La Différence Significative entre Valeurs p et Intervalles de Confiance

Lonni Besançon, Pierre Dragicevic

La différence significative entre valeurs p et intervalles de confiance

Lonni Besançon
Université Paris Saclay
91190, Gif-Sur-Yvette, France
lonni.besancon@gmail.com

Pierre Dragicevic
Inria, Université Paris Saclay
91190, Gif-Sur-Yvette, France
pierre.dragicevic@inria.fr

Résumé
En plus des mauvaises interprétations dont ils sont souvent à l'origine, les tests d'hypothèse apportent souvent une fausse assurance sur les résultats des communications scientifiques. Les techniques d'estimation fournissent plus d'information et se prêtent mieux à des interprétations nuancées. Nous discutons les limitations des tests d'hypothèse, puis offrons des recommandations pratiques sur l'utilisation des techniques d'estimation dans les communications scientifiques, de l'article à la présentation.

Mots Clés
Estimation; NHST; valeur p.

Abstract
In addition to the wrong interpretations it often causes, binary significance testing tends to generate a false impression of confidence in scientific publications. Estimation techniques offer more information and better lend themselves to nuanced interpretations. We discuss the limits of binary significance testing and suggest practical guidelines on how to use estimation techniques in scientific publications, from paper writing to presentation.

Author Keywords
Estimation; NHST; p-value

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction
Le test d'hypothèse nulle (NHST, Null Hypothesis Significance Testing), longtemps considéré comme la pierre angulaire des analyses statistiques, est de plus en plus critiqué dans différentes communautés scientifiques. Les nombreuses limites et faiblesses de cette approche ont été reconnues par les statisticiens et les méthodologistes [4, 7], et cette approche est maintenant critiquée au sein de la communauté IHM [12, 10, 17, 16].

Il est largement admis que les techniques d'estimation (par exemple basées sur les intervalles de confiance) apportent une solution à la plupart des limites du NHST [4, 7]. Sans pour autant bannir l'approche NHST, l'APA (American Psychological Association), une association pourtant conservatrice, recommande fortement l'utilisation de techniques d'estimation pour pallier aux limitations du NHST [25].

Bien que l'utilisation de techniques d'estimation soit encouragée dans de nombreuses communautés, leur utilisation en IHM reste sporadique. L'habitude des lecteurs et relecteurs reste de trouver des tests hypothèses et des valeurs p dans la majorité des articles qu'ils lisent. Ces procédures telles qu'on les a apprises restent profondément ancrées dans nos pratiques et pour beaucoup il peut être difficile de franchir le pas.

Dans cet article, nous rappelons les avantages principaux de l'estimation par rapport au NHST, puis offrons quelques recommandations pratiques sur son utilisation.

Pourquoi faire de l'estimation?
Le terme "estimation" est mal défini et peut prêter à confusion. Comme la plupart des méthodologistes réformistes [7, 24], nous utilisons le terme "estimation" pour référer à une pratique d'inférence statistique qui repose sur deux principes:

1. Le calcul et la communication des tailles d'effet et de leur incertitude sous forme d'**intervalles**. Il s'agit typiquement d'intervalles de confiance [7] mais il peut également s'agir d'intervalles Bayesiens [19].

2. L'interprétation **nuancée** (au lieu de dichotomique) de la force de preuve contenue dans les résultats des analyses statistiques.

Selon cette définition donc, présenter des intervalles de confiance n'est pas une condition *nécessaire* pour faire de l'estimation: d'autres approches comme l'approche Bayesienne se prêtent très bien à l'estimation. Dans cet article nous nous focalisons sur les intervalles de confiance, plus accessibles et plus faciles à calculer. D'autre part, point crucial, présenter des intervalles de confiance n'est pas une condition *suffisante* pour faire de l'estimation. Voici un tableau récapitulatif:

	Valeurs p	Intervalles
Interprétation dichotomique	NHST	NHST déguisé
Interprétation nuancée	p exact	Estimation

L'approche traditionnelle consiste à calculer des valeurs p pour faire des tests d'hypothèse à résultat binaire (en haut à gauche). Passer à l'estimation nécessite d'aller au-delà des valeurs p, mais nécessite aussi d'aller au-delà de la notion de test dichotomique. Nous commençons par rappeler les limites de la notion de test dichotomique.

