Sur les remarques de Claudine Schwartz

1- Notations

Plusieurs remarques sont faites concernant les notations utilisées [CS, p. 1]. En utilisant les notations de Gavarret, nous souhaitions essayer, d'une part de coller au plus près du texte et d'autre part, amener le lecteur à faire un va et vient entre les notations actuelles et celles utilisées à l'époque. Cependant ce parti pris peut se discuter. Page 12 de l'article, la forme moderne est explicitée puisqu'il est écrit : "Le théorème de Poisson revient à prendre $e = 2\sqrt{\frac{2f(1-f)}{n}}$."

C. Schwartz fait remarquer justement que l'utilisation d'une formulation comme "6094 condamnations sur 10 000 accusés" n'est plus utilisée aujourd'hui [CS, p. 2]. Là encore, nous avons conservé l'expression de Gavarret, à la fois pour être plus près du texte mais aussi pour montrer qu'il peut exister d'autres façons d'exprimer un pourcentage. Cette expression nous semble facile à comprendre pour un lecteur moderne.

C. Schwartz pense trouver dans cette formulation, une "réticence à considérer des probabilités qui ne soient pas des nombres rationnels" [CS, p. 2]. Il ne nous semble pas que Gavarret formule ici la valeur numérique prise par une probabilité. Sauf erreur de notre part, dans son texte, quand cette formulation existe, il s'agit toujours de fréquences (sauf pour mesurer le risque de première espèce).

2- Chiffres et idéologie

Les compléments apportés [CS, p. 3] sur la concomitance "polygamie - sex-ratio" sont très intéressants. À notre connaissance, Gavarret ne revendique nulle part de découverte dans ce domaine grâce au test utilisé. Il prend seulement les exemples de son époque pour appuyer la pertinence de la formule de Poisson.

Concernant ce problème du "sex-ratio", nous sommes tout à fait d'accord sur les précautions à prendre et les erreurs à éviter quant à la manipulation des nombres. Nous n'avons pas abordé ce problème dans notre article et il nous semble qu'un renvoi à la bibliographie [CS, p. 3, notes 5 à 7] peut intéresser les lecteurs.

3- Intervalle de confiance et test

C. Schwartz conteste les formulations "*erroné*" et "*faux*" utilisées pour qualifier le procédé qui consiste à comparer des intervalles de confiance pour décider entre l'hypothèse nulle et l'hypothèse alternative [CS, p. 4-5].

Gavarret ne parle pas effectivement de significativité et il se contente de considérer que, lorsqu'on déclare une différence qui n'existe pas, la probabilité de se tromper est si petite qu'elle autorise à énoncer cette proposition comme une impossibilité. Mais cependant, implicitement, il chiffre cette probabilité (ici 0,0047) qui est bien le risque α.

Mais, dans les définitions qu'introduit C. Schwartz, n'y a-t-il pas risque de confusion à parler de α-significativité avec MIC sans préciser que α ne désigne pas ici le risque de première espèce ?

En effet, soit p une proportion inconnue et p^* une proportion de référence. L' intervalle de confiance de niveau $1-\alpha$ pour p, noté I_α, contient (approximativement) toutes les valeurs p^* du paramètre non rejetées avec le test $H_0 : p = p^*$ contre $H_1 : p \neq p^*$ effectué au risque α. Dans ce cas, pour un p^* donné, il y a (presque) équivalence entre le test classique : $H_0 : p = p^*$ contre $H_1 : p \neq p^*$ effectué au risque α et la méthode consistant à rejeter l'hypothèse nulle si $p^* \notin I_\alpha$. Dans les deux cas, le risque de se tromper en rejetant à tort l'hypothèse nulle est α.

