De la mauvaise influence des étoiles : *Khi² versus* taille des effets

Denis CORROYER

Laboratoire de Psychologie Environnementale (LPEnv)

Université Paris 5 - CNRS (UMR 8069)

Objectifs de l'exposé

- Les problèmes posés par une méthodologie centrée sur les tests statistiques
  - Ces tests ne répondent pas à la question de la taille des effets
  - *e.g.* pour un tableau de contingence, le test du *Khi*² n'évalue en aucun cas la taille des effets.
- Quelques propositions pour une autre méthodologie
  - Quels indices permettent de mesurer la taille des effets ?
  - Quelles procédures inférentielles permettent d'évaluer la taille des effets dans la population ?
- En conclusion
  - Comment présenter les résultats autrement ?
  - Pourquoi de telles résistances au changement ?

Exemple : Dossier CIVILITE

Moser, Corroyer (2002)

- Étude d'un comportement de Civilité :
  retient ou non la porte à la personne qui suit,
  à l'entrée d'un grand magasin ?
- Existe-t-il des différences de comportement selon :
  - le sexe du client qui ouvre la porte ?
  - le sexe de la personne qui suit ?
  - la densité de clients ?
  - la ville : Province (Nantes) / Paris ?

Analyse de tableaux de contingence 2 x 2

	Retient	la porte
	Oui	Non
Homme	222	144
Femme	254	260

- Distribution d'effectifs
- Hypothèse : les femmes sont plus civiles que les hommes
- Une méthodologie traditionnelle centrée sur les tests : Test du *Khi*² (d'indépendance)
  - $p > .05$: "La différence (entre H et F) n'est pas significative"
  - $p < .05$: "La différence est significative"
- Ici, test Significatif ($p = .01$)

La « critique des tests »

*Khi*² (χ^2), *T* de Student, *F* de Fisher-Snedecor

- Critique presque aussi ancienne que les tests eux-mêmes
  - Boring (1919), Selvin (1957)…
  - En France : Reuchlin (1962), Rouanet et Lépine (1975).
  - Tukey, Cohen, Rosenthal, Guttman, Yates…
- NHST : Null Hypothesis Statistical Testing
  - Seuil .05 arbitraire
  - Probabilités contre-intuitives
  - Pratiques floues (entre Fisher et Neyman-Pearson)
  - Mal interprétés (Student, 1908)
  - Ne répondent pas à la question de la taille des effets

Critiquer les tests ou critiquer les usages des tests?

- Les NHST constituent d'excellents outils : Ne pas généraliser sans précaution, un résultat observé sur un échantillon à l'ensemble de la population !
- Les NHST sont mal employés et/ou mal interprétés

  Les tests sont utilisés pour répondre à des questions auxquelles ils ne peuvent répondre
- La fourchette est-elle un mauvais outil parce que quelques uns ont essayé, en vain, de l'utiliser pour déguster un potage ?

En psychologie, des avancées…

- 1996: « Task Force on Statistical Inference » de l'American Psychological Association (APA)
- Rouanet (1996)
- Wilkinson & the Task Force on Statistical Inference (1999).
- Publication Manual de l'APA (5th edition, 2001)
- "Mentionner la taille des effets"
  - Taille des effets / Ampleur / Intensité / Grandeur / Force
  - Effect size (ES)

Faut-il s'intéresser à la taille des effets ?

Simple option / Nécessité incontournable ?

- Peut-on se contenter de conclure "Les hommes (en général) retiennent plus souvent la porte que les femmes" sans se préoccuper de l'ampleur de cette différence parente?
- Évidemment on a envie de considérer l'ampleur des différences
(75% contre 30% ou 75% contre 72%?)

