Le paradigme de Neyman-Pearson, dit classique ou fréquentiste[1] de l'inférence statistique

Marc Bourdeau

Louis.Marc.Bourdeau@Gmail.com

R.A. Fisher avait mauvais caractère, très mauvais, et il avait la dent rancunière. Il n'a jamais admis le point de vue des Neyman & Pearson sur l'inférence. Qui le mettait en situation de ne pas avoir tout compris… À savoir que les hypothèses viennent naturellement par paires, ou elles sont vraies (acceptables) ou elles sont fausses (non acceptables), ce qui constitue l'avancée importante sur l'inférence fishérienne[2] qui ne reconnait qu'une hypothèse, tant et si bien que le point de vue des Neyman & Pearson constitue depuis les années trente du siècle dernier le paradigme classique (ou fréquentiste) de l'inférence statistique, qui peut donc à bon droit s'appeler le paradigme de Neyman-Pearson de l'inférence statistique.[3]

Dans la pratique, comment ne pas penser en effet à des paires d'hypothèses? Un nouveau traitement est-il meilleur ou non? Trouve-t-on moins d'imperfections dans une amélioration proposée d'un procédé industriel, moins d'attentes lors d'une réorganisation d'un service, plus de durabilité, etc. Une hypothèse ne vient jamais sans sa contrepartie. C'est du moins le cas en statistique.

Ce qui n'est pas le cas lorsqu'on veut établir, par exemple, la valeur d'un paramètre pour un produit. Dans ce cas, on ne parle pas d'une hypothèse, mais on cherche une fourchette de valeurs possibles pour sa valeur. Ce paramètre a une valeur… variable[4], i.e. plus ou moins incertaine, elle vient invariablement avec un intervalle où on pense qu'elle se situe. On parle d'intervalle de confiance, ou de tolérance, selon le contexte, et ces concepts sont probabilistes. Pas de certitude, un produit particulier peut ne pas être conforme à la valeur spécifiée, la valeur du paramètre ne se situant dans ces intervalles.

---

[1] Qu'on oppose au paradigme bayésien, du nom de Thomas Bayes [1702-1761], parfois laplacien, du nom de Pierre Simon, marquis de Laplace [1749-1827] qui l'a redécouvert. Jerzy Neyman [1894-1981]. Egon Sharpe Pearson (fils de Karl)]1895-1980].

[2] En fait, Fisher a proposé et voulu imposer contre les points de vue de Neyman & Pearson une inférence fiduciale qui fut enterré rapidement dans la controverse. La simplicité et la logique impeccable de l'inférence de Neyman-Pearson l'ont rendue incontournable.

[3] On ne peut que renvoyer, pour les détails historiques assez complexes mais passionnants, aux ouvrages de Droesbeke & Tassi (1990), ainsi que Stigler (1999).

[4] « *There is no true value of anything.* » (W. Edward Deming [1900-1993]). Dans sa préface à Walter A. Shewhart [1891-1937], *Statistical method from the viewpoint of quality control,* 1939 réédité en 1986, New York NY: Dover Publications.

Un nouveau procédé nous amènera à poser une hypothèse pour la valeur de ce paramètre : est-elle conforme ou non à sa spécification? Là on pose une paire d'hypothèses : l'hypothèse dite nulle (la seule que Fisher reconnaissait), et l'hypothèse dite *alternative*.

Si, par exemple, on veut tester une certaine hypothèse de base, dite nulle, pour une moyenne: ou elle est acceptable ou elle ne l'est pas. L'alternative pose qu'elle ne serait pas acceptable. Il faut donc avoir tête une 'expérience' statistique pour la mettre en jeu, la tester.

