Une raison pour ne pas abandonner les tests de signification de l'hypothèse nulle

Bruno Lecoutre[1], Jacques Poitevineau[2], Marie-Paule Lecoutre[3]

[1] ERIS, Laboratoire de Mathématiques Raphaël Salem
UMR 6085 C.N.R.S. et Université de Rouen
Avenue de l'Université, BP 12, 76801 Saint-Etienne-du-Rouvray
bruno.lecoutre@univ-rouen.fr
Internet : http ://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris
[2] ERIS, LAM/LCPE
UMR 7604, C.N.R.S., Université de Paris 6 et Ministère de la Culture
11 rue de Lourmel, 75015 Paris
poitevin@ccr.jussieu.fr
[3] ERIS, Laboratoire Psy.Co, E.A. 1780,
Université de Rouen
UFR Psychologie, Sociologie, Sciences de l'Education
76821 Mont-Saint-Aignan Cedex
marie-paule.lecoutre@univ-rouen.fr

Résumé

On montre que l'on peut directement calculer un intervalle pour un contraste entre moyennes, étant donné seulement la valeur observée du contraste et la statistique de test t ou F associé (ou encore, de manière équivalente le seuil observé correspondant : "*p-value*"). Cet intervalle peut être vu comme un intervalle de confiance *fréquentiste* ou comme un intervalle de crédibilité *bayésien* ou comme un intervalle *fiduciaire*. Cela donne aux utilisateurs des tests de signification usuels la possibilité d'une transition facile vers des pratiques statistiques plus appropriées. On met en avant les liens conceptuels entre les tests et les intervalles de confiance ou de crédibilité

1 Introduction

Un grand nombre d'articles récents ont mis en avant la nécessité de changements dans la présentation des résultats expérimentaux. Une opinion de plus en plus répandue est que des procédures inférentielles qui fournissent une information appropriée sur la taille des effets ("*effect size*") doivent être utilisées en plus ou à la place des tests de signification de l'hypothèse nulle. Ainsi, en psychologie, cela a été rendu officiel par l'"American Psychological Association Task Force on Statistical Inference", qui a proposé des "*guidelines*" pour une révision de la section statistique du manuel de l'American Psychological Association. Ces propositions préconisent l'utilisation systématique d'estimations par intervalles : "interval estimates should be given for any effect sizes involving principal outcomes" (Wilkinson *et al.*, 1999).

Par conséquent un projet salutaire devrait être de fournir aux utilisateurs des tests de signification des outils qui faciliteraient une transition "en douceur" vers les intervalles

d'estimation. Dans cette perspective un résultat étonnement simple et pourtant virtuellement ignoré est la facilité à obtenir un intervalle d'estimation pour une différence entre deux moyennes (et plus généralement pour un contraste entre moyennes) à partir du test t ou F qui lui est associé.

Un tel intervalle d'estimation ("fourchette") peut recevoir différentes justifications et interprétations. Il peut être vu aussi bien comme un intervalle de confiance *fréquentiste*, comme un intervalle de crédibilité *bayésien*, ou encore comme un intervalle *fiduciaire* (Fisher, 1990). Les discussions théoriques sur ces différents cadres d'inférence dépassent l'objectif de cet article. Nous signalerons cependant que l'opinion des auteurs est qu'une approche bayésienne avec une motivation fiduciaire est idéalement adaptée à l'analyse des données expérimentales et à la publication scientifique. Le lecteur intéressé peut se référer à Lecoutre *et al.* (2001) et Rouanet *et al.* (2000). Ici nous utiliserons l'expression "intervalle", laissant le lecteur libre de choisir son cadre de justification et d'interprétation.