Limites des tests dichotomiques
Le NHST est utilisé comme un outil retournant une réponse binaire, ce qui supprime de l'information et présente des dangers d'interprétation [10]. Les résultats se retrouvent catégorisés en tant que statistiquement significatifs ou non selon que p est inférieur ou supérieur à un seuil arbitraire généralement fixé à $\alpha = .05$. Cette vision, faussement rassurante, est simpliste: elle efface les nuances qui peuvent exister dans les données collectées. En effet, la force de preuve dans les données est fondamentalement continue. Avec le NHST, une valeur $p_1 = .052$ sera considérée comme non statistiquement significative, alors qu'une valeur $p_2 = .048$ sera considérée comme statistiquement significative. Supposons par ailleurs que nous obtenons $p_3 = .3$ (non significatif) et $p_4 = .005$ (significatif). Il n'est pas logique que le même traitement soit réservé à p_1 et p_3 d'une part, et à p_2 et p_4 d'autre part, vu que p_1 et p_2 sont de loin les deux résultats les plus similaires.

L'application aveugle d'un couperet mène à d'autres paradoxes et problèmes épistémiques, tels que les pratiques de "p-hacking" et les biais de publication [10, 2]. C'est pourquoi beaucoup de méthodologistes désireux de conserver les valeurs p recommandent d'abandonner la notion de seuil [2, 22]. Cette interprétation des valeurs p comme mesure continue de force de preuve a été dès l'origine pronée par Ronald Fisher [14]. Selon cette école de pensée, $p_1 = .052$ et $p_2 = .048$ sont des résultats indiscernables.

Limites des valeurs p
Même en conservant toute l'information dans les valeurs p sans employer aucun seuil, la quantité d'information véhiculée par p reste limitée. La raison est que p se focalise exclusivement sur l'hypothèse nulle (en général l'hypothèse d'absence d'effet ou de différence) et sa contraposée (il existe un effet ou une différence). Ainsi, p nous aide seulement à déterminer la certitude avec laquelle nous pouvons conclure qu'il existe un effet ou une différence. Utiliser les intervalles de confiance permet non seulement de véhiculer la même information, mais permet en plus de caractériser quelles magnitudes d'effet sont plausibles et quelles magnitudes sont moins plausibles. Par contre, l'approche consistant à rapporter p et la moyenne d'échantillon (sans l'intervalle) est trompeuse, car une valeur de p significative n'implique pas que cette moyenne soit précise [10].

Une analyse graphique d'intervalles de confiance peut se rapporter dans un premier temps à un raisonnement similaire à celui que l'on peut faire avec les valeurs p: plus un intervalle est éloigné de la valeur zéro, plus les résultats sont statistiquement significatifs [20]. Qui plus est, cette analyse graphique fournit également des indications sur la force possible de l'effet. En particulier, les intervalles de confiance fournissent une borne supérieure approximative à la taille probable de l'effet. Ainsi, même un résultat non significatif avec $p \gg .05$ peut être interprété: quand l'intervalle est petit on peut en conclure que l'effet est négligeable. Enfin, les intervalles de confiance sont plus parlants que la valeur numérique d'une valeur p, et ils donnent moins l'impression de précision et de certitude [10]. Alors que les valeurs p sont toujours données sous forme textuelle, les intervalles de confiance peuvent être présentées sous forme graphique afin de faciliter l'interprétation.