Pour une paire d'hypothèses donnée sur un paramètre (c'est la variante principale, on distingue aussi les hypothèses sur des lois), l'hypothèse de base, dite *nulle*, propose une valeur donnée pour un paramètre d'une loi probabiliste — tout paramètre vient en principe avec une loi probabiliste associée[5]. C'est une hypothèse dite *simple.* Les hypothèses dites *alternatives*, sont invariablement *composites* : la vraie valeur du paramètre est située dans un intervalle semi-infini, ou à gauche, ou à droite ou différente (dans ce cas, ce sont deux intervalles semi-infinis) de la valeur présumée dans l'hypothèse de base. Il y a donc deux hypothèses. On doit se décider en faveur de l'une ou de l'autre. Pour prendre un exemple, tester l'égalité de deux moyennes :

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_1 : \mu_1 - \mu_2 < 0, \text{ ou } H_1 : \mu_1 - \mu_2 > 0, \text{ ou } H_1 : \mu_1 - \mu_2 \neq 0.$$

H_0 est l'hypothèse nulle ou de base, H_1 est l'hypothèse alternative. La première alternative est dite unilatérale à gauche, la deuxième à droite, la troisième est dite bilatérale.

On trouve forcément alors deux risques d'erreur qu'il faut savoir quantifier : ou on n'accepte pas à tort l'hypothèse nulle présumée vraie sur la valeur du paramètre en faveur de l'autre hypothèse; ou on ne rejette pas à tort l'hypothèse nulle.

Remarquez que les statisticiens ne sont jamais très affirmatifs, ce sont des maîtres de la litote… On entend souvent les raccourcis de langage 'rejeter' l`hypothèse pour 'ne pas accepter'. C'est un raccourci commode mais non exact (comme l'explique le texte de Nadine Schwartz, cité plus haut).

Ne pas accepter (rejeter) à tort l'hypothèse nulle alors qu'elle est acceptable, est dite l'erreur de première espèce; la seconde, l'erreur de seconde espère,[6] soit ne pas rejeter (accepter) à tort l'hypothèse nulle alors qu'elle n'est pas acceptable. Les risques associés à chacune sont évidemment des probabilités. Il y a deux types d'erreur; il faut les quantifier.

Pour la première, ce sera une fonction d'un risque maximal (ou tolérable) de première espèce, noté α; pour la seconde ce sera une fonction de $\alpha\alpha$, de la taille échantillonnale de l'expérience utilisée pour le test, et d'un écart, δ, présumé à la valeur du paramètre spécifié dans l'hypothèse de base.

Un test statistique sur hypothèse nulle consiste à établir un critère *a priori* qui repose sur la rareté de la réalisation d'une variable aléatoire calculée sur des données issues d'une expérience

---

[5] Pour Karl Pearson [1957-1936], le fondateur avec Fisher de la statistique moderne, les lois de probabilité sont le fondement de la réalité, et non les nombres eux-mêmes, une sorte d'idéalisme qui s'oppose au matérialisme intuitif. *In : The Grammar of* Science, 1892, réédité en 2004, New York NY : Dover Publications.

[6] En français, on parle de 'second' lorsque la liste ne comporte pas de troisième terme… De 'deuxième' autrement.

statistique. Cette variable aléatoire est dite échantillonnale, on l'appelle même le *test*, puisque son usage est de tester une hypothèse nulle sur un échantillon, et elle ne sert qu'à ça.

Pour le critère et la décision à prendre, on doit se donner a priori[7] un risque maximal de première espèce, c'est la probabilité α. Cela étant fait, on définit une variable aléatoire dite échantillonnale qui utilise la valeur présumée du paramètre à tester ainsi que la valeur du risque maximal pour établir la décision. Le critère est un jugement de rareté, c'est la probabilité, associée à la réalisation sur un (seul) échantillon de la variable échantillonnale du test. On décide de ne pas accepter l'hypothèse de base si la réalisation de la variable aléatoire du test sur un échantillon (unique) est plus rare que α.

Cette réalisation a une probabilité de dépassement *p*. Par définition, *p* est la rareté du test, de la variable aléatoire, donc qui sert de test. Cette probabilité de dépassement s'appelle en anglais d'un nom qui ne renvoie pas au concept *p-value*. Et ainsi :

$$P[\text{ Rejeter } H_0 \mid H_0 \text{ supposée vraie }] = p < \alpha,$$

Comme les hypothèses viennent tout naturellement par paires : l'hypothèse dite nulle (la seule que Fisher reconnaissait) et l'hypothèse alternative. L'alternative pose que la première serait fausse, dans le sens précisé par l'alternative (à gauche, à droite, ou bilatéral), et *p* dépend de l'hypothèse alternative ainsi, bien sûr que de la loi de la variable aléatoire échantillonnale qui sert de test, du test pour faire bref.

