'amplitude de la différence entre deux moyennes ?

Imaginons que nous ayons fait passer le même examen (sur 100) aux élèves des classes de Julie et de Bob (que l'on observe dans le graphique de droite de la Figure 1). La moyenne au sein de la classe de Julie est de 76 et, pour Bob, la moyenne est de 80. Pour discuter de la différence entre les deux classes, nous souhaitons calculer l'amplitude de cette différence. Le plus simple est de rapporter la différence brute entre les moyennes (ici, 4 points séparent les deux groupes). Il s'agit d'une taille d'effet *brute* (non standardisée). Il est aussi possible de standardiser la différence sur une échelle universelle (où 0 signifie aucune différence et 3 ou plus, une différence énorme). Plusieurs approches ont été développées, mais nous allons nous concentrer uniquement sur la méthode du *d* de Cohen, qui est la différence entre la moyenne *m* barre des élèves au sein des classes de Julie et de Bob, différence divisée par *s*, qui est l'écart-type de ces deux groupes-classes (ici, *s* vaut 20) :

$$d = \frac{\left(\bar{m}_{Bob} - \bar{m}_{Julie}\right)}{s} = \frac{80 - 76}{20} = 0{,}20$$

Le *d* de Cohen permet d'obtenir un coefficient qui peut être interprété à l'aide de certaines balises : la taille de l'effet est petite si $d \approx 0{,}2$, moyenne si $d \approx 0{,}5$ et grande si $d \geq 0{,}8$. Évidemment, cette nomenclature ne doit pas se substituer au bon jugement du chercheur. Cohen est lui-même sans équivoque en déclarant que ces balises sont « an operation fraught with many dangers » (Cohen, 1977). Il est effectivement difficile de mettre les tailles d'effet dans des cannettes (Baguley, 2009, p. 613).

Une taille d'effet prise isolément ne dit pas si elle est significativement différente de zéro. Il faut pour cela tenir compte de la précision de cette différence (voir le commandement suivant). Prises ensemble, taille d'effet et précision se conjuguent pour donner un test statistique. Une grande taille d'effet, une grande précision, ou les deux, sont à même de produire un *p* significatif. Pour cette raison, il vaut mieux discuter des données (la taille d'effet, la précision) et seulement ensuite procéder à un test statistique, s'il se révèle utile pour conclure l'argument.

En résumé, l'*homo statisticus* ne doit pas seulement rapporter la valeur *p*, il doit aussi rapporter les tailles de l'effet et la précision des effets. Finalement, il

doit aussi veiller à ce que ces résultats aient du sens, qu'ils soient intelligibles et lui permettent d'apprendre quelque chose de nouveau. Cela implique qu'il se tienne à jour sur ce qui est publié !

5. RAPPORTER LES INTERVALLES DE CONFIANCE, TU FERAS

Comme on l'a vu, les résultats présentés dans un graphique simple et efficace (Commandement 2) sont le point de départ de tout argument. On y repère des tendances et on tente ensuite d'estimer l'effet réel dans la population à partir de l'effet observé dans l'échantillon. La première question que l'on doit se poser concernant l'estimé[2] est : l'effet est-il important ? (voir Commandement 4). La seconde question est : l'estimé de cet effet est-il précis ? La précision d'un estimé est presque aussi importante que l'estimé lui-même, car à quoi sert une statistique qui n'a aucune précision ? Cumming et Fidler (2009) argumentent que des estimés précis permettent de mieux répondre à des questions de recherche que des valeurs *p*. Par exemple, avec le graphique de droite de la Figure 1, on peut deviner une tendance vers le haut, chose qu'un résultat classique (tel que $F(2, 15) = 3{,}21$, $p = {,}03$) n'aurait pas permis de découvrir.