2 Du rapport *F* à l'intervalle pour un contraste entre moyennes

A titre d'illustration considérons une expérience avec deux facteurs croisés *Âge* et *Traitement*, chacun à deux modalités. Les moyennes observées des quatre conditions expérimentales (avec 10 sujets pour chacune) sont respectivement 5.77 (a1,t1), 5.25 (a2,t1), 4.83 (a1,t2) et 4.71 (a2,t2). On trouve dans une revue expérimentale internationale les commentaires typiques suivants, basés sur les tests F usuels de l'analyse de variance :

> "the only significant effect is a main effect of treatment (F[1,36]=6.39, p=0.016), reflecting a substantial improvement",

et encore

> "clearly, there is no evidence ($F_{[1,36]} = 0.47$, $p = 0.50$) of an interaction".

De tels commentaires sont fréquents dans les publications expérimentales. Il est fortement suggéré à un lecteur, peu au fait de la rhétorique accompagnant l'usage des tests de signification, que l'on a démontré à la fois l'existence d'un effet important du traitement et l'absence d'effet d'interaction. Mais il n'en est rien !

La différence des moyennes observées pour les deux traitements est :

$$d = \frac{1}{2}(5.77 + 5.25) - \frac{1}{2}(4.83 + 4.71) = +0.74$$

et l'effet d'interaction peut être caractérisé par la différence des différences :

$$d = (5.77 - 4.83) - (5.25 - 4.71) = +0.40$$

Un résultat simple et général est que l'intervalle $100(1-\alpha)\%$ pour la différence vraie δ peut être déduite du rapport F (avec un et q degrés de liberté). Cet intervalle est (en supposant $d \neq 0$) :

$$\left[d - (|d|/\sqrt{F})t_{q;\frac{\alpha}{2}}\ ,\ d + (|d|/\sqrt{F})t_{q;\frac{\alpha}{2}}\right]$$

où $t_{q;\frac{\alpha}{2}}$ est le $(\frac{\alpha}{2})\%$ percentile supérieur de la distribution de Student à q degrés de liberté (rappelons que le carré de $t_{q;\frac{\alpha}{2}}$ est le $\alpha\%$ percentile supérieur de la distribution

F à un et q degrés de liberté). En outre une bonne approximation peut être directement obtenue (c'est-à-dire sans se référer à des tables statistiques) en remplaçant $t_{q;\frac{\alpha}{2}}$ par $1.96\sqrt{q/(q-2)}$, ou encore plus simplement par 2 quand q est grand.

Ce résultat met en avant la propriété fondamentale de la statistique de test F d'être un estimateur de la précision expérimentale, *conditionnellement à la valeur observée d*. De manière plus explicite d^2/F estime la variance d'erreur d'échantillonnage de d. Le même résultat s'applique aux tests t de Student usuel, en remplaçant $|d|/\sqrt{F}$ par d/t.

A partir de $t_{36;0.025} = 2.028$, nous obtenons ici les intervalles 95% [+0.15 , +1.33] pour la différence entre les deux traitements et [-0.78 , +1.58] pour l'effet d'interaction. Cela montre clairement que l'on ne peut pas conclure à la fois à un effet important du traitement et à un effet d'interaction faible, ou du moins relativement négligeable (et encore moins à l'absence d'interaction).

3 Seuils de signification et intervalles d'estimation

Les statistiques de test t ou F peuvent être calculées à partir du seuil de signification observé ("*p-value*"). Par conséquent des intervalles peuvent être déduits directement du seuil observé p (supposé connu avec une précision suffisante). Il s'ensuit que, *étant donné la valeur observée d*, le seuil p est aussi un estimateur de la précision expérimentale. D'où, intuitivement, plus le résultat est significatif (plus p est petit par rapport à α), plus δ devrait être proche de d. Il est éclairant de remarquer que l'intervalle $100(1-\alpha)\%$ peut encore être écrit comme

$$[d - d_\alpha, d + d_\alpha]$$

où $d_\alpha = (|d|/\sqrt{F})t_{q;\frac{\alpha}{2}}$ est la valeur critique (positive) de d telle que le test est déclaré significatif au seuil bilatéral α si $|d|$ est supérieur à d_α.