Il y a aussi l'erreur de seconde espèce, ne pas rejeter à tort l'hypothèse nulle en faveur de l'alternative supposée alors vraie. Pour cela, il faut préciser l'alternative, la rendre simple tout comme l'hypothèse de base, il faut se donner une certaine écart, δ, à la valeur de celle-ci :

$$P[\text{ Ne pas rejeter } \mid H_0 \text{ supposée fausse avec écart } \delta\ ] = \beta(\ \alpha\ ;\ n\ ;\ \delta\ ).$$

La probabilité β s'appelle la probabilité de non détection d'un écart donnée à l'hypothèse nulle.

1-β de ce fait s'appelle la puissance du test. La puissance de *détection* dépend bien sûr du triplet (α; n; δ). Pour un α donné, elle augmente avec n, et δ.

Incidemment la valeur 1-α s'appelle l'efficacité du test : la probabilité de rejeter l'hypothèse de base quand elle n'est pas acceptable (fausse), la probabilité de détecter une hypothèse nulle non acceptable. On voit immédiatement que

$$\beta(n;\ \alpha;\ \delta=0) = 1 - \alpha.$$

On ne calcule pas souvent, dans les applications, la valeur du β, car on ne peut pas bien spécifier l'écart δ, et la puissance désirée à cet écart.

C'est ainsi que la théorie probabiliste nous amène à fixer un seuil tolérable pour l'erreur qui consiste à rejeter à tort la valeur posée dans l'hypothèse nulle ou de base. C'est le α. On trouve tolérable d'avoir une probabilité α maximale de se tromper (α est en général assez petite), et si une expérience nous donne un probabilité $p < \alpha$ on rejette l'hypothèse de base. On porte le jugement que l'hypothèse est plutôt non crédible. Pas de certitude, jamais de certitude : on dit qu'au vu des données, on ne peut accepter l'hypothèse nulle, ne pas être en mesure de l'accepter, et pas qu'elle est fausse — les statisticiens pratiquent beaucoup la litote…

[7] On insiste sur cet a priori. De nombreux chercheurs ont une pratique *post hoc*… Voir Kline (2013).

Pour cela on établit une règle de décision fondée sur le résultat d'une expérience statistique qui est calculable au seuil α. Si la probabi1.lité p de la réalisation de l'expérience statistique est inférieure à α, on peut se permettre de ne pas accepter l'hypothèse nulle (de base) tout en courant un risque maximal α de se tromper, ce faisant : P[rejeter H_0 | H_0 soit vraie] = p < α.

Le risque de rejeter à tort l'hypothèse (nulle ou de base) a une probabilité qui alors est calculable, c'est la valeur p qui est inférieure ou non à α, le risque maximal de première espèce tolérable. p est le risque observé, α est une valeur déterminée, p est une observation d'une variable aléatoire, dite le *test*. Alpha est dite le *risque de première espèce*. On n'accepte par l'hypothèse de base si p < α

Il est à remarquer qu'on ne peut dans ce contexte probabiliste éliminer totalement la probabilité du risque de première espèce, à moins de choisir α = 0, ce qui est absurde : à ne jamais rejeter H_0, bien sûr on ne se trompe jamais quand elle est vraie …mais quand elle est fausse? On se retrouve avec β = 1 pour tous les *n* et *δ*, la puissance est nulle ! On échappe certainement à Charybde mais on tombe invariablement sur Scylla.

On ne peut éliminer tous les risques dans un processus décisionnel, à prendre la décision ou à ne pas la prendre. Comme dans la vraie vie! C'est Neyman & Pearson qui ont quantifié et précisé tout cela dans l'inférence statistique. La théorie des tests statistiques nous amène au seuil de la théorie (probabiliste) de la décision. Optimiser les décisions quand les risques sont quantifiables avec des coûts associés à chaque possibilité.