La précision d'un estimé dépend de deux facteurs[3] : (i) la taille de l'échantillon : plus l'échantillon est grand, plus l'estimé a des chances d'être précis ; (ii) la variabilité des sujets dans la population. Si les sujets observés sont très variables les uns des autres, il y a des chances que l'estimé sera moins précis. Mis ensemble, ces deux facteurs peuvent s'énoncer comme suit : *la précision d'un estimé (i) diminue lorsque la variabilité observée entre les sujets devient plus grande et (ii) augmente lorsque la taille de l'échantillon devient plus grande.*

Il existe plusieurs méthodes pour chiffrer la précision. Une première est l'erreur type ; une autre, très connue, est l'intervalle de confiance à 95 % (des méthodes plus récentes incluent l'intervalle crédible à 95 % ou encore l'intervalle de vraisemblance à 8 contre 1, mais nous n'irons pas dans cette direction).

Par exemple, si vous voulez connaitre la précision d'une moyenne, vous pouvez calculer l'erreur type de la moyenne (ou en anglais, *standard error of the mean*, *SEM*). Celle-ci s'obtient avec la formule

$$SEM = {}^{s}/_{\sqrt{n}}$$

où s est l'écart type de l'échantillon et n est la taille de l'échantillon.

L'erreur type peut être utilisée pour établir une fourchette de valeurs où se trouve possiblement la moyenne de la population. Par exemple, on écrira que la moyenne du premier groupe de la Figure 1, droite, est 76 ± 4, où 4 serait l'erreur type de la moyenne, d'où une fourchette possible de 72 à 80. Cependant, cette fourchette est étroite et il y a près d'une chance sur trois qu'elle se

révèle erronée. L'intervalle de confiance à 95 % augmente cette fourchette de valeurs pour être plus précautionneux : elle n'a qu'une chance sur 20 d'être erronée. Dans l'exemple, on écrira « la moyenne dans la classe de Julie est de 76, avec un intervalle de confiance à 95 % de [68, 84] ».

De façon générale, l'intervalle de confiance à 95 % est près de deux fois plus large que l'erreur type. Une bonne manière de répondre à la question « Vos résultats sont-ils précis / fiables ? », est de donner l'intervalle de confiance à 95 %. Plus cet intervalle est étroit, plus votre résultat est précis. Plus votre résultat est précis, plus il est possible de voir les tendances qui se dégagent entre les groupes étudiés. Savoir que des groupes ne sont pas identiques est une chose, savoir de quelle façon ils se distinguent en est une autre beaucoup plus importante. Un test statistique peut répondre à la première question, alors que les intervalles de confiance permettent de visualiser la seconde.

6. T'AVEUGLER PAR LA VALEUR *P*, TU NE TE LAISSERAS PAS

Une autre statistique fréquemment rapportée est la valeur *p*. Cette valeur *p* est produite automatiquement par les logiciels et c'est trop souvent la première chose que l'on regarde, en espérant un *p* plus petit que ,05. Or, cette statistique est très trompeuse, pour plusieurs raisons :

1. Il est possible d'obtenir un *p* bien inférieur à ,05 alors que l'effet observé peut être en réalité négligeable. En effet, avec un immense échantillon, n'importe quelle différence, aussi infime soit-elle, peut devenir statistiquement significative.
2. Un effet important peut ne pas être statistiquement significatif si la précision de votre échantillon est faible. Ça ne veut pas dire qu'il ne faut pas en parler.
3. Le seuil à ,05 est entièrement arbitraire. Pourquoi un résultat ayant un *p* de ,051 devrait-il être complètement ignoré, alors qu'un autre avec un *p* de ,049 devrait-il être célébré ?

Dans les faits, la valeur *p* ne donne aucun indice à savoir si les résultats sont intéressants et d'une grandeur significative *dans la vraie vie*. La seule chose que la valeur *p* indique, c'est une réponse à la question « le hasard aurait-il pu produire ce résultat si l'hypothèse nulle était vraie ? » Elle ne donne pas la probabilité de l'hypothèse nulle, comme beaucoup le croient à tort (Cohen, 1994 ; voir Commandement 3 ci-haut).

Une autre raison de se méfier de la valeur *p* est que cette valeur est très peu reproductible. Si vous refaites une nouvelle fois la même expérience, il y a très peu de chance que le nouveau *p* ressemble au *p* précédemment obtenu (ce que Cumming [2009] appelle *la danse de la valeur* p).