Bien que les intervalles de confiance soient représentés de façon binaire (une valeur se situe dans l'intervalle ou non), la démarche d'estimation impose de ne pas les interpréter de façon binaire. Par exemple, déclarer les résultats significatifs ou non-significatifs selon que l'intervalle de confiance à 95% contient zéro revient à observer si p est en-dessous ou au dessus de $\alpha = .05$. Il s'agit essentiellement de la méthode NHST (en haut à droite du tableau précédent).

Le rôle central de l'interprétation

La nature subjective des interprétations dans la démarche d'estimation est une critique courante de l'estimation et constitue une barrière importante au changement. Cette subjectivité met mal à l'aise. Elle est cependant inévitable. Nous pensons néanmoins que les chercheurs en IHM sont en bonne position pour évoluer d'une vision mécaniste de l'analyse statistique à une vision qui reconnaît l'humain.

L'Interaction Homme-Machine se concentre sur l'être humain, ses capacités et ses besoins. La recherche en IHM a pour but d'assister, d'augmenter l'être humain, jamais de le suppléer. Néanmoins, lorsqu'il s'agit d'étudier les résultats d'expériences conduites avec des êtres humains, la majorité des chercheurs utilisent encore un algorithme, un outil statistique informatique qui leur retourne une réponse binaire: les résultats sont ou non significatifs. Cette étonnante contradiction se retrouve dans un grand nombre de travaux acceptés dans les meilleures conférences d'IHM. Sans vouloir en aucune manière remettre en question ces travaux, il serait souhaitable de mieux intégrer le jugement humain dans la boucle de l'analyse des résultats.

Il existe de nombreuses définitions des statistiques. Selon Wikipédia, la statistique est "l'ensemble des méthodes qui ont pour objet la collecte, le traitement et l'*interprétation* de données d'observations relatives à un groupe d'individus ou d'unités" [1]. Selon le Larousse, l'interprétation consiste à donner un sens personnel parmi d'autre possibles [21]. L'interprétation de résultats statistiques doit donc reposer sur une analyse personnelle (et donc nécessairement subjective) des résultats par l'humain. Une telle interprétation peut se dérouler à trois niveaux:

1. Celui du/des auteur(s) lorsqu'ils discutent leurs résultats dans leur article ou leur présentation,
2. Celui des relecteurs lorsqu'ils évaluent une soumission,
3. Celui des lecteurs d'une publication, pour les aider à décider si les résultats méritent d'être réutilisés, ou pour discuter de ces résultats.

Chacun de ces "utilisateurs" possède sa propre expérience, ses propres domaines d'expertise, et ses propres objectifs. Une approche qui laisserait la place à l'interprétation humaine permettrait à chacun d'évaluer l'importance, la force, et l'impact de résultats en fonction de son contexte.

Une interprétation humaine permet également de souligner l'incertitude inhérente aux résultats expérimentaux et à leurs analyses statistiques [11, 15]. Souligner l'incertitude statistique valorise et encourage la reproduction (replication), activité d'une importance capitale dans les sciences [8]. Malgré l'importance de la reproduction, il est encore très difficile de publier des articles reproduisant des études déjà menées. Une raison potentielle est le fait que la présentation actuelle des résultats statistiques dans les articles démontre une assurance, une certitude, quant à la viabilité des résultats. Un groupe de chercheurs conduisant donc la même expérience ne fera que confirmer ou infirmer des résultats "déjà connus" et leur communication aura de fortes chances d'être rejetée. À l'inverse, une philosophie basée sur l'estimation pourrait permettre de mettre l'accent sur la nature cumulative des expériences utilisateur.

L'estimation dans la pratique

Publier un article qui utilise l'estimation sans aucune valeur p n'est pas entièrement sans risque, mais est tout à fait possible (voir http://www.aviz.badstats pour des exemples d'articles). Nous fournissons ici quelques pistes.