Voici l'ensemble du paradigme de Neyman-Pearson des tests d'hypothèse sur paramètres

1. On fixe les deux hypothèses, la première, l'hypothèse de base H_0, est simple pour la valeur d'un paramètre; la seconde, H_1, est composite.
2. On se donne un seuil α (toujours a priori), une probabilité maximale pour l'erreur de première espèce
3. On détermine une variable échantillonnale en termes de *n*, une taille échantillonnale, et de α
4. Cela nous donne un critère de décision de non acceptation de l'hypothèse de base en faveur de H_1, pour que l'erreur de première espèce ne dépasse pas la valeur maximale α
5. On cueille un échantillon de taille n de valeurs pour le paramètre à tester.
6. On applique le critère de décision calculé sur l'échantillon pour décider

Si on a une exigence de détection d'écarts à la valeur du paramètre à respecter avec une probabilité prescrite, une puissance prescrite donc, on peut

7. Calculer la taille échantillonnale à cueillir pour satisfaire à ces exigences (on appelle cela le contrôle des β par la taille échantillonnale

Abaques de puissance

La terminologie est passablement confuse. Si on se place dans une perspective de contrôle de qualité, pour le contrôle/vérification/surveillance des lots plus particulièrement, on parle alors des courbes d'efficacité. En anglais on utilise le terme : « *Operating characteristic (OC) curves* »[8]

Dans notre perspective de la théorie des tests à la Neyman-Pearson, il s'agit d'utiliser des graphiques des erreurs de seconde espèce, β, pour obtenir des puissances, en fonction des trois paramètres (α; n; δ), et comprendre en profondeur le fonctionnement du paradigme classique :

$1 - \beta(\alpha ; n ; \delta)$, dont on sait déjà que, quel que soit n : $1 - \beta(\alpha ; n ; \delta = 0) = 1 - (1 - \alpha) = \alpha$.

Dans la pratique, on utilise les abaques de puissance surtout pour contrôler les erreurs de seconde espèce par un bon choix de taille échantillonnale. Cela complète le développement de ce paradigme, c'est très utile pour la planification des expériences statistiques.

On développe le détail de l'utilisation de ces abaques sur un cas particulier. Pour les autres cas, nous ne présentons que les abaques (Bayer, 1968; chap. IV.5, p. 290-292).

On obtient les puissances par complémentation à 1 : $1 - \beta(\alpha ; n ; \delta)$. On note que $\beta(\alpha ; n ; \delta = 0) = 1 - \alpha = 0{,}95$, donc la puissance de 5% avec l`hypothèse nulle exacte. On note que $\beta(\alpha ; n ; \delta)$ diminue avec l'écart δ croissant (et Δ aussi) quel que soit n la taille échantillonnale. Pour un δ (Δ) fixé, $\beta(\alpha ; n ; \delta)$ diminue avec n, et donc la puissance augmente avec la taille échantillonnale.

[8] On pourra consulter la page d'assistance du logiciel Minitab : http://support.minitab.com/fr-fr/minitab/17/topic-library/quality-tools/acceptance-sampling/acceptance-sampling-graphs/operating-characteristic-oc-curve/

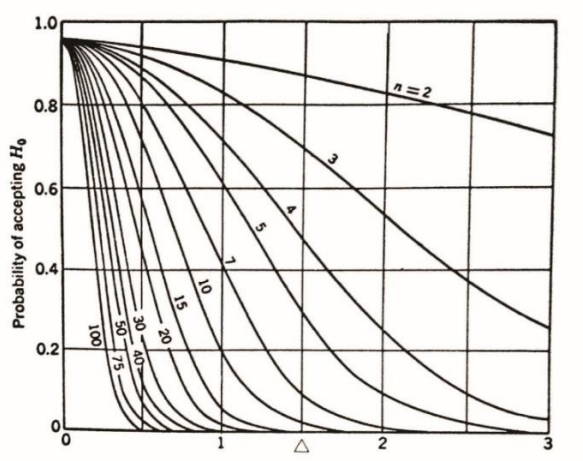

Fig. 1 — Les erreurs de seconde espèce $\beta(\alpha ; n ; \delta)$ pour le cas du test-T sur une paire de moyennes, $H_0 : | \mu_1-\mu_2 | = 0$, l'hypothèse alternative bilatérale $H_1: | \mu_1-\mu_2 | \neq 0$, avec variance inconnue. Le cas $\alpha = 0{,}05$. $\Delta = \Delta =| \mu_1-\mu_2 | / 2\sigma$, un écart δ un écart à la moyenne présumée (ou différence entre deux moyennes) normalisé par 2σ.