Finalement, la valeur *p* donne un aperçu très limité des résultats. Décrire des résultats uniquement avec des *p*, c'est comme explorer un immeuble uniquement avec les yeux fermés. C'est beaucoup plus facile en les ouvrants !

Plutôt que de centrer son argument sur la valeur *p*, il est préférable de le centrer sur un graphique simple et efficace montrant les résultats et leurs intervalles de confiance à 95 %. Avec le graphique, il est possible de voir les tendances et les résultats qui se démarquent. Si vous décrivez bien le graphique à votre lecteur, la valeur *p* n'ajoutera sans doute que très peu à votre argument (et, pour cette raison, il est recommandé de rapporter les valeurs *p* à la fin, après avoir décrit longuement les graphiques).

Le culte du ,05 devrait être abandonné ; il n'y a aucun mal à discuter de résultats prometteurs, même si le *p* excède ,05, car les scientifiques doivent être à l'affut de résultats nouveaux (Wasserstein et Lazar, 2016). De façon plus extrême, plusieurs auteurs recommandent l'abandon complet de la valeur *p*, comme Cumming (2014) et son *New Statistics*. Dans ce sens, certaines revues interdisent même de rapporter des valeurs *p*, comme la *Basic and Applied Psychology* (Trafimow et Marks, 2015).

Pourquoi en est-on venu à donner une telle importance à la valeur *p*, alors qu'il s'agit d'une statistique très limitée ? Sans doute un critère externe (le fameux ,05) est-il rassurant, car il se commet à notre place ; mais savoir si on a un *p* inférieur à un seuil arbitraire ne constitue pas un bon argument. Un bon argument, c'est (i) un graphique simple et efficace, (ii) une explication plausible des résultats trouvés, (iii) la reproduction des résultats les plus surprenants. La seule utilité de la valeur *p* se trouve sans doute ici : si le résultat d'une certaine condition a une valeur *p* très faible, c'est cette condition qu'il faut reproduire en priorité.

7. LES PLUS RÉCENTES ÉTUDES, TU LIRAS

Comme nous l'avons déjà dit, les développements en statistiques s'accélèrent depuis quelques années. Plusieurs raisons peuvent expliquer cela : la performance accrue des ordinateurs, le rapprochement des communautés de chercheurs de différentes disciplines, l'accès à des logiciels d'analyse puissants et gratuits tels que R, etc. Pour cette raison, il est important de se tenir à jour avec les plus récentes percées méthodologiques en statistiques, de manière à maximiser ses analyses de données et, en particulier, à utiliser les meilleures statistiques et éviter les moins informatives. Cela permet d'éviter les mauvaises habitudes, ce qui peut être parfaitement illustré par l'usage de l'alpha de Cronbach.

Le cas de l'alpha de Cronbach

Le coefficient alpha de Cronbach (1951) est un cas d'étude intéressant. Laveault (2012) écrivait que « le coefficient alpha est sans doute l'une des mesures les plus répandues de la fidélité » (p. 2). Par contre, plusieurs auteurs ont déjà

soulevé plusieurs limites inhérentes à ce coefficient (Cortina, 1993 ; Dunn, Baguley et Brunsden, 2014 ; Laveault, 2012 ; Revelle et Zinbarg, 2009 ; Sijtsma, 2009). Par exemple, Dunn, Baguley et Brunsden (2014) ont démontré que le coefficient alpha est biaisé. Sijtsma (2009), de son côté, déclarait que l'alpha de Cronbach n'est pas une mesure de la consistance interne ou du degré d'unidimensionnalité d'un test. Ce chercheur va même plus loin : "the only reason to report alpha is that top journals tend to accept articles that use statistical methods that have been around for a long time such as alpha" (p. 119).

