Les statistiques et les fours à chaleur tournante: un outil plutôt qu'une finalité en soi

Markus Brauer [1]

Des conversations informelles avec de nombreux collègues de partout dans le monde révèlent l'existence d'une certaine insatisfaction concernant les connaissances en statistiques des étudiants de psychologie sociale à la fin de leurs études. En effet, la plupart d'entre eux racontent des anecdotes où ils avaient affaire à un-e étudiant-e en quatrième ou cinquième année qui ne savait pas qu'une analyse de corrélation donne le même résultat qu'une analyse de régression simple, que la valeur F est égale à la valeur t élevée au carré ou qu'une ANOVA 2 X 2 implique trois comparaisons pertinentes (les deux effets principaux et l'effet d'interaction). Au premier abord, on pourrait penser que ce manque de connaissances est dû à un volume trop faible d'enseignement en statistiques. Mais une inspection des curricula révèle que ce n'est pas le cas. La plupart des étudiants en psychologie sociale suivent de nombreux cours en statistiques.

Qu'est-ce qui explique alors ce manque de connaissances ? L'hypothèse que je voudrais avancer dans cet article est qu'une des causes de ce phénomène est l'orientation de l'enseignement censé transmettre ces connaissances. Ce que nous enseignons à nos étudiants ce sont les *statistique*s. Ce que nous devrions leur enseigner c'est *l'analyse des données*. Cette distinction peut paraître artificielle mais elle reflète une attitude sous-jacente envers l'objectif de l'enseignement. Les statistiques sont une « branche des mathématiques appliquées qui a pour objet l'étude des phénomènes mettant en jeu un grand nombre d'éléments » *(Le Dictionnaire de Notre Temps, Hachette 1989)*. L'analyse des données c'est ce que nous faisons en tant que psychologues: après avoir réalisé une étude visant à tester une certaine hypothèse, nous souhaitons interroger nos données pour savoir si l'hypothèse est confirmée ou non. Dans le cas des statistiques, l'enseignement est focalisé sur les bases théoriques des différents tests statistiques et la dérivation des formules. Dans le cas de l'analyse des données, on enseignera aux étudiants comment ces tests statistiques nous permettent d'atteindre notre objectif, c'est-à-dire comment ils nous permettent d'obtenir une réponse à nos questions théoriques.

Pour illustrer ce point, prenons la métaphore d'une personne qui aime cuisinier (comme l'auteur). L'objectif de cette personne est de préparer des plats succulents pour sa famille et ses amis. Pour atteindre cet objectif, elle devra se servir de certains outils dans sa cuisine tels que la balance (électrique, dans le cas présent), le robot (350 tours par minute), ou le four (à chaleur tournante, bien sûr). Pour n'en prendre qu'un exemple il est donc indispensable que cette personne sache se servir du four. « Se servir du four » cela ne veut pas dire connaître les processus thermiques et électriques qui se déroulent à l'intérieur du four, savoir à quoi servent les différents câbles à l'arrière du four, ou connaître par cœur la formule mathématique décrivant la propagation d'air chaud à l'intérieur d'un espace fermé. En revanche, « se servir du four » cela veut dire savoir choisir les différentes cuissons selon les plats à cuisiner. Pour les brioches c'est la « cuisson fournil », pour les tartes c'est la « chaleur tournante », pour le rôti c'est la

[1] Université Blaise Pascal, Clermont-Ferrand.

cuisson « gril », et ainsi de suite. Si la personne ne sait pas quelle cuisson va avec quel plat, elle ne réussira jamais à préparer des plats succulents. Après tout, une brioche non levée ne sera jamais très bonne, même avec un mélange extraordinaire d'ingrédients (tout comme une expérience psychologique exemplaire ne sert à rien si le chercheur ne sait pas analyser les données).