Enfin, le croisement d'une verticale pour un Δ et d'une horizontale pour un β prescrits permet de déterminer approximativement une taille échantillonnale minimum exigée pour satisfaire ces *desiderata.*

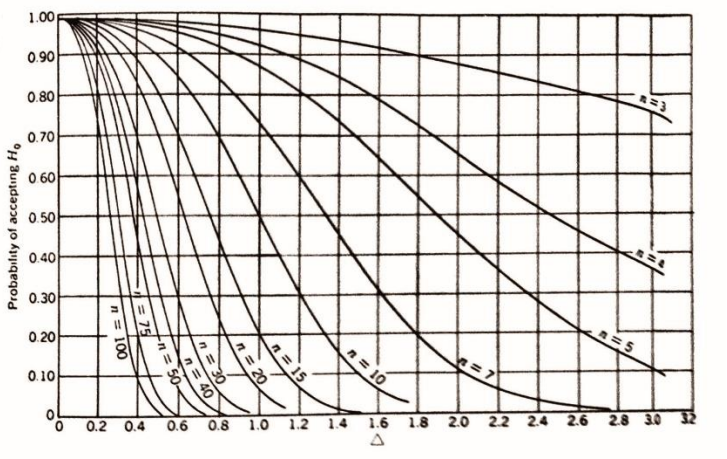

Fig. 2 — La même que la Fig. 1, avec cette fois $\alpha = 0{,}01$. Ainsi à une demie-différence normalisée présumée de 0,4 écarts types pour $\delta = | \mu_1-\mu_2 | / 2$, i. e. $\Delta =| \mu_1-\mu_2 | / 2\sigma = 0{,}4$, on a une puissance de 90% (0,9) avec $n = 100$, 60% avec $n = 50$, et seulement de 20% avec n=20. Pour une puissance désirée supérieure à 90% à 0,2 écart type de $\delta = | \mu_1-\mu_2 | / 2$ ou $\Delta = 0{,}2$, il faut calculer la taille échantillonnales requise à l'aide d'une formule que nous ne développons pas ici…

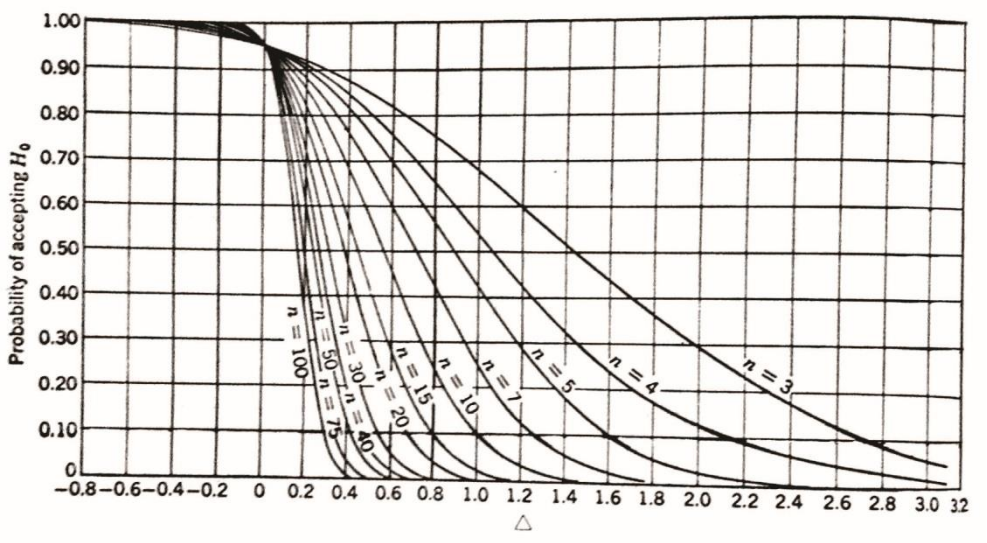

Fig. 3 — Le cas du test-T pour H_0 : $|\mu_1-\mu_2| = 0$, l'hypothèse alternative H_1 : $=|\mu_1-\mu_2| > \text{ou} < 0$ et $\alpha = 0{,}05$.