Heureusement, quelques alternatives existent. À titre d'exemple, le coefficient omega (McDonald, 1999) est vivement recommandé par Revelle et Zinbarg (2009), ainsi que par Dunn, Baguley et Brunsden (2014). Malheureusement, l'omega n'est pas disponible dans SPSS et le chercheur doit, par exemple, se tourner vers la librairie R psych pour l'utiliser. De plus, le coefficient omega est rapporté de façon marginale dans la littérature en sciences de l'éducation, même si l'on connaît sa supériorité vis-à-vis de l'alpha de Cronbach depuis plusieurs années. Cet exemple montre que s'approprier les plus récentes études en statistiques permettra de réfléchir avant d'utiliser un tel coefficient qui est, somme toute, insatisfaisant.

8. DE SPSS, TU N'ABUSERAS PAS

En statistiques, on pourrait dire que le logiciel SPSS est l'arbre qui cache la forêt. Même si ce logiciel est d'une grande utilité, il ne permet pas un accès à toutes les méthodes d'analyses de données. Rappelons, à titre d'exemple, que le coefficient omega, qui doit être préféré au coefficient alpha pour mesurer la consistance interne d'un test, n'est pas intégré dans SPSS. De plus, certaines statistiques qui se retrouvent dans ce logiciel sont parfois erronées (tel le epsilon de Huynh-Feldt, voir Dalgaard, 2007 ; Lecoutre, 1991).

Que doit-on retenir de tout cela ? D'une part, l'*homo statisticus* doit être ouvert d'esprit et être prêt à utiliser plusieurs logiciels pour trouver l'analyse qui se prêtera le mieux à ses données. Par exemple, celui-ci pourra utiliser R pour utiliser les modélisations de la théorie de la réponse aux items et utiliser MPlus pour faire des régressions multiniveaux. Pouvoir utiliser un grand nombre de logiciels nécessite de nombreux et fréquents apprentissages et réclame beaucoup d'adaptation. Ces constantes mises à jour peuvent, il faut l'avouer, avoir quelque chose de décourageant. Par contre, nul chercheur n'a besoin de se convertir en un yogi pour avoir la patience que requièrent ces exigences, car de plus en plus de sites internet et d'ouvrages de vulgarisation sont publiés pour assister l'utilisateur et parce que la communauté à laquelle il appartient peut l'aider à se mettre à jour de façon autonome.

9. VÉRIFIER LES POSTULATS DERRIÈRE UNE ANALYSE, TU FERAS

Un modèle statistique peut se comparer à une paire de lunettes. Comme le dit l'économiste Joseph Stiglitz (2003), « une idéologie fournit une lentille à travers laquelle voir le monde, un ensemble de croyances auxquelles on adhère si fort qu'il n'est presque pas besoin de confirmation empirique » (cité dans Saul, 2006, p. 9) ; nous savons qu'un modèle approprié permettra à l'analyste d'avoir une description claire des données. À l'opposé, un modèle inapproprié transmettra, au mieux, une vision floue des données ; au pire, l'analyse pourrait supporter l'explication inverse.

Les données doivent respecter certains postulats, afin d'être analysées adéquatement par certains modèles statistiques. Vérifier ces postulats permet de s'assurer que les analyses transmettront une description adéquate des données. Aussi, l'*homo statisticus* vérifiera systématiquement les postulats. Par exemple, il prendra le temps de valider la normalité des données avant de procéder à un test *t* ou à une ANOVA et il étudiera la distribution des erreurs de mesure lors d'une analyse de régression linéaire. Il n'oubliera pas que certains postulats sont plus importants que d'autres et fera preuve de jugement lorsqu'il choisira ses analyses en toute connaissance de cause. Mieux : il dira quelques mots sur ces postulats dans le cadre de ses publications et de ses communications.