Justifications
L'estimation est peu répandue en IHM, les lecteurs et relecteurs étant habitués à trouver des tests hypothèses et des valeurs p. Par conséquent, afin de rassurer les lecteurs et de ne pas donner l'impression que la section des résultats manque de "rigueur statistique", il est conseillé de présenter dans un premier temps les raisons pour lesquelles les auteurs décident d'utiliser des techniques d'estimation. Un petit paragraphe avec deux ou trois références suffit (voir par exemple les paragraphes de justification dans [6, 5]). Il arrive par contre que certains relecteurs choisissent d'entièrement ignorer ce paragraphe et ses références.

Explication des intervalles de confiance
Il peut être utile d'expliquer au lecteur non averti comment lire et interpréter les intervalles de confiance. Il existe une définition exacte des intervalles de confiance, et plusieurs définitions approximatives mais utiles [7, 10]. Supposons que nous cherchions à estimer une moyenne. Un intervalle de confiance à N% est issu d'une procédure qui produit des intervalles qui capturent la "vraie" moyenne (la moyenne de la population) N% du temps lors de réplications répétées de la même expérience. Certains intervalles de confiance peuvent être également définis comme l'ensemble des valeurs qui ne sont pas rejetées comme hypothèse nulle dans un test statistique (dont le seuil est $\alpha = .05$ pour un intervalle à 95%). Ces définitions sont néanmoins difficiles à digérer dans un article.

Selon une interprétation plus intuitive dite "Bayesienne", qui offre une approximation raisonnable dans la grande majorité des cas, un intervalle de confiance indique une plage de valeurs plausibles pour la moyenne de la population [7, 10, 24]. Pour ne pas encourager une interprétation binaire, il peut être utile d'ajouter que les valeurs situées en dehors de l'intervalle ne sont pas impossibles, et que les valeurs proches de l'estimation ponctuelle (le point) sont plus plausibles que les valeurs situées près du bord.

Rapport avec les valeurs p
Pour les lecteurs habitués aux valeurs p et au repère offert par le seuil conventionnel de $\alpha = .05$, il peut être utile de rappeler rapidement et à titre indicatif les correspondances entre les intervalles de confiance et p.

Souvent, il est nécessaire d'interpréter l'écart ou le chevauchement entre deux intervalles de confiance (par exemple, entre la moyenne de la technique A et la moyenne de la technique B). Cumming [7] fournit une règle simple: pour deux échantillons indépendants (variable inter-sujets), si le chevauchement est de moins 1/3 de la longueur moyenne des deux intervalles, alors la différence est significative à $\alpha = .05$. Le danger est cependant d'encourager le lecteur à penser de façon dichotomique. Krzywinski et Altman [20] offrent une figure (reprise par Dimara et al [9] et reproduite en Figure 1) qui offre d'autres valeurs de α comme repères. Encore une fois, ces correspondances valent uniquement pour une variable inter-sujets.

Le repère $\alpha = .05$ est utile à l'interprétation vu son rôle majeur dans l'histoire des statistiques. Malgré tout, les chevauchements doivent être interprétés de manière nuancée plutôt que via des seuils stricts. L'écart ou le chevauchement entre deux intervalles de confiance nous permet de quantifier de façon continue la certitude avec laquelle on peut affirmer qu'une différence existe entre les deux moyennes. Plus ces intervalles se rapprochent et se chevauchent, plus la preuve sera faible. Un chevauchement important (Figure 1) devra mener à la conclusion que les résultats ne permettent pas de conclure (mais en aucun cas que les moyennes sont identiques ou similaires).

Si les articles ont tendance à présenter leur tests d'hypothèse

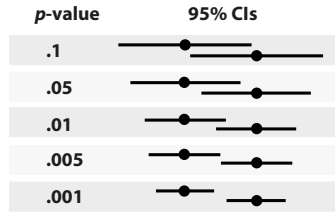

Figure 1: Correspondance entre chevauchement des CIs et la valeur p, pour des échantillons indépendants [9, 20].