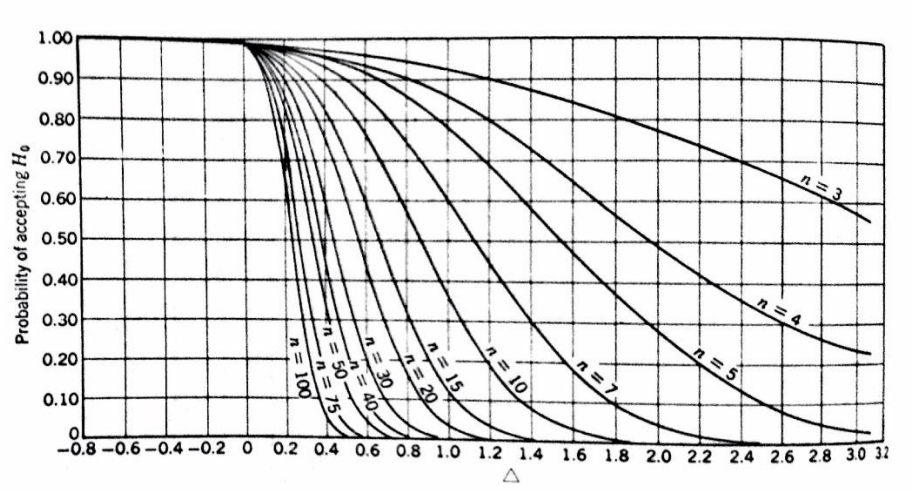

Fig. 4 — Le cas analogue à celui de la Fig. 3, avec $\alpha = 0{,}01$.

Le lecteur intéressé à visualiser ce que signifie un écart normalisé à l'hypothèse nulle en unité de σ pourra utiliser l'animation suivante.

Le *d* de Cohen est précisément le double de la valeur de cet écart normalisé : $\Delta =|\mu_1-\mu_2| / 2\sigma = d/2$, obtenu de deux moyennes expérimentales. Par définition du *d* de Cohen, $d =| X_1bar - X_2bar | / \sigma$, i.e. que *d* s'exprime en termes de l'écart type entre deux valeurs moyennes estimées. Par exemple, d = 1, signifie que distance $| X_1bar - X_2bar |$ vaut 1 écart type (on suppose ici que les moyennes ont le même écart type). Il y a des variantes du *d* de Cohen qui utilisent les écarts types regroupés des deux échantillons (*pooled sigmas*), etc., ce qui ne change pas grand-chose.

Et maintenant…

Nous avons fait le tour du paradigme de Neyman-Pearson. Mais est-ce bien tout ? C'est sans doute plus complexe que d'enseigner seulement la version Fisher, incomplète et illogique finalement, car s'il y a un risque de se tromper à ne pas accepter l'hypothèse nulle — qui ne porte pas ce nom dans l'inférence de Fisher —, il y a indéniablement un risque de se tromper à ne pas la rejeter… Peut-être à cause de la querelle entre Fisher et Neyman-Pearson assez virulente parfois, prenant modèle sur la naissance orageuse du paradigme de Neyman-Pearson (Droesbeke, 1990) et qui a fait couler beaucoup d'encre, ou simplement à cause de la difficulté à bien comprendre les incertitudes/risques liées à la non acceptation ou l'acceptation de la paire d'hypothèses (tout enseignant peut en témoigner…), peut-être aussi à cause des litotes (prudence) utilisées dans ce contexte, le paradigme classique n'a pas cessé d'engendrer des controverses assez virulentes elles aussi parfois. Au point que ces dernières années, au moins une revue scientifique, *Basic and Applied Social Psychology*, a décidé de bannir de ses pages toute référence à cette façon de faire. Mais pas d'inférences pas de statistique !..

Cela a engendré une crise sans précédent dans le monde scientifique. Dans le fichier sous-jacent nous avons décortiqué ce problème à l'occasion de ce bannissement. C'est l'ignorance encore ici qui mène le bal.[9] Il ne faut pas en demander trop aux paradigmes de décision en présence d'incertitude. Même le paradigme bayésien est très questionnant. Des certitudes il n'y en a pas, il n'y en aura pas. On ne peut échapper à l'erreur[10] !