Cette (presque) équivalence entre les deux méthodes qui existe quand on souhaite comparer une proportion inconnue à une proportion de référence n'existe plus quand on souhaite comparer deux proportions inconnues p et $p^{'}$. En effet, soit I_α (resp. $I^{'}_\alpha$), l'intervalle de confiance de niveau $1-\alpha$ pour p (resp. $p^{'}$). Si on rejette $H_0 : p = p^{'}$ quand I_α et $I^{'}_\alpha$ sont disjoints, le risque de se tromper en rejetant à tort l'hypothèse nulle (ce qui signifie que p et $p^{'}$ ont alors une valeur commune) n'est plus α mais un $\alpha^{'}$ qui dépend des tailles des échantillons et de la valeur commune de p et $p^{'}$. Au passage, remarquons que la difficulté qui concerne le risque de deuxième espèce, à savoir dépendre de beaucoup de paramètres inconnus, se trouve alors présente dans le calcul de $\alpha^{'}$.

Dans le test MT, lorsqu'on dit que la différence est significative au risque 0,05, cela signifie que le risque de se tromper en affirmant une différence est plus petit ou égal à 0,05, mais dans le test MIC, lorsqu'on dit que la différence est significative, au risque 0,05, cela signifie que le risque de se tromper en affirmant une différence est plus petit ou égal à $\alpha^{'}$ avec $\alpha^{'} < 0{,}05$ sans donner plus de renseignement sur $\alpha^{'}$.

Pour notre part, nous pensons que le problème majeur de cette notion de différence fortement significative n'est pas sa difficulté à la généraliser à la comparaison globale de plusieurs fréquences (c'est hélas le même problème pour le test MT), mais l'impossibilité de "contrôler" le risque de première espèce.

Les qualificatifs "*erroné*" et "*faux*" concernaient seulement un procédé qui aurait pu laisser croire que le risque de première espèce restait le même et que les deux méthodes étaient équivalentes en terme de risques. Dans la mesure où il est précisé que le α de l'α-significativité ne correspond pas au risque de première espèce, les mots utilisés par nous sont sans doute un peu forts !

4- A. Quetelet

C. Schwartz fait remarquer que les ponts éventuels (ou l'absence de ponts) entre Gavarret et Quetelet n'ont pas été explorés [CS, p. 5]. Nous soupçonnons que Gavarret et Quetelet se sont servis de données existant dans Poisson. C'est en effet une piste intéressante qu'il faudrait approfondir.

A propos de la comparaison entre ce qui se passe avant 1830 et ce qui se passe en 1830, C. Schwartz dit "il faudrait dire qu'il teste si la nouvelle législation a eu ou non un effet significatif" [CS, p. 5].

Nous pensons que l'exemple est davantage utilisé pour illustrer la phrase « Toutes les fois que deux observateurs se placeront dans des circonstances bien évidemment différentes pour chercher la chance moyenne de production du même phénomène, on peut prédire que leurs résultats offriront une différence supérieure à la *limite* des oscillations des compatibles avec l'invariabilité des causes possibles » [G, p. 84] que pour en décider une différence. En effet, dans le cas traité, il s'agissait bien de "circonstances bien évidemment différentes" et il ne s'agit pas ici de tester si la nouvelle legislation a eu ou non en effet significatif. Plus loin dans le texte, Gavarret explique que « si, par contre, il est permis, dans certaines questions, de signaler l'intervention d'une perturbation notable et bien évidente ; il n'en est pas toujours ainsi » [G, p. 85]. C'est alors ici, comme par exemple dans le fameux problème du sexe des enfants illégitimes, qu'un test est utile pour chercher si une différence est "significative" (au sens de Gavarret c'est à dire avec $\alpha = 0{,}0047$). La "significativité" de cette différence amène à décider ici que les "circonstances" sont différentes, mais comme le fait remarquer C. Schwartz, ne permet en aucun cas de décider quelle est la nature de ces "circonstances".

5- Claude Bernard

L'étude des réticences sur l'utilisation des statistiques en médecine dans la deuxième moitié du 19$^{\text{ème}}$ siècle est une piste que nous nous proposons d'approfondir.

6- Au fil de la lecture de [LTT]

Nous sommes en accord avec le fait qu'il est un peu osé de parler d'une intuition attribuée à Gavarret de ce que l'on nomme aujourd'hui des données censurées [CS, p. 7]. En effet, aujourd'hui ce terme bien précis ne désigne que des données dont on connait, avant la date de l'étude, la présence d'une pathologie, mais pour lesquelles, à la date de l'étude il est impossible de connaître l'issue. Il ne semble pas que Gavarret envisage ce cas.