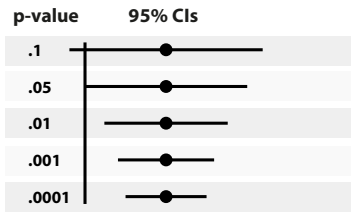

Figure 2: Correspondance entre le chevauchement d'un CI par rapport à l'hypothèse nulle et la valeur p, pour un échantillon simple.

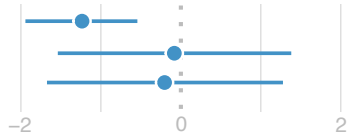

Figure 3: Exemple d'intervalles de confiance pour une différence.

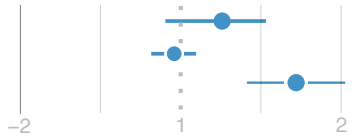

Figure 4: Exemple d'intervalles de confiance pour un quotient.

comme une vérité absolue, il s'agit avec les intervalles de confiance d'accepter l'incertitude dans nos données et de souligner le fait que certains résultats peuvent être plus certains que d'autres. Ainsi, l'utilisation de tournures de phrases telles que "La figure présente une forte preuve que la technique A est en moyenne plus rapide que la technique B, alors que la différence entre A et C est beaucoup plus incertaine" est tout à fait acceptable, pourvu que les intervalles de confiance soient présentés assez clairement pour permettre aux lecteurs de juger par eux-mêmes. L'approche NHST les prive de cette liberté, et la remplace par un sentiment trompeur d'objectivité et de rigueur scientifique.

Tailles d'effet

Si les intervalles de confiance sur les moyennes individuelles sont informatives, il est souvent conseillé d'aller plus loin en reportant les intervalles de confiance sur les tailles d'effet. Une taille d'effet permet de quantifier la question de recherche par une valeur unique, qui ne requiert pas de comparer des moyennes [7]. Par exemple, si l'objectif de l'étude est de mettre en compétition deux techniques, la taille d'effet peut être une mesure qui quantifie la différence de performance moyenne entre ces deux techniques. L'interprétation doit se concentrer sur cette taille d'effet et son intervalle de confiance (Figure 2) en complément des moyennes individuelles, surtout si la variable est intra-sujets [7].

Une taille d'effet peut être standardisée ou simple. L'utilisation de la première étant rarement indispensable voire souvent déconseillée [3], il suffira en général de calculer des tailles d'effet simples. Une taille d'effet simple peut être une différence, ou un quotient, entre deux moyennes, avec les intervalles de confiance associés. Une différence est idéalement présentée sur un graphique avec zéro pour origine (Figure 3), alors que pour un quotient, l'origine est 1 (Figure 4). Dans les deux cas, l'intervalle de confiance permettra de juger visuellement la direction probable de l'effet, ainsi que la plage des magnitudes plausibles.

Exemples d'interprétation

Revenons sur les figures 3 et 4. Supposons que la figure 3 représente la différence de précision entre deux techniques $(A - B)$, pour trois tâches différentes. Dans le premier cas, on voit que l'intervalle de confiance est clairement loin de zéro. Cela nous permet d'affirmer que l'on dispose d'une preuve forte que la technique A est plus précise que la technique B pour la première tâche. Pour les deux autres tâches, les intervalles de confiance sont à cheval sur la valeur zéro. Nous sommes donc obligés d'admettre que pour ces deux tâches, les résultats ne sont pas concluants, c'est-à-dire que nous ne pouvons pas déterminer quelle technique est plus précise que l'autre, et nous avons une forte incertitude sur la différence entre A et B. En effet, les intervalles étant larges, nous ne pouvons ni déduire que les techniques sont différentes, ni qu'elles sont comparables.