Nous ne développerons pas ici les difficultés de l'approche classique. Mentionnons-en quelques-unes qu'on va élaborer quelque peu plus bas. Quelle valeur se donner pour α ? Que signifie réellement la valeur fixée dans l'hypothèse nulle ? Si on augmente la taille échantillonnale, on peut augmenter la probabilité de ne pas accepter/rejeter une hypothèse nulle, ce qui est le plus cher désir des chercheurs — pas d'effet significatif, pas de publication, pas de subvention… — alors la tentation est forte de se laisser aller jusqu'au rejet de l'hypothèse nulle !

Quelle valeur critique du risque de première espèce se donner a priori? Conventionnellement maintenant, on pose, arbitrairement α =0,05, ce nombre a varié dans le temps, il fut de 0,01 à une époque, mais ce nombre apparaît trop petit, car toute réduction de ce risque implique, à une taille échantillonnale donnée, une augmentation du risque de seconde espèce, β, de ne pas rejeter à tort une hypothèse nulle, pour une distance fixée à l'hypothèse nulle, et pas de publication car on n'accepte pas de moins en moins l'hypothèse nulle ... Par contre beaucoup pensent que α = 0,05 est trop risqué (ce qui n'est pas le cas des parieurs : qui voudrait prendre un pari qu'il estime à une cote 1 contre 20…). Citons ici Fisher (1935):

> « *Personally, the writer prefers to set a low standard of significance at the 5 percent point. […] A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.* »

Dans le « *rarely fails to give this level of significance* », Fisher fait référence à la *nécessaire* réplication des expériences statistiques. Mais c'est justement les problèmes de non-réplicabilité qui ont engendré la crise dont nous avons parlé plus haut.

---

[9] La transmission des connaissances statistiques est, elle aussi, en crise.

[10] Voir la note 4 plus haut.

Il ne faut pas oublier la valeur de la probabilité de dépassement, *p*, observée de l'expérience statistique est elle-même une variable aléatoire, donc sujette à incertitude, ce qui ajoute une incertitude supplémentaire au jugement 'les données observées ne permettent pas d'accepter l'hypothèse nulle… Il s'agit invariablement en effet d'une probabilité de dépassement, qu'on peut noter à juste titre, $p_{dép}$,[11] dont la valeur doit être inférieure à α pour ne pas accepter l'hypothèse nulle et justifier ce jugement.

Mais plus au fait[12], l'organisation scientifique actuelle, avec ses politiques de ne publier que des études montrant un effet significatif est un incitatif à la malhonnêteté (et l'impératif de Fisher, alors ?..). Changer quelques petites données d'un chouia, ne pas garder certaines observations, et hop ! on passe d'un $p_{dép}$ = 0,06 à un léger dépassement du critère …et à une publication !

Comme le souligne David Donoho (2015), un des statisticiens les plus importants des récentes décennies, tout cela ne renvoie pas à un problème statistique mais à un problème de l'*organisation* scientifique. Et il n'est pas près de se régler en dépit de toutes les directives de publication des revues. Le problème en est un d'honnêteté des chercheurs. Les vertus se perdent ! Ce ne sont pas les directives, si strictes soient-elles, telles celles de l'*AERA (American Educational Research Association)*, qui en viendront à bout. Non seulement sont-elles facilement contournables, elles ne sont même pas respectées (Ellis, 2010 révisé 2013, chap. 4).

Le second problème qui saute aux yeux dans l'approche par hypothèses pour l'inférence statistique, est celui de la signification d'une hypothèse sur un paramètre, ainsi pour $H_0 : \mu = \mu_0$. Comme on l'a dit à la note 10 plus haut (!), les valeurs exactes n'existent pas. Tout est sujet à erreurs et incertitudes. Le μ_0 vient donc avec sa marge d'erreur, typiquement un intervalle de confiance auquel on associe une probabilité donnée, typiquement 95%. C'est dire (on doit se rapporter à la théorie classique des tests pour bien comprendre le sens de cette confiance.[13]

Souvent, quand on n'accepte pas une hypothèse nulle, on se donne la peine de calculer sur la base de l'échantillon testé un nouvel intervalle de confiance pour le paramètre testé. Dont la valeur centrale sert de base pour les nouvelles valeurs des hypothèses nulles postérieures, ainsi de suite.