	Oui	Non	
Homme	75 %	25 %	100 %
Femme	30 %	70 %	100 %

	Oui	Non	
Homme	75 %	25 %	100 %
Femme	72 %	28 %	100 %

	Oui	Non	
Homme	75 %	25 %	100 %
Femme	30 %	70 %	100 %

	Oui	Non	
Homme	75 %	25 %	100 %
Femme	72 %	28 %	100 %

- Difficile de soutenir que l'ampleur des différences importe peu, et de considérer ces répartitions comme équivalentes.
- Pourtant…
- … les pratiques usuelles qui s'appuient sur le test du *Khi*², reviennent à ne pas s'intéresser à l'ampleur des différences.

Le test du *Khi*²

- Ni la valeur du *Khi*^{2}, ni le seuil *p,*
  ne constituent des indicateurs de la taille de l'effet (ni de l'effet observé ni de l'effet parent).
- Rappel :
  - « Significatif » ne veut pas dire effet important
  - « Non significatif » ne veut pas dire effet nul, ni effet faible, contrairement à l'interprétation la plus fréquente d'un test non significatif

Mais à quoi sert le *Khi*² ?

- Sur un tableau 2 x 2, le *Khi*² permet seulement de conclure, s'il est significatif, que la différence parente est de même sens que la différence observée
- Sur un tableau *J* x *K*, le *Khi*² permet seulement de conclure, s'il est significatif, que dans la population,
  il existe un lien (non nul, de grandeur inconnue)
  entre les deux variables.

Le *Khi*² ne nous renseigne pas sur l'ampleur de l'effet observé (dans l'échantillon)

On peut obtenir un *Khi*² significatif :

- certes avec un effet observé important

- mais aussi avec un effet observé faible

On peut obtenir un *Khi*² non significatif :

- certes avec un effet observé faible

- mais aussi avec un effet observé important

Effet observé important

	Oui	Non	
g1	75%	25%	100 %
g2	33%	67%	100 %

	Oui	Non
g1	12	4
g2	4	8

	Oui	Non
g1	6	2
g2	2	4

$Khi^2 = 4.86 \quad p = .03$ (S)

$Khi^2 = 2.43 \quad p = .12$ (NS)

Effet observé faible

	Oui	Non	
g1	62%	38%	100 %
g2	66%	34%	100 %

	Oui	Non
g1	620	380
g2	330	170

	Oui	Non
g1	1240	760
g2	660	340

$Khi^2 = 2.30 \quad p = .13$ (NS)

$Khi^2 = 4.59 \quad p = .03$ (S)

Attention !

"Le *Khi*² ne nous renseigne pas
sur l'ampleur de l'effet observé"

Ne veut pas dire

"Le *Khi*² ne dépend pas de l'effet observé"

On a plus de chances d'obtenir un test significatif
avec un effet observé important

Pourquoi le *Khi*² n'est pas un bon indicateur de la taille des effets ?

Sa valeur dépend de deux éléments :

- la taille de l'effet observé
- la taille de l'échantillon (*n*)

*Khi*² = Taille de l'effet observé x *n*

Un bon indice doit être indépendant de l'effectif

- Khi^2 = Taille de l'effet observé x n
- Taille de l'effet observé = Khi^2 / n
  = $Phi^2 = \Phi^2$
  = *Carré moyen de contingence*
  cf. Pearson (19.. ?)
- Critère :
  - Si on double l'effectif, la valeur de l'indice ne doit pas changer.
  - En doublant les effectifs d'un tableau de contingence, on change le Khi^2, on ne change pas le Phi^2, ni les pourcentages en ligne…

Le *Khi*² et le prix des tomates

- Prendre le *Khi*² comme indice de taille d'effet

  équivaut à prendre le prix payé comme indice de prix,

  (quelle que soit la quantité)

  et non le prix au kg

  (prix à quantité constante)

- Prix payé = Prix/kg x poids
- Prix/kg = Prix payé / poids

Pour toute statistique de test…

Statistique de test = Taille de l'effet observé × *f* (n)

- Tableau de contingence

$$Khi^2 = Phi^2 \times n$$

- Comparaison de deux groupes indépendants

$$t = EC \times \sqrt{\frac{n}{2}}$$

- Régression

$$F = \frac{Variance\ prédite}{Variance\ résiduelle} \times (n-2)$$

- Quel(s) indice(s) de taille d’effet ?