Supposons maintenant que ce même amateur-cuisinier développe une telle passion pour son passe-temps favori qu'il souhaite en faire son occupation principale. Peut-être même avec l'espoir, pourquoi pas, de devenir un jour un grand chef. L'objectif devient alors de préparer des plats d'une qualité extraordinaire lui permettant de décrocher des étoiles dans le guide Michelin. Après avoir ouvert un restaurant, cette personne travaille à nouveau dans sa cuisine et doit se servir à nouveau de son four à chaleur tournante. Alors que le choix des différentes cuissons restera sa préoccupation principale, il peut lui être utile de connaître des détails techniques sur le four et la façon dont il fonctionne. Par exemple, si elle sait que la cuisson « gril » implique le fonctionnement de la résistance de voûte et que cette résistance ne produit pas de chaleur mais émet des rayons infrarouges, elle peut être amenée à tantôt fermer la porte du four pour que les plats cuisent en profondeur par conduction ou, au contraire, laisser la porte entrouverte pour que les infrarouges puissent saisir la surface de l'aliment (très important pour les crèmes brûlées!). En connaissant le taux de rayons infrarouges reflétés par les différents aliments elle peut déterminer exactement à quel point l'aliment cuit en profondeur. Notons que ces connaissances sur les processus physiques, mécaniques et techniques du four peuvent se révéler utiles pour un futur grand chef. En revanche, elles ne sont pas indispensables pour l'amateur-cuisinier qui aime faire la cuisine pour sa famille et ses amis

La situation décrite ci-dessus est similaire à celle des étudiants de psychologie sociale au cours de leur formation. Les statistiques sont un outil qui permet aux étudiants d'atteindre un objectif, l'objectif étant de vérifier des hypothèses théoriques. Pour atteindre cet objectif, les étudiants n'ont pas besoin de connaître les théorèmes mathématiques ayant servi aux statisticiens pour développer ou raffiner tel ou tel test statistique. La Figure 1 contient un formulaire ressemblant à ceux qui sont distribués dans de nombreuses universités. Pour vérifier leurs hypothèses théoriques, les étudiants n'ont pas besoin de savoir que la variable aléatoire $Y = \sum_{i \leq n} X_i$ obéit à une loi de Poisson de paramètre $\sum_{i \leq n} \lambda_i$. Tout comme ils peuvent se passer des formules démontrant l'approximation d'une loi binomiale par une loi normale si $k \leq n$ (voir Figure 1).

En revanche, pour pouvoir tester leurs hypothèses théoriques, les étudiants doivent savoir comment analyser les données de leurs études. En d'autres mots, ils ont besoin de savoir quel test statistique est utilisé pour quel type de données. Si la variable dépendante et les variables indépendantes sont catégorielles, c'est un Chi2. Si la variable dépendante et les variables indépendantes sont continues, c'est une analyse de régression. Si la variable dépendante est continue et les variables indépendantes sont catégorielles, c'est une ANOVA. Si les participants sont exposés à plus d'une modalité de l'une des variables indépendantes, c'est une ANOVA avec mesures répétées. Et ainsi de suite. La Figure 2 représente un graphique qui pourrait être distribué aux étudiants et qui pourrait servir comme point de départ pour un tel enseignement. Les étudiants doivent aussi savoir comment on saisit les données, quel menu du logiciel de traitement de données permet d'effectuer le test statistique voulu, et comment on lit et interprète un output. Après avoir acquis ces connaissances, nos étudiants sont parfaitement

1 - Loi uniforme :

Définition : Soit $X : \Omega \to E$ une variable aléatoire avec $E = \{x_1, x_2, \ldots, x_n\}$.
On dit que X obéit à la loi uniforme sur E si pour chaque x_i on a $P(X=x_i) = 1/n$.

2 - Loi hypergéométrique :

Définition : Soit N, N' et n trois entiers vérifiant $n \leq N' \leq N$. Posons alors $p=N'/N$ et $q=1-p$.
Une variable aléatoire X obéit à la loi hypergéométrique de paramètre n, N' et N si elle prend les valeurs entières comprises entre 0 et n avec la loi $P(X=k) = C^k_{N'} C^{n-k}_{N-N'} / C^n_N$. On écrit alors $X \in H(N, N', n)$ et on a $E(X) = np$ et $Var(X) = npq[1-(n-1)/(N-1)]$.

