Pour que l'effet trouvé soit statistiquement significatif, il faudrait qu'il soit plus grand que 1,96 fois l'écart-type, c'est-à-dire au-dessus de 8,4% (l'intervalle de confiance ne contiendrait alors pas 0). Exprimé autrement: la probabilité de rejeter à tort la valeur nulle est de 5%.

Mais est-ce bien raisonnable d'envisager des effets significatifs aussi grands que 8,4%? Non, bien sûr. C'est ce que l'on nomme une erreur de magnitude. L'étude est construite de telle façon que tout résultat statistiquement significatif surestime le véritable effet (qui ne peut dépasser 1%). S'y ajoutent des erreurs de signe quand l'estimation trouvée est de signe opposé au véritable effet (ici: «les parents beaux ont plus de garçons que de filles»). Ainsi, on distingue deux types d'effet: positif, si les parents «beaux» ont plus de filles que les autres, et négatif, s'ils en ont moins.

Pour illustrer les probabilités associées à ces erreurs, envisageons quatre scénarios fondés sur des écarts-types égaux à 4,3% (*voir l'encadré ci-contre*). Ils montrent qu'une étude à partir de cette taille d'échantillon (3000 couples environ) n'est pas pertinente pour estimer des petits effets de l'ordre du 1%. Cela est dû notamment à la valeur de l'écart-type qui est particulièrement élevée. C'est pourquoi les études de la répartition des sexes des nouveau-nés utilisent des échantillons beaucoup plus grands, fondés sur de vastes bases de données démographiques, comptant plus d'un million d'individus.

L'ANALYSE BAYÉSIENNE

On résume tout ce que l'on sait sur l'effet à analyser, en utilisant des sources extérieures déjà connues, par une distribution *a priori*. Cette distribution sera modifiée (conditionnée) par le résultat de l'expérience pour donner la distribution *a posteriori*. Si l'on ne savait rien d'avance, on pourrait prendre une distribution *a priori* non informative et le résultat correspondrait à celui de l'approche fréquentiste. Ici, la distribution obtenue *a posteriori* serait alors approximativement une gaussienne, de moyenne 4,7% et d'écart-type 4,3%, ce qui correspondrait à une probabilité de 86% environ que l'effet réel soit positif. En général, plus la distribution *a priori* est concentrée autour de zéro (hypothèse selon laquelle l'effet réel sur la différence des sexes est faible), plus la probabilité *a posteriori* est proche de 50%.

Choisissons, par exemple, une distribution *a priori* gaussienne (en forme de cloche), centrée sur zéro avec une forme telle que la différence réelle de probabilité d'avoir une fille (selon que les parents sont beaux ou non) soit proche de zéro, avec des probabilités 50%, 90% et 94% d'être respectivement dans les intervalles: $[-0{,}3; 0{,}3]$, $[-1; 1]$, et $[-3; 3]$. Pourquoi centrer la distribution sur zéro? Parce que nous n'avons pas d'*a priori* sur le signe de la différence réelle de probabilité de naissance de filles en fonction de la beauté des parents.

À l'étape suivante, nous calculons, à partir de cette distribution *a priori* et des données, la distribution *a posteriori* de l'effet. Pour résumer, la distribution *a posteriori* donne une probabilité que l'effet soit positif (les parents beaux ont plus de filles) de seulement 58%, dont 45% que cette différence positive soit inférieure à 1%. Cette analyse dépend – mais peu – de la distribution *a priori*; par exemple, si l'on décide d'élargir la courbe de distribution autour de zéro, la probabilité que l'effet réel soit positif n'augmente que de 7% (pour atteindre 65%). Le fait de changer de famille de courbes de distribution a peu d'effets sur les résultats: les effets réels restent faibles, ce que confirment les données.

L'idéal scientifique dans l'inférence sur une quantité (ici, le lien entre la beauté des parents et le rapport des sexes à la naissance) est la >

QUATRE SCÉNARIOS POUR COMPRENDRE LES ERREURS

On peut étudier les erreurs liées aux résultats de Kanazawa en déterminant les probabilités associées *via* quatre scénarios. Dans le premier, que nous nommerons «Différence exactement nulle», supposons qu'il n'y a aucune corrélation entre la beauté des parents et le sexe de leur enfant (il n'y a pas de différence entre les deux groupes de parents). Il existe toujours une petite probabilité – 5% – que le résultat trouvé soit statistiquement significatif. Mais ce résultat sera quand même faux, on a cinq% de chances de se tromper sur la valeur et ce quel que soit le signe de l'effet, positif ou négatif.