Tableau de contingence *2* × *2*

- *Phi*²
- *Rphi*= $r\phi$ = coefficient point tétrachorique
- Différence des pourcentages
- Rapport des pourcentages
- Écart relatif entre les pourcentages
- Odds ratio
- PEM
- (…)

Tableaux de contingence $J \times K$

- Écart global à l'indépendance
  - Phi^2

    Phi^2max = minimum (J, K) – 1
  - V^2 de Cramér = Phi^2 / Phi^2max

    Varie entre 0 et 1

    (…)
- Écarts locaux (par case) à l'indépendance
  - Écarts à l'indépendance (écarts bruts),
  - Taux de liaison (écarts relatifs),
  - Densités
  - Contributions au Phi^2
  - PEM
  - (…)

Le choix des indices

- « Quel est le meilleur indice ? »
- Y a-t-il un meilleur indice ? Faut-il choisir ?
- Souvent intéressant de calculer plusieurs indices.
  - Plusieurs points de vue sur les données
  - *e.g.* les revenus des français : revenu moyen, revenu médian ou revenu le plus fréquent (mode)

Des valeurs-repères

- J'ai obtenu un V^2 de Cramér de 0.30, un taux de liaison de 20 % est-ce peu ? est-ce beaucoup ?
- Qu'est-ce qu'un effet…
  - Faible / petit / négligeable / *small* ?
  - Fort / grand / notable / *large* ?
- Des valeurs-repères indicatives

  (cf. Cohen, Corroyer & Rouanet) :
  - Pour les indices de type corrélation (*e.g.* r_{BP}, *rPhi*…)
    - .20 et .40
  - Pour les indices de type part de variance (*e.g.* V^2 de Cramer)
    - .04 (4 %) et .16 (16 %)

- Comment évaluer l'ampleur des différences parentes?

Les deux cadres inférentiels

- L'inférence bayésienne
  - S'applique à une plus grande variétés de situations
  - Interprétations plus naturelles
  - (cf. J.-M. Bernard)
- L'inférence fréquentiste (*e.g.* les tests classiques)
  - Il existe des solutions pour évaluer la taille des effets parents, dans le cadre fréquentiste classique.
  - Les intervalles de confiance.

L'intervalle de confiance (IC)

Une procédure sous-employée

Tableau de contingence 2 x 2

- L'intervalle de confiance donne une estimation de la taille de l'effet dans la population (effet parent).
- Quel que soit l'indice utilisé pour mesurer l'ampleur de l'effet observé
  (différence de pourcentage, rapport de pourcentages, odds ratio…)
  calculer systématiquement un IC sur l'effet parent.
  - Différence des % → IC sur la différence parente
  - Rapport des % → IC sur le rapport parent

Différence selon la densité à l'entrée du magasin ?

	Oui	Non
Faible	280	160
Forte	196	244

	Oui	Non	
Faible	64%	36%	100 %
Forte	45%	55%	100 %

- Analyse et conclusion avec un *Khi*² classique
  - Dans l'échantillon : plus de comportements civils lorsque la densité est faible
  - *Khi*² significatif ($p < .001$)
  - Dans la population, plus de comportements civils lorsque la densité est faible (au seuil .001)

Ampleur de la différence

	Oui	Non	
Faible	64%	36%	100 %
Forte	45%	55%	100 %

- Ampleur de la différence observée, *Dobs* :
  - *Dobs* = +19 pts
- Ampleur de la différence parente, *Dpar* ?
  - IC (.05) = [+13 pts … +26 pts]
  - Limite inférieure = 13 pts : la différence est d'au moins 13 pts
    - IC = Ensemble des valeurs "possibles" de la différence parente (au seuil .05) .
    - valeur 0 extérieure à l'intervalle ? la différence parente est de même sens que la différence observée (même conclusion que le *Khi*²).