3 - Loi de Poisson :

Définition : Une v.a. X obéit à la loi de Poisson de paramètre λ ($\lambda > 0$) si X prend ses valeurs dans $\mathbb{N}$ avec la loi $P(X=k) = \exp(-\lambda)\, \lambda^k / k!$. On écrit alors $X \in P(\lambda)$ et on a $E(X) = Var(X) = \lambda$.

Propriété : Soit $X_1, X_2, \ldots, X_n$ un v.a. indépendantes de loi respective $P(\lambda_i)$.
Alors la variable aléatoire $Y = \sum_{i \leq n} X_i$ obéit à une loi de Poisson de paramètre $\sum_{i \leq n} \lambda_i$.

4 - Loi avec densité :

Définition : Soit une fonction $f : \mathbb{R} \to \mathbb{R}^+$ possédant certaines propriétés de régularité et soit X une v.a. dont la fonction de répartition est F_X. On dit que X admet f pour densité si pour tout réel u on a

$$F_X(u) = P(X \leq u) = \int_{-\infty}^{u} f(x)dx$$

Propriétés : 1 - $\int_{-\infty}^{+\infty} f(t)dt = 1$.

2 - $P(X = a) = 0$ pour tout réel a.

3 - $P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = \int_a^b f(x)dx = F_X(b) - F_X(a)$.

4 - F_X est une fonction continue et on a $\lim_{u \to -\infty} F_X(u) = 0$ et $\lim_{u \to +\infty} F_X(u) = 1$.

5 - Loi normale $N(m,\sigma)$:

Définition : une v.a. X obéit à la loi normale de moyenne m et d'écart-type σ ($\sigma \neq 0$) si elle admet pour densité la fonction f définie sur R par $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$. On écrit alors $X \in N(m,\sigma)$.

Propriétés : si $X \in N(m,\sigma)$ alors $\frac{X-m}{\sigma} \in N(0,1)$.

6 - Approximation d'une loi binomiale par une loi normale :

En prenant dans le théorème précédent $X \in B(p)$, on a $Y_n = \sum_{i \leq n} X_i \in B(n,p)$ et $Y_n = n\overline{X}_n$.