Terminons cet exposé (il faudra y revenir), par la question fondamentale : on peut toujours s'organiser pour ne pas accepter une hypothèse nulle avec des écarts minimes échantillonnés à la valeur du paramètre hypothétique, avec n'importe quel α si petit ou grand soit-il prévu a priori. Scientifiquement parlant, on est justifié de le faire. Mais ne faudrait-il pas distinguer entre la signification pratique et la signification statistique. Ainsi, avoir un nouveau traitement pour le psoriasis qui passe d'un taux moyen de guérison complète de μ_1=35% avec l'ancien à μ_2=35,4% statistiquement avéré, est-il important dans la pratique au point de totalement abandonner la production et la prescription du premier[14] ? Avec toutes les incertitudes qui risquent de se déclarer au cours du temps, les coûts de substitution, etc.

---

[11] En anglais, on utilise une expression qui ne veut absolument rien dire, *p-value*. 'value' de quoi?

[12] « *The fact of the matter »*, est une expression difficile à traduire en français…

[13] On donne la référence Wikipedia en anglais, plus complète que celle en français…

[14] On pourra lire avec profit, le texte « Autisme et antidépresseurs », qui développe ces questions épidémiologiques et de santé publique. On pourrait ouvrir ici un chapitre sur la Théorie de la décision statistique.

Ainsi il y a des distances $|\mu_1 - \mu_2|$ qui sont totalement inintéressantes pratiquement dans certains contextes et majeurs dans un autre. On rapporte cette distance en termes d'écarts types, en divisant cette distance par un écart type approprié. Cohen (1968) a proposé de dénoter cette grandeur, *appelée* grandeur (ou taille) de l'effet : $\delta = |\mu_1 - \mu_2| / \sigma$, où l'écart type est choisi selon le type de paires d'hypothèses envisagées.[15] Pour une illustration intéressante. D'après Cohen (1968, 1988), on peut penser à des balises pour les estimations de $D = |X_1\text{-}bar - X_2\text{-}bar| / S$ de δ, où les variables aléatoires sont calculées sur les échantillons de l'expérience statistique :

$$d = |x_1\text{-}bar - x_2\text{-}bar| / s.$$

Les balises de Cohen du type « tailles de chemise », P, M, G, sont plutôt controversées. Ainsi on dit que $d = 0{,}2$ (soit la distance entre les deux moyennes est égale à 0,2 écarts types), est dite P(etite) ; 0,5 de M ; 0,8 G. À la réflexion, cela doit bien dépendre des contextes, non ?

Pour une discussion de ces problèmes, et des liens entre les *d* et la puissance des tests des expériences statistiques, on pourra se reporter à Ellis (2010, chap. 4).

À notre avis, même chez Ellis, les choses ne sont pas toujours bien claires. Nous devrons y revenir…

— 27 mars 2017

Références

AERA (American Educational Research Association). (2006). *Standards for reporting on social science research in AERA publications.* Consulté le 22 mars 2017, sur http://www.aera.net/Publications/Standards-for-Research-Conduct

Beyer, W. H. (Éd.). (1968). *Handbook of tables for probability and statistics* (éd. 2). Cleveland OH: The Chemical Rubber Co.

Cohen, J. (1968). *Statistical power analysis for the behavioral sciences* (éd. 1re). Hillsdale NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (éd. 2e). Hillsdale NJ: Lawrence Erlbaum Associates.

Donoho, D. (2015). *50 years of data science.* Consulté le 23 mars 2017.

Droesbeke, J.-J., & Tassi, P. (1990). *Histoire de la statistique.* Paris F: Presses universitaires de France, Coll. Que sais-je? No 2527.

[15] Dans le Wikipedia en français, on a le lien suivant pour Taille d'effet

Elllis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results.* Cambridge MA: Cambridge University Press.

Fisher, R. A. (1935). *The design of experiments.* Edinburg, UK: Oliver and Boyd.

Kline, R. B. (2013). *Beyond significance testing. Statistics reform in the behavioral sciences* (éd. 2e). Washington DC: American Phychological Association.

Liu, X. S. (2014). *Statistical power analysis for the social and behavioral sciences. Basic and advanced techniques.* New York NY: Routledge.

Stigler, S. M. (1999). *Statistics on the table. The history of statistical concepts and methods.* Cambridge MA: Harvard University Press.