Comme autre illustration, considérons une étude planifiée pour tester l'efficacité d'une nouveau médicament en comparant deux groupes (nouveau médicament *vs.* placebo) de 20 sujets chacun. le nouveau médicament est considéré comme efficace (cliniquement intéressant) par les experts du domaine si la différence est supérieure à +2. Quatre cas possibles de résultats ont été construits en croisant le résultat du test t (significatif, $p = 0.001$ *vs.* non significatif, $p = 0.60$, bilatéral) et la différence observée d entre les deux moyennes (grande, $d = +4.92$ *vs.* petite, $d = +0.84$).

Les intervalles 95% pour la différence vraie δ sont donnés dans le Tableau 1. Ce tableau illustre le passage de la connaissance de d et p (ou t) à une conclusion sur la grandeur de δ (l'efficacité du nouveau médicament). A partir de ce tableau, il devient facile d'éviter des conclusions erronées basées sur des interprétations hâtives du test de signification. Les règles générales suivantes peuvent s'en déduire.

Table 1 - Intervalles 95% pour δ dans les quatre cas de résultats ($t_{38;0.025} = 2.024$, d'où $d_{0.05} = 2.024d/t$)

cas	t	p	d	$d_{0.05}$	95% *intervalle*	*conclusion*
1	+3.566	0.001	+4.92	2.79	[+2.13 , +7.71]	efficace
2	+3.566	0.001	+0.84	0.48	[+0.36 , +1.32]	inefficace
3	+0.529	0.60	+4.92	18.83	[-13.91 , +23.75]	pas de conclusion
4	+0.529	0.60	+0.84	3.22	[-2.38 , +4.06]	pas de conclusion

Cas 1 (test significatif, différence d positive et grande). De tels résultats semblent généralement très favorables aux utilisateurs des tests. Cela est justifié ici, car $d - d_{0.05}$ est supérieur à +2. Cependant il faut souligner que conclure à une différence grande (notable) nécessite certaines précautions. Le test doit être "suffisamment significatif", c'est-à-dire p suffisamment inférieur à α, afin d'impliquer une grande valeur $d - d_\alpha$. En effet, dans le cas limite où d est positive et où le test est juste significatif au seuil bilatéral α, on peut seulement conclure que δ est positive.

Cas 2 (test significatif, différence d positive et petite). Puisque $0 < d_\alpha < d$, ces conditions impliquent que d_α et $d - d_\alpha$ sont petits. De plus, dans les résultats considérés ici, $d + d_{0.05}$ également petit (inférieur à +2), de sorte que l'on peut conclure à une différence petite. Parce qu'il y a apparemment un conflit entre la différence observée petite et le résultat du test statistiquement significatif, ce cas apparaît généralement aux utilisateurs des tests comme embarrassant. Cependant il n'y a pas de paradoxe, car cela peut seulement se produire quand la précision expérimentale est "très bonne" (c'est-à-dire quand la variance d'erreur d'échantillonnage est faible). C'est donc en fait un cas privilégié. Mais, en conséquence le test est très puissant (d_α petit), de sorte que même une différence observée faible peut être statistiquement significative.

Cas 3 (test non significatif, différence d positive et grande). En règle générale, on ne peut pas tirer de conclusion inductive ferme : il est bien entendu hors de question de pouvoir conclure à une différence faible. En fait ces résultats indiquent une précision expérimentale insuffisante et par conséquent ne sont pas réellement contradictoires. (seule une différence observée très grande devrait être statistiquement significative). Cependant, beaucoup d'utilisateurs des tests trouvent ce cas embarrassant parce qu'ils ne peuvent pas généraliser la conclusion descriptive d'une différence grande.