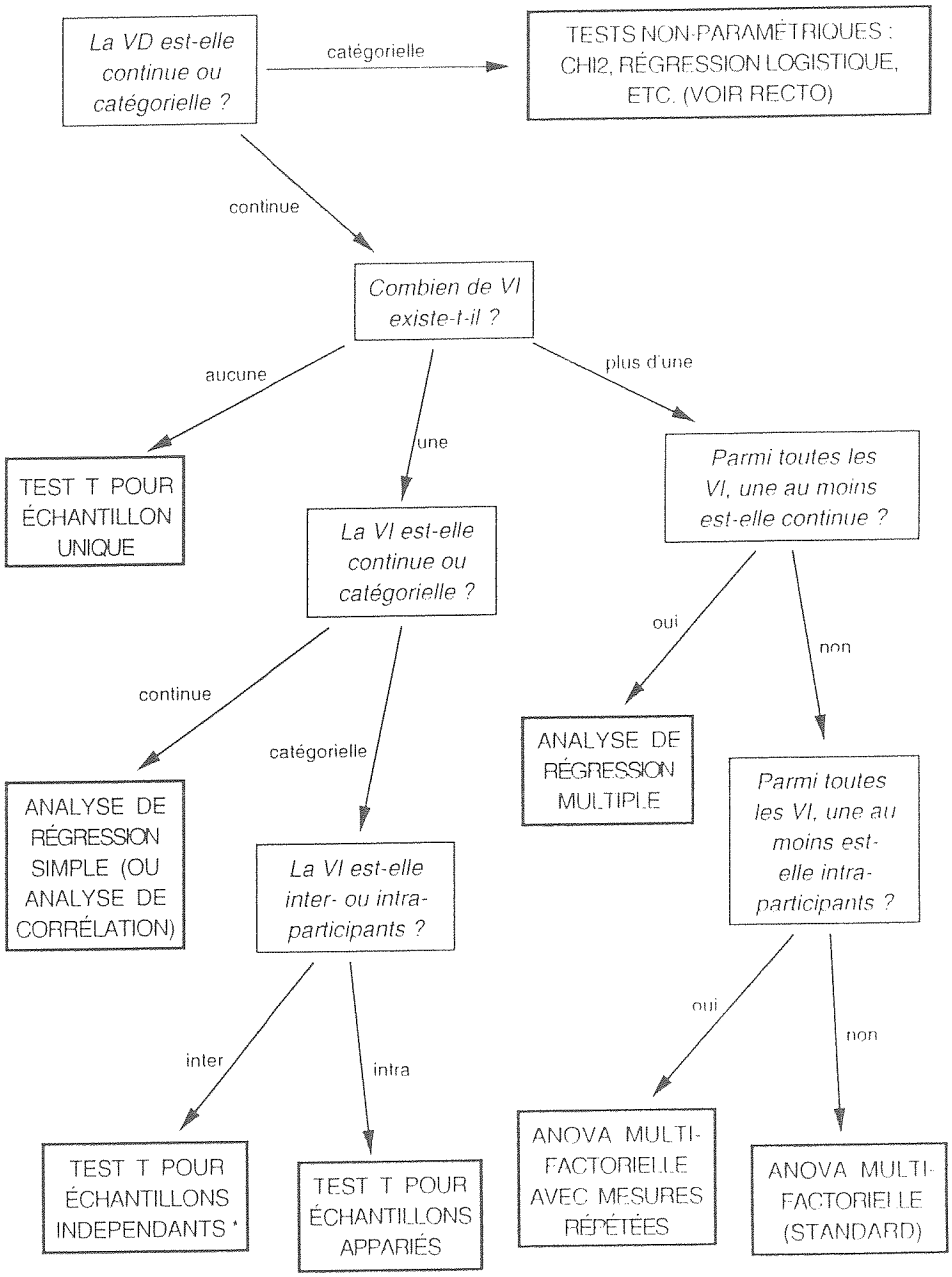
Pour n assez grand on a donc $\overline{X}_n \approx N\left(p, \sqrt{pq}/\sqrt{n}\right)$ et donc $Y_n = n\overline{X}_n \approx N(np, \sqrt{npq})$.

soit encore $\frac{Y_n - np}{\sqrt{npq}} \approx N(0,1)$.

D'où pour $k \leq n$: $P(Y_n \leq k) = P\left(\frac{Y_n - np}{\sqrt{npq}} \leq \frac{k-np}{\sqrt{npq}}\right) \approx F\left(\frac{k-np}{\sqrt{npq}}\right)$.

Figure 1
Notions de base

Quel test statistique pour quelles données ?



* ou ANOVA unifactorielle si la VI a plus de deux modalités

Figure 2

Quel test staistique pour quelles données ?

qualifiés pour le monde de travail où il s'agit de tester l'efficacité d'interventions diverses ou de tester des différences entre plusieurs groupes ayant reçu des traitements différents.

Il est vrai qu'à un niveau plus élevé, il convient de disposer de connaissances plus détaillées sur les bases théoriques des statistiques. Un doctorant en psychologie sociale en fin de thèse doit connaître la théorie d'échantillonnage sur laquelle sont fondés la plupart des tests statistiques que nous utilisons. La connaissance des formules des différents indicateurs statistiques (r, F, χ^2, η^2, etc.) lui permettra de comprendre ce que fait réellement le test statistique que son ordinateur effectue pour lui. Par exemple, une fois qu'il réalise que la valeur F exprime le rapport entre la variance expliquée (par les variables indépendantes) et la somme de la variance expliquée et la variance résiduelle, il peut être amené à modifier sa procédure expérimentale pour minimiser la variance résiduelle *(Judd & McClelland, 1989)*. Une fois qu'il réalise que l'ANOVA est un cas particulier de l'analyse de régression, il va peut-être décider de ne plus dichotomiser ses variables indépendantes continues *(Brauer, sous presse)*. Et une fois qu'il sait interpréter un coefficient de régression dans une analyse de régression multiple, il réalise la nécessité d'effectuer certaines analyses avant de pouvoir affirmer que les données sont consistantes avec le modèle médiationnel prédit par sa théorie *(Brauer, 2000)*. Alors que ces connaissances peuvent s'avérer utiles pour un spécialiste, elles sont relativement superflues pour un étudiant en premier ou deuxième cycle.