10. DE LA VULGARISATION LORS DE LA DIFFUSION, TU ABUSERAS

Trop souvent, le spécialiste des méthodes quantitatives et le spécialiste de telle science humaine ou sociale sont des personnes distinctes. Borsboom (2006) soulève le fossé communicationnel existant entre ceux-ci qui détiennent l'expertise en méthodes quantitatives et ceux-là qui ont des objets de recherche axés sur une discipline. Henson, Hull et Williams (2010) caricaturaient d'ailleurs cette idée ainsi :

> We hope to see the day when education researchers no longer rely on a wizard methodologist who retires to a back room with a computer, conjures the spirits of Spearman, Fisher, Cohen, and Cattell, and emerges with a Results section for the next publication. (p. 237)

Deux voies doivent être favorisées pour maximiser la vulgarisation lors de la diffusion de résultats quantitatifs. Premièrement, il est important de rapporter d'abord des graphiques simples et efficaces en prenant bien soin de donner toutes les explications nécessaires à leur bonne compréhension. Ensuite, et seulement ensuite, les statistiques inférentielles peuvent être présentées de façon compréhensible. Deuxièmement, nous croyons qu'il faut faire un effort supplémentaire pour développer des stratégies de diffusion originales afin de capter l'attention d'un auditoire. Si cela est plus difficile à faire à l'écrit (rappelons que les règles imposées par les revues scientifiques et les maisons d'édition sont contraignantes), il nous semble important de favoriser des voies originales (par exemple, des blogues, des articles avec *chat room*, des articles généralistes, etc.) pour faire comprendre au profane le caractère parfois obscur de ces analyses.

CONCLUSION

Les articles critiques à l'endroit des statistiques telles qu'elles ont été pratiquées dans les décennies postfisherienne (de 1940 aux années 2000) sont nombreux, parfois acerbes ou caustiques. Leur nombre très important ces dernières années montre cependant la chose suivante : la communauté des chercheurs est prête à entendre ce message, et elle est prête à changer ses pratiques (Cumming, 2014). L'enseignement des statistiques est aussi sur le point de subir des changements importants, avec une réduction de la place des tests d'hypothèses nulles et une plus grande place aux statistiques descriptives, aux tailles d'effets et aux barres d'erreurs (Wasserstein et Lazar, 2016). Ces changements nous forcent à sortir de notre zone de confort forgée par les habitudes, mais, à terme, une meilleure science, plus transparente et plus innovante, prendra place.

NOTES

1. Pour reprendre les paroles de Guillaume d'Ockham : *pluralitas non est ponenda sine necessitate* (la pluralité ne doit pas être posée sans nécessité) est aussi un principe valable quand vient le temps de concevoir un graphique.
2. L'estimation est le processus par lequel on obtient un estimé, soit, une quantité numérique.
3. On peut aussi mentionner un troisième facteur : la qualité de l'instrument de mesure. En sciences de l'éducation, on considère généralement que l'erreur de mesure est négligeable par rapport à la variabilité de la population.

RÉFÉRENCES

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603-617.

Bédard, P. O. (2013) *La mobilisation des savoirs scientifiques par les analystes de politiques québécois : analyse de cheminement contrefactuelle et essai épistémologique d'interprétation causale* (Thèse de doctorat). Université Laval, Québec, QC.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440. doi:10.1007/s11336-006-1447-6

Chanquoy, L., Tricot, A. et Sweller, J. (2007). *La charge cognitive : théorie et applications*. Paris, France : Armand Collin.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY : Academic Press.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003. doi: 10.1037/0003-066X.49.12.997

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cumming, G. (2009, 3 mars). *Dance of the* p *values* [vidéo en ligne]. Repéré à https://www.youtube.com/watch?feature=player_embedded&v=ez4DgdurRPg

Cumming, G. (2014) The new statistics: Why and how, *Psychological Science, 25*, 7-29. doi: 10.1177/0956797613504966

Cumming, G. et Fidler, F. (2009) Confidence intervals: Better answers to better questions, *Journal of Psychology, 217*, 15-26. doi: 10.1027/0044-3409.217.1.15

Dalgaard, P. (2007) New functions for multivariate analysis. *R News, 7*, 2-7. Repéré à http://www.r-project.org/doc/Rnews/Rnews_2007-2.pdf

Dunn, T. J., Baguley, T. et Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399-412.

Giguère, G., Hélie, S. et Cousineau, D. (2004). Manifeste pour le retour des sciences en psychologie. *Revue québécoise de psychologie*, *25*, 117-130.