Concernant la formule de Liapounov citée à la page 25 de notre article, il aurait été plus judicieux de parler du théorème central limite avec la condition de Liapounov [CS, p. 7]. Nous modifions notre texte dans ce sens. Rappelons ce dont il s'agit :

Soient $X_1, X_2, \cdots, X_n$ une suite de variables aléatoires définies sur le même espace de probabilité et indépendantes.

On suppose que $\mathbb{E}(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$ et $\mathbb{E}(|X_i - \mu_i|^3)$ sont finis, pour tout $i = 1, ..., n$

On pose : $m_n = \sum_{i=1}^{n} \mu_i$, $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$ (moyenne et variance de $X_1 + X_2 + \cdots + X_n$) et $r_n^3 = \sum_{i=1}^{n} \mathbb{E}(|X_i - \mu_i|^3)$.

La condition de Liapounov s'énonce ainsi : $\lim \dfrac{r_n}{s_n} = 0$ quand n tend vers $+\infty$.

Alors, si la condition de Liapounov est vérifiée : $\dfrac{X_1 + X_2 + \cdots + X_n - m_n}{s_n}$ tend en loi vers $\mathcal{N}(0, 1)$.

Rappel de l'annexe :

On pose $X_i^{(1)} \sim \mathcal{B}(p_1)$ pour $1 \leq i \leq n_1$ (resp. $X_k^{(2)} \sim \mathcal{B}(p_2)$ pour $1 \leq k \leq n_2$) la variable qui vaut 1 si E est arrivé dans la première série (resp. dans la deuxième série).

Soit $f_{n_1} = \dfrac{1}{n_1} \sum_{i=1}^{n_1} X_i^{(1)}$ (resp. $f_{n_2} = \dfrac{1}{n_2} \sum_{k=1}^{n_2} X_k^{(2)}$) la fréquence d'apparition de E dans la première série (resp. la deuxième série).

Soit $Y_j = \dfrac{1}{n_1} X_j^{(1)}$ si $1 \leq j \leq n_1$ et $Y_j = -\dfrac{1}{n_2} X_{j-n_1}^{(2)}$ si $n_1 + 1 \leq j \leq n_1 + n_2$.

On pose encore :

$$W_{(n_1,n_2)} = \frac{Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2} - \mathbb{E}(Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2})}{\sqrt{Var(Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2})}}$$

Mais ici $\mathbb{E}(Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2}) = m_n$ et
$\sqrt{Var(Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2})} = s_n = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
d'autre part $r_n = \left(\frac{p(1-3p_1+4p_1^2-2p_1^3)}{n_1^2} + \frac{p(1-3p_2+4p_2^2-2p_2^3)}{n_2^2} \right)^{\frac{1}{3}}$.
Il est aisé de montrer que la condition de Liapounov est vérifiée et donc :

$$\frac{(f_{n_1} - f_{n_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \text{ tend en loi vers } \mathcal{N}(0,1)$$

Il ne reste plus qu'à estimer s_n par $\sqrt{\frac{f_{n_1}(1-f_{n_1})}{n_1} + \frac{f_{n_2}(1-f_{n_2})}{n_2}}$ pour retrouver la formule de Poisson.

Concernant les tests choisis a posteriori, la remarque de C. Schwartz [CS, p. 7] est très judicieuse et nous avons complété la phrase "Mais on sait qu'il n'est pas équivalent de tester tous les groupements de jours deux à deux et de tester globalement l'équirépartition ;" par "il n'est pas non plus équivalent de tester globalement l'équirépartition et d'effectuer des tests sur des groupes de jours choisis a priori par le statisticien, même si ceux-ci semblent refléter des situations extrêmes".

La critique de l'activité 1 [CS, p. 7] consiste à dire que, puisque que l'on peut tomber deux fois sur la même personne, les échantillons ne seraient plus indépendants. Les problèmes engendrés par une absence d'indépendance ne se poseraient que si les calculs, à notre avis exacts, étaient suivis d'une inférence statistique.
Ici un des objectifs de l'activité est seulement d'observer si les "erreurs" se compensent ou non.

Denis Lanier, Jean Lejeune, Didier Trotoux