Bien sûr, dans un monde optimal, on voudrait que les étudiants en psychologie sociale possèdent les deux types de connaissances en fin d'études: d'un côté l'analyse des données, c'est-à-dire le savoir-faire concernant le choix du test approprié, la réalisation de ce test sur ordinateur, et l'interprétation des résultats; de l'autre côté les statistiques, c'est-à-dire les bases théoriques des tests statistiques. Mais l'expérience montre que c'est difficilement possible. Etant donné que nous semblons être dans l'obligation de choisir (au moins partiellement), l'enseignement de l'analyse de données me paraît plus important que celui des statistiques. Si nous avions affaire à deux étudiants de maîtrise, lequel préférerions-nous: L'étudiant A qui connaît la formule du coefficient de corrélation par cœur, ou l'étudiant B qui sait effectuer une analyse de corrélation sur son ordinateur et ensuite interpréter le coefficient de corrélation ? L'étudiant A qui connaît les bases théoriques du modèle linéaire général, ou l'étudiant B qui sait quel test statistique va avec quel type de données ? L'étudiant A qui connaît l'algèbre des matrices, ou l'étudiant B qui sait effectuer une analyse factorielle sur les données de son projet de recherche et qui peut nous dire que l'échelle qu'il a développée semble satisfaisante car elle ne mesure qu'un seul concept psychologique ? Il est indéniable que personnellement, j'aurais une préférence pour l'étudiant B car il réussira probablement mieux son travail de recherche. D'ailleurs, je suis convaincu que c'est lui qui s'en sortira mieux dans le monde professionnel.

La position défendue dans cet article est volontairement extrême pour faire un contraste avec l'orientation actuelle de l'enseignement des statistiques pour les psychologues sociaux. Comme c'est souvent le cas, la solution idéale se trouve probablement au milieu. Le but de cet article est de stimuler des discussions sur ce point. Après tout, l'analyse des données est pour les psychologues sociaux un outil et non pas une finalité en soi. Au moins au niveau des premier et deuxième cycles, essayons de focaliser l'attention des étudiants sur la façon dont on se sert de cet outil plutôt que sur les aspects théoriques de cet outil *(voir Doise, Clémence, et Lorenzi-Cioldi, 1992, pour un bon exemple)*. Aussi, en intégrant l'enseignement de l'analyse des données avec celui de la méthodologie, les étudiants se rendront davantage compte que ces deux aspects de l'expérimentation sont intimement liés.

Il reste à voir si la distinction entre les termes *analyse des données* et *statistique* est pertinente. Le terme analyse de données a pour le moins l'avantage de refléter ce que nous faisons en tant que psychologues sociaux: notre objectif principal n'est pas de faire des statistiques mais d'analyser et d'interpréter nos données pour vérifier nos hypothèses théoriques.

Références

— BRAUER, M. (sous presse): L'analyse des variables indépendantes continues et catégorielles, *L'Année Psychologique*.

— BRAUER, M. (2000): L'identification des processus médiateurs dans la recherche en psychologie, *L'Année Psychologique*, 100, p. 661-681.

— DOISE, W., CLÉMENCE, A., & LORENZI-CIOLDI, F. (1992): *Représentations sociales et analyse de données*, Grenoble, Presses Universitaires.

— JUDD, C. M., & MCCLELLAND, G. H. (1989): *Data analysis: A model comparison approach*, San Diego (CA), Hareourt Brace Jovanovich.