Cas 4 (test non significatif, différence d positive et petite). Ces conditions impliquent seulement que d est plus petit que d_α. Mais cela peut correspondre aussi bien à des valeurs $d - d_\alpha$ et $d + d_\alpha$ grandes ou petites. Dans les résultats considérés ici, $d + d_{0.05}$ est nettement supérieur à +2, de sorte qu'on ne peut tirer aucune conclusion ferme. Néanmoins, comme dans le premier cas, la convergence apparente entre la différence observée et le résultat du test semble souvent favorable aux utilisateurs des tests qui tendent à conclure à tort que le médicament est inefficace.

4 Conclusion

En un sens, le seuil observé p ne peut pas être regardé comme une mesure rationnelle du degré de certitude (voir notamment Hacking, 1965 ; Spielman, 1974). Il faut aussi souligner que p *en lui-même* ne dit rien sur la grandeur de l'effet. Cependant, il apparaît que dans beaucoup de cas usuels la statistique de test, ou de manière équivalente le seuil observé p, peut être directement combiné avec une statistique descriptive pour obtenir un intervalle d'estimation. A la différence de la puissance cet intervalle est directement et facilement interprétable en termes de grandeur de l'effet. Seldmeier & Gigerenzer (1989) déploraient le peu d'utilisations faites de la puissance dans les publications expérimentales. Face aux mauvais usages des tests de signification, ils énonçaient que "given such misconceptions, the calculation of power may appear obsolete because intuitively the level of significance already seems to determine all we want to know" (page 314). Un énoncé plus pertinent

apparaît être "given such misconceptions, the calculation of power may appear obsolete because *formally* the level of significance *may determine what we want to know*". Cela confirme le constat de Goodman et Berlin (1994) que "for interpretation of observed results, the concept of power has no place" (cela ne signifie pas que la puissance ne peut pas être utile pour les calculs d'effectifs).

En particulier, en regard des mauvais usages répandus des tests de signification (voir en particulier Lecoutre *et al.*, 2003), il s'ensuit qu'un résultat "largement significatif" permet le plus souvent de généraliser le résultat descriptif. Toutefois, en fonction de la grandeur de l'effet observé, cela peut conduire aussi bien à conclure à un effet grand, moyen ou petit. Au contraire un résultat "largement non significatif" ne conduira à conclure à un effet petit que si l'effet observé est lui-même très petit. En pratique un tel résultat correspondra le plus souvent à une constat d'ignorance.

En conclusion, même si bannir les tests de signification des publications expérimentales serait sans aucun doute une thérapie de choc (voir Shrout, 1997), les statistiques t, les rapports F et les seuils observés p demeureraient utiles, au moins pour les calculs des intervalles d'estimation et pour les réanalyses de résultats précédemment publiés. Ironiquement, les fournir avec une précision suffisante apparaît alors être une bonne pratique en vue des analyses ultérieures sur les tailles des effets.

Références

- Fisher R. A. (1990) – *Statistical Methods, Experimental Design, and Scientific Inference* (Re-issue). Oxford : Oxford University Press.
- Goodman, S.N. & Berlin, J.A. (1994) – The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, **121**, 200–206.
- Hacking, I. (1965) – *The Logic of Statistical Inference*. Cambridge, England : Cambridge University Press.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2001) – Uses, abuses and misuses of significance tests in the scientific community : won't the Bayesian choice be unavoidable ? *International Statistical Review*, **69**, 399-418.

Lecoutre M.-P., Poitevineau J., & Lecoutre B. (2003) – Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology* **38**, 37-45.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., & Le Roux, B. (2000) – *New Ways in Statistical Methodology : From Significance Tests to Bayesian Inference*, 2^nd edition. Bern, CH : Peter Lang.

Spielman, S. (1974) – The Logic of Tests of Significance. *Philosophy of Science*, **41**, 211–226.

Seldmeier, P. & Gigerenzer, G. (1989) – Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, **105**, 309–316.

Shrout, P.E. (1997) – Should significance tests be banned ? Introduction to a special section exploring the pros and cons. *Psychological Science*, **8**, 1–2.

Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999) – Statistical Methods in Psychology Journals : Guidelines and Explanations. *American Psychologist*, **54**, 594–604.