Supposons maintenant que la figure 4 représente le quotient du temps nécessaire pour accomplir les trois mêmes tâches (A/B). On peut l'interpréter ce graphique de la façon suivante. Pour la première tâche, le faible chevauchement sur la valeur 1 suggère que B est probablement plus rapide que A en moyenne, mais la preuve est très loin d'être aussi forte que, par exemple, le premier intervalle sur la figure 3. Pour la seconde tâche, les résultats sont à nouveau non concluants: on ne peut rien dire quant à la direction de l'effet. Par contre, l'intervalle est très court, et on peut donc en déduire que les techniques sont quasiment équivalentes. Cette déduction aurait été impossible avec juste une valeur p. Pour la troisième tâche, l'éloignement de l'intervalle de confiance par rapport à la valeur 1 nous permet non seulement d'affirmer avec quasi certitude que

la technique B est plus rapide que A, mais aussi que B est très probablement de 1.5 à deux fois plus rapide.

Calcul des intervalles de confiance
Calculer des intervalles de confiance est en général aisé avec des outils comme R. La méthode idéale peut varier selon le type de la variable estimée. Les tests binaires de distribution comme les tests de normalité sont peu utiles [10]: il est préférable de réfléchir sur les propriétés de la variable mesurée et/ou de visualiser leur distribution.

Une variable temporelle (par exemple, le temps de complétion d'une tâche) est nécessairement strictement positive et sa distribution va typiquement présenter une dissymétrie. Dans presque tous les cas, cette dissymétrie peut être corrigée par une transformation logarithmique [23]. Ceci fait, l'intervalle de confiance classique basé sur la distribution t (retourné par la fonction t.test dans R) peut être calculé. Les résultats (moyenne et limites de l'intervalle) sont ensuite anti-logués avant d'être présentés. En conséquence, la moyenne reportée ne sera pas arithmétique mais géométrique. Cette approche présente l'avantage supplémentaire de réduire l'influence des mesures extrêmes (il n'est donc pas nécessaire de les éliminer) [10].

D'autres variables peuvent être approximativement normalement distribuées, en particulier si leur domaine de définition couvre l'ensemble des réels. Dans ces cas, l'intervalle retourné par t.test peut être utilisé sans transformation.

Pour les variables dont la distribution s'éloigne de façon importante de la distribution normale ou dont la distribution est incertaine (par exemple, des variables bornées), une bonne solution "passe-partout" est le bootstrapping, une méthode basée sur la simulation [26, 18]. Le bootstrapping est une méthode non-déterministe mais fonctionne pour une grande variété de distributions. L'utilisation de la bibliothèque "boot" dans R permet de calculer les intervalles de confiance selon cette méthode.

Nous avons inclus un exemple d'analyse de données réelles avec le code source R dans http://www.aviz.fr/ci.

Répondre aux relecteurs
Bien que l'approche d'estimation soit louée pour ses avantages et de plus en plus utilisée, il demeure que son utilisation est parfois mal reçue par les lecteurs ou relecteurs. Si les relecteurs bien informés sur les débats méthodologiques actuels sont souvent très enthousiastes, il n'est pas rare de recevoir des relectures où l'approche d'estimation est jugée insuffisante ou pas assez rigoureuse: *"La section résultat est plutôt vague. S'il vous plaît reportez une analyse statistique de vos résultats (valeur-p)"*. Dans le cas où la conférence propose aux auteurs d'écrire un rebuttal, si la justification suggérée précédemment est déjà incluse dans l'article, l'auteur pourra insister à nouveau dessus dans le rebuttal en pointant vers les articles ou ouvrages prônant l'utilisation de techniques d'estimation.

Idéalement, une argumentation solide devrait être incluse dans chaque article pour ne pas avoir à faire face à ce problème, mais les contraintes de nombre de pages forcent à un compromis. Si la place le permet, l'utilisation du graphique de Krzywinski et Altman (Figure 1) sur la correspondance entre l'estimation et les valeurs p peut permettre de réconcilier les deux mondes. Sinon, une simple référence vers cet article en expliquant qu'une lecture utilisant la logique des tests d'hypothèses peut être faite à partir des intervalles de confiances présentés dans le papier.