Henson, R. K., Hull, D. M. et Williams, C. S. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Educational Researcher*, *39*, 229-240. doi:10.3102/0013189X10365102

Hunt, E. (2013). Calls for replicability must go beyond motherhood and apple pie. *European Journal of Personality, 27*,126-127.

Laveault, D. (2012). Soixante ans de bons et mauvais usages du alpha de Cronbach. *Mesure et évaluation en éducation, 35*(2012), 1-7.

Lecoutre, B. (1991). A correction for the epsilon_tilde approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics*, *16*, 371-372. doi: 10.3102/10769986016004371

Marmolejo-Ramos, F. et Cousineau, D. (sous presse). Perspectives on the use of null hypothesis statistical testing (Part II): Is null hypothesis statistical testing an irregular bulk of masonry? *Educational and Psychological Measurements*.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ : Lawrence Erlbaum.

Revelle, W. et Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145-154. doi:10.1007/s11336-008-9102-z

Saul, J. (2006). *Mort de la globalisation*. Paris, France : Payot.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*, 572-582.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120. doi:10.1007/s11336-008-9101-0.

Sijtsma, K. (2015). Playing with data- or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*, 1-15.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129-138. doi:10.1037/h0042769

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238-241.

Stiglitz, J. (2003). Quand le capitalisme perd la tête. Paris, France : Payot.

Trafimow, D. et Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1-2. doi: 10.1080/01973533.2015.1012991

Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CT : Graphics Press

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, PA : Addison-Wesley.

Tversky, A. et Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Tversky, A. et Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-292.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804. doi: 10.3758/BF03194105

Wasserstein, R. L. et Lazar, N. A. (2016). ASA's statement on statistical significance and p-values. *The American Statistician*. doi: 10.1080/00031305.2016. 1154108

Watier, N., Lamontagne, C. et Chartier, S. (2011). What does the mean mean? *Journal of Statistics Education*, *19*, 1-20. Repéré à http://www.amstat.org/publications/jse/v19n2/watier.pdf

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics, 19*, 3-28.

Wilkinson, L. (2005). *The grammar of graphics* (2e éd). New York, NY : Springer.

SÉBASTIEN BÉLAND est professeur adjoint au département d'administration et fondements de l'éducation, à l'Université de Montréal. Ses intérêts de recherche suivent deux axes : la mesure dans le domaine de l'éducation et l'évaluation des apprentissages dans les programmes d'études en art, au niveau postsecondaires. sebastien.beland@umontreal.ca

DENIS COUSINEAU est professeur de l'Université d'Ottawa à l'école de psychologie et fondateur de la revue *The Quantitative Methods for Psychology* (www.tqmp.org). Ses recherches portent sur les modèles de temps de réponse, l'attention visuelle et la recherche visuelle. Il travaille aussi en mathématique sur l'estimation de paramètres pour les distributions de Weibull et de Pareto. Denis.Cousineau@uottawa.ca

NATHALIE LOYE est professeure agrégée au département d'administration et fondements de l'éducation, à l'Université de Montréal. Ses intérêts de recherche portent sur les modèles de mesure appliqués à des données issues de tests ou de questionnaires dans les domaines de l'éducation et de la santé. Elle s'intéresse particulièrement à la validité des données et des instruments. nathalie.loye@umontreal.ca

SÉBASTIEN BÉLAND is adjunct professor in the département d'administration et fondements de l'éducation at the Université de Montréal. His research interest focuses on two main axes: measurement in the field of education and postsecondary learning assessment in arts. sebastien.beland@umontreal.ca

DENIS COUSINEAU is professor at the Université d'Ottawa School of psychology and founder of *The Quantitative Methods for Psychology* (www.tqmp.org). His areas of research include response time models, attention, and visual search. He also worked in mathematics on parameter estimation for the Weibull and the Pareto distributions. Denis.Cousineau@uottawa.ca

NATHALIE LOYE is associate professor in the département d'administration et fondements de l'éducation at the Université de Montréal. Her research interests include measurement models applied to data from questionnaires or tests in the fields of education and health. She is particularly interested in the validity of data and instruments. nathalie.loye@umontreal.ca