Dans le scénario dit de 0,3% exactement, les parents beaux ont 0,3 % de probabilité supplémentaire d'avoir une fille – valeur plausible. On a toujours une probabilité de 5% de trouver un résultat statistiquement significatif à l'issue de l'étude. Comment se répartissent ces 5% de part et d'autre de la valeur réelle? Dans ces conditions, on a une probabilité de 3% d'observer un effet statistiquement significatif positif et 2% d'observer un effet statistiquement significatif négatif. Mais dans chaque cas, l'effet estimé – au moins 8,4% pour qu'il soit significatif – sera beaucoup trop éloigné de la valeur réelle de 0,3%. Ainsi, l'erreur de magnitude sera importante. Par ailleurs, la probabilité d'aller dans le mauvais sens – erreur de signe – sera égale à 2/5. Imaginons que l'on se retrouve entre les bornes de l'intervalle de confiance à 95%, donc proches de la valeur réelle. Dans ce cas, l'erreur de magnitude serait faible, mais la probabilité de faire une erreur de signe n'est pas négligeable: 47,5 %, proche des 50% observés en l'absence d'effet. Ainsi, le signe observé n'apporte pratiquement aucune information sur le sens d'une éventuelle différence.

Troisième scénario: si les parents séduisants ont une probabilité d'avoir une fille supérieure de 1% (ce qui semble être l'effet maximal possible), alors on a réciproquement des probabilités de 4 et 1% d'avoir des effets positif et négatif statistiquement significatifs. Globalement, il y a une probabilité de 40% de faire une erreur sur le signe de l'estimation; là encore, l'estimation donne peu d'informations sur le signe ou l'ordre de grandeur de l'effet réel.

Enfin, envisageons le quatrième scénario: même si la différence réelle était de 3% – valeur improbable –, il n'y aurait encore que 10% de probabilité d'obtenir un résultat statistiquement significatif. Dans ce cas, nous aurions une probabilité de 24% de faire une erreur de signe. Ce type d'étude est donc peu informatif.

> description de l'incertitude qui est résumée ici par une distribution de probabilités. Des chercheurs peuvent collecter des données ou analyser des données qui ont déjà été publiées de façon créative (comme l'a fait Kanazawa) et publier leurs résultats. Des métaanalyses peuvent être faites pour revoir tous ces résultats conjointement; elles lisseront les variations qui sont inhérentes à ces études sur de petits échantillons dans lesquelles la probabilité d'un effet positif peut passer de 50% à 58%, puis peut-être redescendre à 38% et ainsi de suite.

Comment reconnaître les données fiables? En collectant toujours plus de données. Chaque année, le magazine américain *People* publie une liste des 50 plus belles célébrités mondiales. Nous avons répertorié le sexe de leurs enfants pour les numéros parus entre 1995 et 2000 (*voir l'encadré page 30*). En 1995, par exemple, on a dénombré 32 naissances de filles pour 24 garçons, soit 57,1% de filles, ce qui correspond à 8,6% de plus que dans la population générale (48,5%). Un résultat en accord avec l'hypothèse de Kanazawa. Mais l'écart-type étant de 6,7%, l'estimation de 8,6% n'est pas statistiquement significative. Pour le confirmer, nous avons comparé ce résultat avec ceux des années suivantes. Entre 1995 et 2000, les plus belles personnes selon People ont eu 157 filles sur un total de 329 enfants, soit 47,7% de filles (pour un écart-type de 2,8%), ce qui est seulement 0,8% inférieur au chiffre obtenu pour la population générale. On ne peut rien en conclure...