Il arrive parfois que certains relecteurs demandent d'inclure les valeurs p obtenues pour chacun des résultats de la communication. Bien que cela ne soit pas particulièrement chronophage et que cela n'accroisse pas la place prise par

l'analyse des résultats sur le papier, il n'en reste pas moins que cette information est redondante, et peut rendre la lecture de l'article difficile [10]. Ces problèmes peuvent être mentionnés dans le rebuttal et l'auteur peut inclure un lien vers le graphique de Krzywinski et Altman dans la communication elle même, ou dans le rebuttal.

Malgré toutes ces précautions, un article peut se voir sévèrement attaqué pour son approche non traditionnelle d'analyse et d'interprétation des résultats, par un relecteur qui ne se laisse pas convaincre. Cependant, nous n'avons pour le moment pas encore eu de cas où l'une de nos soumissions s'est vue refusée pour cette raison uniquement.

Présentation
Depuis plusieurs années, les techniques d'estimation sont utilisées par notre équipe dans nos articles et présentations. Nous avons chacun adapté les méthodes à notre façon. Dans une présentation, pour de raisons de temps, il est difficile de communiquer tous les résultats et tous les graphiques de l'article, il est donc nécessaire de simplifier. Certains d'entre nous présentent uniquement les quelques tailles d'effet qui répondent aux questions de recherche principales. Une autre approche (adoptée par le premier auteur, suite à de nombreux retours et questions en conférences ou en réunions) consiste à présenter les moyennes individuelles. S'il est vrai que le calcul de la taille d'effet et sa représentation graphique ne sont pas spécialement complexes, cette approche demande une certaine habitude. Il y a de fortes chances qu'une partie de l'assistance ne puisse pas interpréter le graphique (qui ne sera présent à l'écran que pendant un temps prédéfini et en général court). Une simple analyse des moyennes et des intervalles de confiance permet déjà à l'assistance de se faire un avis. Le présentateur peut néanmoins tout de même parler d'effets forts ou faibles durant sa présentation, en gardant les graphiques avec taille d'effet sur des slides de secours. En effet, le moment des questions est d'une part en général plus propice à l'affichage sur une durée plus longue d'un graphique et d'autre part permet au présentateur de prendre le temps d'expliquer à son auditoire comment lire le graphique présenté. Si l'utilisation de techniques d'estimation venait à se démocratiser dans les articles IHM, il n'est pas à exclure de pouvoir par la suite inclure les tailles d'effet directement dans les présentations.

Conclusion
Nous avons présenté les avantages principaux de l'estimation en statistiques et souligné le rôle central de l'interprétation humaine dans les communications scientifiques. Nous avons proposé quelques pistes et exemples pratiques afin de faciliter l'utilisation de l'estimation dans les publications. Nous espérons que cela permettra à d'autres chercheurs et étudiants d'utiliser à leur tour une approche d'estimation dans leurs articles. Préférer l'estimation à la pensée dichotomique permettrait de souligner le fait qu'une seule étude ne suffit pas à atteindre la certitude et que des études répliquées sont nécessaires à la validation et au raffinement de résultats expérimentaux. Cela mettrait également l'accent sur l'importance de l'humain dans l'interprétation, et en particulier celui de l'utilisateur final qui est le lecteur de nos communications scientifiques. Pour citer Fisher [13]:

"Nous avons le devoir de formuler, de résumer et de communiquer nos conclusions sous une forme intelligible, en reconnaissance du droit des autres esprits libres de les utiliser pour prendre leurs propres décisions."

Remerciements
Merci à Y. Jansen pour ses commentaires sur cet article.