Différence selon le sexe ?

	Oui	Non
Homme	222	144
Femme	254	260

	Oui	Non
Homme	61%	39%
Femme	49%	51%

- Analyse et conclusion avec un Khi² classique
  - Dans l'échantillon, les hommes sont plus civils que les femmes
  - Test *Khi*² significatif ($p = .001$)
  - Dans la population, les hommes sont plus civils que les femmes (au seuil .001)

Ampleur de la différence ?

- Ampleur de la différence observée, *Dobs ?*
  - *Dobs* = +12 pts
- Ampleur de la différence parente, *Dpar ?*
  - IC = [+5 pts … +18 pts]
  - Limite inférieure = 5 pts
  - Si on considère que 5 pts n'est pas une grande différence (cf. valeurs-repères),
    « On ne peut pas conclure à une différence parente importante ($IC_{.05}$ = [+5 pts … +18 pts]) ».

Différence selon le sexe de la personne qui suit

	Oui	Non
H	56%	44%
F	53%	47%

- Test du *Khi*² non significatif
  - On ne peut pas conclure à l'existence de différence parente, selon que la personne qui suit est H ou F.
  - Attention: Ne veut pas dire absence de différence parente.
- *Dobs* = 3 pts
- IC = [- 5 pts … +10 pts]
  - Inclut des valeurs de *Dpar* autres que 0
  - Inclut des valeurs non négligeables (10 pts)
  - Test NS ne veut pas dire différence parente nulle ni différence parente négligeable.

Des intervalles de confiance pour beaucoup d'indices

- Données multivariées numériques
  - Corrélations, Corrélations partielles, Coefficients de régression (…)
- Données expérimentales
  - Différence entre 2 groupes indépendants, Effet d'interaction,
  - Différence entre 2 groupes appariés,
  - Tendance linéaire, quadratique (…)

La présentation des résultats

- Pas de "différence significative"
  - C'est le test qui est significatif, pas la différence
- Utiliser systématiquement les indices mesurant la taille de l'effet.
  - *Phi*², *V*² de Cramér, Contributions au *Phi*²…
- Plus radicalement :
  - Ne plus utiliser les tests classiques ?
    - Pas de *Khi*², de *T* de Student, de *F*...
    - Plus de "différence significative"
  - Mentionner systématiquement l'intervalle de confiance,
    - IC (.05) indique ce que serait le résultat du test (Selon que 0 est ou non à l'intérieur de l'IC)
    - IC informe sur la taille de l'effet parent.
- Cf. Rothman (1978) éditeur de la *New England Journal of Medicine*

Les résistances aux changements

Le poids du nombre…

- Le nombre d'années
  - Un siècle d'analyses centrées sur les tests
- Le nombre d'utilisateurs
  - Des millions (?) d'utilisateurs
- Le nombre de tests
  - Des milliards de tests effectués depuis un siècle
- "A way of thinking that has survived decades of ferocious attacks is likely to have some value."

  [un reviewer anonyme, *in* Frick, 1996]

Le poids des habitudes…

- "We need statistical thinking, not rituals." (Gigerenzer, 1998)
- On continue à faire avec les outils informatiques ce que l'on faisait "à la main" il y a 20 ans.
- Depuis un siècle, en psychologie (en SHS ?),
  - recherche scientifique = test
  - résultat intéressant = test significatif
- Une (petite) révolution copernicienne à venir pour les pratiques statistiques

Quelques citations…

- "[NHST] has not only failed to support the advance of psychology as a science but also has seriously impeded it." (Cohen, 1994)
- "The earth is round ($p < .05$)" (Cohen)
- Autres citations sur :

http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.html

Le poids du nombre (bis)

- "C'est pas parce qu'ils sont nombreux à avoir tort qu'ils ont raison"

Michel COLUCCI