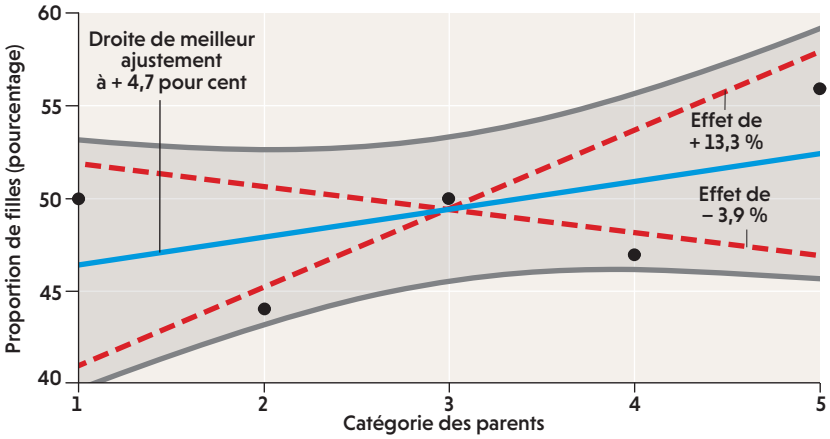
Pourquoi perdre notre temps à étudier des erreurs statistiques que personne n'a repérées? Pour deux raisons. D'abord, les résultats qui semblent avoir un sens sans être statistiquement significatifs sont les plus problématiques. Ensuite, certains médias et diverses publications scientifiques, par leur intérêt pour certains sujets de sociologie et leur sélection des résultats, biaisent la recherche en sciences sociales.

De fait, les résultats de Kanazawa ont tout de suite suscité l'intérêt des médias, jusqu'à des blogs du *New York Times*. Cette publication dans une revue à comité de lecture semblait être une caution suffisante balayant les doutes éventuels.

UN BRUIT ASSOURDISSANT

Qui plus est, l'effet n'a cessé d'augmenter. Ainsi, l'estimation – statistiquement non significative – de 4,7% que nous avons faite est passée à 8% dans l'analyse de Kanazawa (comparaison du groupe le plus beau à la moyenne des quatre groupes les moins séduisants), pour atteindre la valeur de 26% après une étude complémentaire introduisant d'autres corrections, avant de grimper à 36% pour des raisons encore floues!

Cette inflation nous surprit, ce chiffre étant 10 à 100 fois supérieur à tous les rapports filles-garçons publiés dans la littérature. Nous en avons conclu que, dans cette étude, le bruit (les données parasites) était supérieur au signal pertinent. La puissance statistique désigne la capacité de détecter une différence lorsqu'elle existe. Les études avec des échantillons plus importants ont toujours plus de puissance. Ainsi, si l'on veut affirmer quelque chose à propos d'effets de l'ordre de 1%, on a tout intérêt à partir de données pertinentes et à réaliser des tests qui exploitent bien les données. Cet exemple illustre bien le fait que les études qui manquent de puissance statistique ont peu de chances de parvenir à une pertinence statistique et, plus important encore, elles surestiment la taille des effets. Autrement dit, avec ces études, le bruit devient plus fort que le signal, c'est-à-dire l'effet observé.



Une analyse fréquentiste des données détermine la courbe de meilleur ajustement *(la droite bleue)* aux données de S. Kanazawa *(points en noir)*, pour un écart-type de 4,3%. Un intervalle de confiance à 95% montre que des effets aussi faibles que – 3,9% et aussi grands que 13,3% sont compatibles avec les données, car ils encadrent l'estimation à 4,7%.

Comment échapper à ce type de problèmes en sociologie? Aujourd'hui, la plupart des sujets de sociologie ont été passés au crible, et les chercheurs en sont donc réduits à étudier les petits effets. L'étude du rapport des sexes à la naissance est un sujet de société proche de nos préoccupations. Présenté sous forme d'une «vérité politiquement incorrecte», le résultat de Kanazawa, parce qu'il concerne les naissances, touche à des questions sensibles telles que l'avortement, le congé parental, le rôle de l'homme et de la femme dans la société.

On a vu que les études dont la pertinence statistique est insuffisante produisent des résultats aléatoires, parfois statistiquement significatifs, mais le plus souvent intuitifs. C'est un des points faibles de la psychologie évolutionniste: elle interprète des résultats aléatoires sans reconnaître la fragilité des explications qu'elle donne. Par exemple: les personnes jugées séduisantes auraient plus de chances d'être en bonne santé, riches et issues de groupes ethniques dominants, et plus