Bibliographie

[1] 2004. Statistique. Statistique — Wikipedia, The Free Encyclopedia. (2004). https://fr.wikipedia.org/wiki/Statistique

[2] Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. 2017. *The earth is flat (p> 0.05): Significance thresholds and the crisis of unreplicable research*. Technical Report. PeerJ Preprints.

[3] Thom Baguley. 2009. Standardized or simple effect size: What should be reported? *British Journal of Psychology* 100, 3 (Aug. 2009), 603–617. DOI: http://dx.doi.org/10.1348/000712608X377117

[4] Monya Baker. 2015. Statisticians issue warning over misuse of P values. *Nature* 531, 7593 (March 2015), 151. DOI: http://dx.doi.org/10.1038/nature.2016.19503

[5] Lonni Besançon, Mehdi Ammi, and Tobias Isenberg. 2017a. Pressure-Based Gain Factor Control for Mobile 3D Interaction using Locally-Coupled Devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1831–1842.

[6] Lonni Besançon, Paul Issartel, Mehdi Ammi, and Tobias Isenberg. 2017b. Mouse, tactile, and tangible input for 3D manipulation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4727–4740.

[7] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological Science* 25, 1 (Jan. 2014), 7–29. DOI: http://dx.doi.org/10.1177/0956797613504966

[8] Michael J Curtis and Darrell R Abernethy. 2015. Replication–why we need to publish our findings. *Pharmacology research & perspectives* 3, 4 (2015). DOI: http://dx.doi.org/10.1002/prp2.164

[9] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2017. Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5475–5484.

[10] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer International Publishing, Cham, Switzerland, Chapter 13, 291–330. DOI: http://dx.doi.org/10.1007/978-3-319-26633-6_13

[11] Pierre Dragicevic. 2017. Statistical Dances: Why no Statistical Analysis is Reliable and What to do About it. Séminaire recherche reproductible, GRICAD, Grenoble. (2017). https://tinyurl.com/gricad-dance

[12] Pierre Dragicevic, Fanny Chevalier, and Stéphane Huot. 2014. Running an HCI Experiment in Multiple Parallel Universes. In *Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, 607–618. DOI: http://dx.doi.org/10.1145/2559206.2578881

[13] Ronald Fisher. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)* (1955), 69–78.

[14] Gerd Gigerenzer. 2004. Mindless statistics. *The Journal of Socio-Economics* 33, 5 (2004), 587–606.

[15] Roger Giner-Sorolla. 2012. Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science* 7, 6 (2012), 562–571.

[16] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016a. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1081–1084.

[17] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016b. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4521–4532.

[18] Kris N Kirby and Daniel Gerlanc. 2013. BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior research methods* 45, 4 (2013), 905–927.

[19] John K Kruschke and Torrin M Liddell. 2015. The Bayesian new statistics: two historical trends converge. *SSRN Electronic Journal* (2015).

[20] Martin Krzywinski and Naomi Altman. 2013. Points of significance: error bars. *Nature methods* 10, 10 (2013), 921–922.

[21] Larousse. 2017. Interprétation. (2017). http://www.larousse.fr/dictionnaires/francais/interpréter/43813

[22] Michael J Lew. 2013. To P or not to P: On the evidential nature of P-values and their place in scientific inference. *arXiv preprint arXiv:1311.0081* (2013).

[23] Jeff Sauro and James R. Lewis. 2010. Average task times in usability tests: What to report?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 2347–2350. DOI: http://dx.doi.org/10.1145/1753326.1753679

[24] John E Hunter Frank L Schmidt, John E Hunter, L Harlow, S Mulaik, and J Steiger. 1997. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. (1997).

[25] Gary R. VandenBos. 2009. *Publication Manual of the American Psychological Association* (6th ed.). American Psychological Association, Washington, DC. http://www.apastyle.org/manual/

[26] Michael Wood. 2005. Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods* 8, 4 (2005), 454–470. DOI: http://dx.doi.org/10.1177/1094428105280059